Automated Ultrasound Data Processing for Defect Detection and Characterisation Through Machine Learning

Vedran Tunukovic

Department of Electronic and Electrical Engineering

Centre for Ultrasonic Engineering

University of Strathclyde

A thesis submitted for the degree of

Doctor of Philosophy

May 2025

Copyright

Signed: 

Date: 16th May 2025

Acknowledgements

This work would not have been possible without many people's selfless help, contribution, and support.

Firstly, I would like to thank Dr. Ehsan Mohseni for his guidance and patience, which have been incredibly important to me, both in my research and personal life. My thanks equally extend to Prof. Gareth Pierce, Prof. Gordon Dobie, and Prof. Sandy Cochran, for encouraging me to explore different ideas and for pushing me to take on challenges I might not have approached on my own.

I am deeply grateful to everyone in the SEARCH lab, students and staff alike, for making this PhD a memorable experience. Special thanks go to Amine Hifi, whose energy and enthusiasm are highly contagious, and to Shaun McKnight, who was always willing to help and guide me. Without his advice and support, I wouldn't have been able to achieve so much during these years.

I'm also thankful to the Future Ultrasonic Engineering Centre for Doctoral Training for giving me the opportunity to pursue this research, and to industrial sponsor Spirit AeroSystems and Tom O'Hare for being open to new ideas and sharing their expertise.

I owe a heartfelt thanks to my parents, Ivana and Silvio, for the opportunities they have given me, and my siblings, Jan and Ema, for their unconditional support. Finally, I am eternally grateful to my wife, Nina, for standing by me through all the ups and downs and for providing me with the stability and peace I needed to see this journey through.

To Robin, this work is dedicated to you. You have brought us immeasurable joy that words cannot fully express.

Abstract

The growing adoption of Carbon Fibre Reinforced Plastic (CFRP) composites in safety-critical structures, such as aircraft fuselages and wind turbine blades, requires thorough inspection process to ensure material integrity and prevent catastrophic failures. These inspections are conducted using Non-Destructive Evaluation (NDE), a collective term for methods that assess the quality of materials without causing damage. Among these methods, Ultrasonic Testing (UT) stands out as a preferred inspection technique in the aerospace industry. The field of NDE has experienced significant advancements with the introduction of advanced sensor technologies and robotic manipulators, which have automated and accelerated data acquisition processes. However, data analysis and interpretation remain predominantly manual, making the process time-consuming and prone to human error, thus creating a bottleneck in manufacturing. Recent advancements in Artificial Intelligence (AI) present new opportunities to automate these tasks.

This thesis explores the application of AI techniques to analyse UT datasets obtained from reference CFRP samples representative of those used in the aerospace industry. The initial research focused on evaluating the performance of supervised AI methods, specifically object detection models, in defect detection tasks using ultrasonic C-scan images, which represent the top cross-sectional view of the inspected materials. As a baseline, both a traditional signal thresholding technique and an enhanced statistical thresholding method, based on theoretical mathematical distributions fitted to the observed data, were examined. The primary contribution of this work is the demonstration of the superior performance of AI models in this context over thresholding methods. Additionally, supervised training was conducted exclusively on simulated data, thereby addressing the data scarcity challenge.

Building on this, an unsupervised AI method in the form of anomaly detection was explored. This approach addressed the scarcity of datasets containing defective indications and the challenges of relying on large-scale simulations, which require significant computational resources and extensive manual effort to generate ground truths for supervised training. A two-step workflow was developed, comprising an automated signal gating method based on unsupervised clustering and an autoencoder

model serving as an anomaly detector applied to ultrasonic B-scans (which represent cross-sectional images of the material). The key advantages of this method include a streamlined development process focused on the use of pristine data (in this thesis, the term pristine refers to samples that contain no intentional or unintentional manufacturing defects that are detectable using the ultrasonic inspection setup employed, within the limits of its resolution), which is more readily available, and the elimination of ground truth generation requirements. This workflow was successfully applied to samples with both uniform thickness and complex geometries. Additionally, this research investigated the impact of human factors on AI results and highlighted the challenges posed by inconsistent data quality during scans.

The final stage of this research focused on strategies for integrating the developed AI models into NDE data analysis, driven by the recent evolution towards NDE 4.0 (a concept that combines digitalisation, automation, and connectivity to modernise NDE) focused on enabling fully automated systems. Despite significant research in this field, implementation strategies are often underexplored, and clearly defined automation levels achievable with AI remain lacking. To address these gaps, four levels of data analysis automation using AI were defined and evaluated. Additionally, the synchronous use of multiple AI models, each designed to process distinct views of the ultrasonic data, enabled cross-validation between models. This approach enhanced trust in the automated system while offering mechanisms to mitigate potential performance degradation of NDE operators using such systems. Furthermore, this strategy aligns with NDE 4.0 objectives, transitioning operators into supervisory roles while delegating repetitive tasks to the AI systems. The developed methods were evaluated in a case study involving complex geometry samples, demonstrating their effectiveness for potential industrial applications.

Overall, this thesis proposes methods to assist in automating NDE data analysis processes for UT of CFRP composites, with the primary goal of enhancing accuracy and reducing analysis time to address the current bottleneck in aerospace manufacturing.

Contents

List of Figures

List of Tables

Abbreviations

CFRP	-	Carbon Fibre Reinforced Plastics
NDE	-	Non-Destructive Evaluation
UT	-	Ultrasonic Testing
AI	-	Artificial Intelligence
GPU	-	Graphics Processing Unit
PAUT	-	Phased Array Ultrasonic Testing
FMC	-	Full Matrix Capture
TFM	-	Total Focusing Method
3D	-	Three-Dimensional
2D	-	Two-Dimensional
TOF	-	Time of Flight
TCG	-	Time Compensated Gain
ML	-	Machine Learning
DL	-	Deep Learning
NN	-	Neural Network
Tanh	-	Hyperbolic Tangent
ReLU	-	Rectified Linear Unit
GeLU	-	Gaussian Linear Unit
ELU	-	Exponential Linear Unit
MSE	-	Mean Squared Error
SGD	-	Stochastic Gradient Descent
ADAM	-	Adaptive Moment Estimation
CNN	-	Convolutional Neural Network
VGG	-	Visual Geometry Group
ResNet	-	Residual Network
FEA	-	Finite Element Analysis
AE	-	Autoencoder
KNN	-	K-Nearest Neighbours
SVM	-	Support Vector Machine
DBSCAN	-	Density-Based Spatial Clustering of Applications with Noise
LSTM	-	Long Short-Term Memory

YOLO	-	You Only Look Once
SSD	-	Single Shot Detector
VAE	-	Variational Autoencoder
GAN	-	Generative Adversarial Network
PC	-	Personal Computer
FT	-	Force-Torque
TCP/IP	-	Transmission Control Protocol/Internet Protocol
FBH	-	Flat Bottom Hole
CPU	-	Central Processing Unit
A-scan	-	Amplitude Scan
B-scan	-	Brightness Scan
RAM	-	Random Access Memory
R-CNN	-	Region-Based Convolutional Neural Networks
PDF	-	Probability Density Function
CDF	-	Cumulative Density Function
COCO	-	Common Objects in Context
RPN	-	Region Proposal Network
CSP	-	Cross Stage Partial
PAN	-	Path Aggregated Network
FPN	-	Feature Pyramid Network
IoU	-	Intersection Over Union
AUC	-	Area Under Curve
NMS	-	Non-Maximum Suppression
CAD	-	Computer-Aided Design
FPR	-	False Positive Rate
TPR	-	True Positive Rate
RMS	-	Root Mean Square
ROC	-	Receiver Operating Characteristic
UDP	-	User Datagram Protocol
ROS	-	Robotic Operating System
SSL	-	Self-Supervised Learning

Chapter 1: Introduction

1.1 Industrial Motivation and Research Context

Composite materials have become integral in industries such as renewable energy, biomedicine, aerospace, sports, construction, and the automotive sector. Among these composites, Carbon Fibre Reinforced Plastics (CFRPs) stand out due to their lightweight structure, strength-to-weight ratio, and resistance to fatigue and corrosion [1], [2], [3]. Aerospace and renewable energy are two main sectors that use CFRPs to construct safety-critical components.

In the aerospace industry, one of the earliest recorded applications of composites dates back to the introduction of the F-14 and F-15 military fighter jets in 1976, where boron-reinforced composites were utilised to construct empennages. By 1983, commercial aircraft like the Airbus A300 and A310 incorporated composites for secondary components [1]. In the following years, the use of composites moved from secondary to primary structural components. Nowadays, composite materials account for 53% and 50 % of structural weight for flagship aircraft models such as Airbus A350 XWB and Boeing 787 Dreamliner, respectively [4], [5]. The use of CFRPs also contributed to significant ecological benefits, with the Boeing 787 Dreamliner achieving 21% greater fuel efficiency than its predecessors [2].

An overview of several commercial and military aircraft models with their respective structural mass percentages of composite materials used is presented in Table 1. Figure 1 illustrates the increase in the adoption of composites in Airbus and Boeing commercial aircraft over the years.

Table 1 Examples of composite material usage in commercial and military aircraft, including the manufacturer, model, year, and structural percentage.

Manufacturer	Type	Model	Year	Structural %	Source
Airbus	Commercial	A320	1988	28%	[6]
Boeing	Commercial	777	1995	12%	[2], [7]
Airbus	Commercial	A380	2007	20 – 22%	[1], [8]
Boeing	Commercial	787	2011	50%	[2], [4]
Airbus	Commercial	A350 XWB	2015	53%	[5]
McDonnell Douglas	Military	F-15	1976	2%	[1], [9]
McDonnell Douglas	Military	F-18	1983	19%	[1], [10]
Eurofighter	Military	Typhoon	2003	40%	[11]
Lockheed Martin	Military	F-22	2005	24%	[2], [12]
Lockheed Martin	Military	F-35	2015	35%	[2], [13]

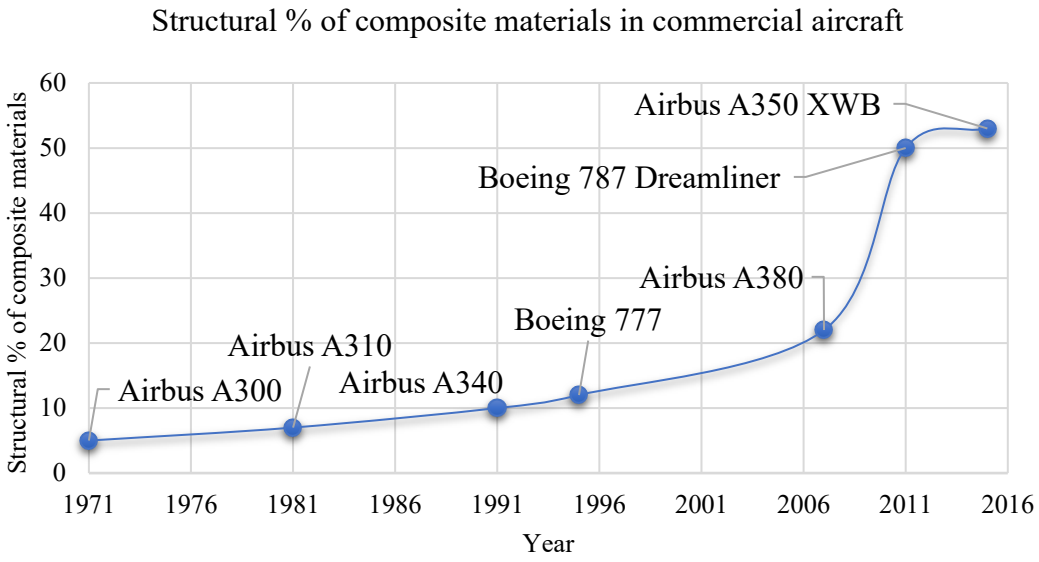


Figure 1 Structural percentage mass of composites used in Airbus and Boeing commercial aircraft models throughout the years.

However, while composites offer great benefits to industries mentioned above, several downsides have been recognised. These materials are highly anisotropic and prone to complex failure modes such as delamination, fibre breakage, matrix cracking, and porosity. Many of these defects may form during manufacturing or in-service use and often remain hidden beneath the surface, making visual inspection ineffective. Furthermore, composites typically fail in a brittle and sudden manner, without clear warning signs [14], [15].

There have been several important examples of composite material failure that illustrate the critical need for detailed inspection. For instance, Boeing 787 Dreamliner fleet underwent additional inspections after delamination issues were identified in composite fuselage sections. These flaws did not lead to catastrophic failures, but they highlighted the challenges of detecting and addressing defect early in service life [16].

Another widely discussed example was the crash of American Airlines Flight 587 in 2001 due to the vertical stabiliser failure. Although the primary cause was determined to be human error in turbulent flight conditions, prior report had noted delamination in the same structural component, which drew the attention to the need for thorough inspection procedures [17].

Given these risks, the use of composite materials in high-value, safety-critical components requires comprehensive Non-Destructive Evaluation (NDE), a process of examining materials without causing damage or altering their functionality. Common NDE techniques include Ultrasonic Testing (UT), radiographic testing, eddy current testing, and visual testing. The choice of a suitable NDE method depends on various factors, such as the material's characteristics, geometry, safety considerations, and the specific defects being targeted. UT has become a preferred modality in aerospace due to its ability to provide volumetric inspection, assess internal structure, and offer flexibility and ease of use, making it ideal for detecting a wide range of defects. The approximate market size of the NDE industry in The United States in 2021 was 2.43 billion dollars, with UT representing a major portion, approximately 660 million dollars [18]. Growth projections indicate that the NDE market could reach 16.66 billion dollars by 2029, driven primarily by the aerospace industry's high standards for material quality and the oil and gas industry's need for robust pipeline inspection methods.

Advancements in NDE sensor technology have been paralleled by improvements in robotics, particularly through the integration of sensors with industrial manipulators for deployment. As a result, precision and repeatability of NDE processes greatly increased, reducing the reliance on manual inspection skills of highly trained certified inspectors [19], while accelerating the data acquisition process by generating large volumes of data in a very short time. However, data interpretation remains

predominantly manual and lags behind acquisition speeds, increasing the risk of human error and creating bottlenecks in manufacturing [20], [21]. Therefore, the development of automated data interpretation tools to aid the decision-making process would complement the automated robotic systems currently used within the industry, serving as a key enabler for realising the full potential of industrial automation in line with the NDE 4.0. Artificial Intelligence (AI) technology has been recognised as a key enabler for resolving data interpretation bottlenecks and tackling repetitive tasks currently performed by humans. However, the uptake and implementation in industry settings is still limited, particularly in safety-critical sectors such as aerospace. This is due to several factors, including the relatively early stage of research into the application of AI for NDE, and the lack of consensus around the fundamental requirements AI systems must meet to be considered reliable and trustworthy for such applications [22]. Furthermore, safety-critical industries are typically cautious in adopting new technologies as the consequence of failure can be severe. Lastly, while broader AI standards are beginning to emerge (e.g., first legal framework on the use of AI in the European Union the "AI Act" came into force in August 2024 and will gradually be implemented in practice, starting with February 2025 [23], [24]), similar regulatory guidelines for AI enabled NDE are lacking. Some efforts do exist, most notably a set of technical reports and recommended practices published by the European Network for Inspection and Qualification (ENIQ) [25], as well as the "Handbook of Nondestructive Evaluation 4.0" [26], a 2025 publication that is only beginning to explore the potential impact and opportunities of AI and how it may shape the future NDE landscape.

In the last decade, the field of AI has experienced a surge in research interest, particularly in computer vision with seminal works such as the development of powerful model architectures like AlexNet [27] and ResNet [28], as well as the use of large datasets like ImageNet [29] for model training. One of the key drivers of this surge is the demonstrated ability of AI models to outperform humans in complex and high-dimensional tasks. For example, the AlphaGo and AlphaZero fundamentally changed the way how professional Go and chess players approach their respective game, and no human has matched their performance since [30], [31]. AI models have also achieved diagnostic accuracy that exceeds that of experienced trained medical

professionals in specific medical tasks [32] and are showing increasing promise in outperforming NDE professionals [21]. Additionally, AI model AlphaFold has resolved difficulty solvable problems like protein structure prediction, significantly accelerating research in the biomedical sciences.

Another factor behind the increased interest in AI is the limitation of traditional rule-based system when applied to real-world data, which is often noisy and subjective to variability. As discussed in Chapter 4, such systems tend to struggle under these conditions due to their rigidity. In contrast, AI models can learn directly from data and adapt to patterns that may be difficult or impossible to manually specify.

Finally, significant progress in the Graphics Processing Unit (GPU) market, alongside the availability of open-sourced datasets (e.g., Kaggle competitions [33]), and development of AI frameworks (such as PyTorch and TensorFlow) significantly lowered the barrier for entry for research.

Although evaluating the market size for AI technologies is challenging, some reports suggest that the market, valued at 757 billion dollars in 2025, is expected to grow approximately fivefold by 2034 [34].

The outlined driving factors for the future production of CFRPs, combined with the growing adoption of automated robotic setups for NDE, and the emergence of new AI technologies, provide strong motivation to explore and research automated solutions for addressing data interpretation bottlenecks.

1.2 Aims and Objectives

The main aims and objectives of this PhD thesis are:

- To investigate and understand the current state of the art in UT methods used for NDE of CFRPs, with a specific focus on the aerospace industry.
- To investigate the current state of the art in research on the application of AI technologies within NDE workflows and assess their adoption and use in industrial practices.
- To acquire and analyse UT data from various CFRP reference samples using a robotic manipulator setup that mirrors industrial practices.

- To develop and evaluate AI-driven workflows for the automated processing of ultrasonic inspection data, with the goal of improving defect detection and characterisation, and accelerating analysis.
- To explore both supervised and unsupervised AI approaches, including the use of synthetic data and domain-specific augmentation strategies for model training.
- To investigate multi-model analysis across multiple ultrasonic scan views, and to propose a tiered framework for integrating AI into existing NDE workflows.

1.3 Outline of Thesis Structure

The remainder of the thesis is structured as follows:

- Chapter 2 provides an overview of UT, CFRPs, AI, and the application of AI methods in NDE. This chapter serves as a background research and literature review to identify gaps in the knowledge and outline the motivation for the present work.
- Chapter 3 describes the automated NDE robotic setup used to capture the UT data used in this thesis. It details the UT and robotic setups, data processing techniques, CFRP reference samples, and the used hardware.
- Chapter 4 presents a comparative study of traditional and AI methods for defect detection in ultrasonic amplitude C-scans.
- Chapter 5 introduces a two-stage unsupervised approach for ultrasonic Brightness scan (B-scan) analysis, based on an automated gating method and an anomaly detection AI model.
- Chapter 6 explores strategies for implementing developed AI models into NDE workflows, with a focus on the simultaneous processing of different ultrasonic views to achieve comprehensive analysis.
- Chapter 7 concludes the main findings and limitations of the thesis and discusses potential future research trajectories.

1.4 Contributions to Knowledge

The work presented in this thesis is focused on the automated analysis of UT data from CFRP samples. The key contributions are summarised:

- Successful training of object detection AI models using purely synthetic ultrasonic C-scan data, addressing the lack of open, labelled datasets in this domain. Transfer learning techniques were employed by incorporating domain-specific augmentations derived from noise profiles of real inspection data, helping bridge the gap between simulated and real-world conditions. The models were validated on real inspection data and showed generalisable performance. A comparative analysis was conducted against traditional statistical thresholding methods, which remain standard practice in industry.
- Development of a two-step data analysis workflow for ultrasonic B-scans, which includes an automated signal gating method to remove prominent geometrical features from the captured data using unsupervised clustering, and an AI-based anomaly detection model trained exclusively on pristine (in this thesis, the term pristine refers to samples that contain no intentional or unintentional manufacturing defects that are detectable using the ultrasonic inspection setup employed, within the limits of its resolution) experimental data, thus reducing the need for labour-intensive ground truth labelling.
- Proposal and discussion of integration strategies for the developed AI models, focusing on automation levels ranging from operator assistance tools to highly automated solutions that are adaptable to various risk profiles. While similar integration challenges and solutions have been explored in different fields (e.g., biomedical and medical domains), this thesis contributes the framework for implementing such strategies in NDE data analysis workflows.
- Development of a multimodal, multiview ensemble approach to defect detection that synchronously combines outputs from three AI models analysing different ultrasonic data views, including B-scans, C-scans, and 3D volumetric inspections. This ensemble design provides a cross-validation mechanism and flexible implementation strategies, addressing limitations in earlier NDE research which typically does not leverage multiple data views in combination.

1.5 Lead author publications arising from this thesis

[1] **Vedran Tunukovic**, S. McKnight, E. Mohseni, S. G. Pierce, R. Pyle, E. Duernberger, C. Loukas, R. K.W. Vithanage, D. Lines, G. Dobie, C. N. MacLeod, S. Cochran, T. O'Hare, "A study of machine learning object detection performance for phased array ultrasonic testing of carbon fibre reinforced plastics", *NDT & E International, Volume 144, 2024, 103094, https://doi.org/10.1016/j.ndteint.2024.103094.*

[2] **Vedran Tunukovic**, S. McKnight, R. Pyle, Z. Wang, E. Mohseni, S. G. Pierce, R. K. W. Vithanage, G. Dobie, C. N. MacLeod, S. Cochran, T. O'Hare, "Unsupervised machine learning for flaw detection in automated ultrasonic testing of carbon fibre reinforced plastic composites", *Ultrasonics, Volume 140, 2024, 107313, https://doi.org/10.1016/j.ultras.2024.107313.*

[3] **Vedran Tunukovic**, S. McKnight, A. Hifi, E. Mohseni, S. G. Pierce, R. K.W. Vithanage, G. Dobie, C. N. MacLeod, S. Cochran, T. O'Hare, "Human-machine collaborative automation strategies for ultrasonic phased array data analysis of carbon fibre reinforced plastics", *NDT & E International, Volume 154, 2025, 103392, https://doi.org/10.1016/j.ndteint.2025.103392.*

1.6 Co-author publications arising from this thesis

[1] Shaun McKnight, C. MacKinnon, S. G. Pierce, E. Mohseni, **Vedran Tunukovic**, C. N. MacLeod, R. K. W. Vithanage, and T. O'Hare, "Three-Dimensional Residual Neural Architecture Search for Ultrasonic Defect Detection," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, Volume 71, 2024, doi: 10.1109/TUFFC.2024.3353408.*

[2] Shaun McKnight, **Vedran Tunukovic**, S. G. Pierce, E. Mohseni, R. Pyle, C. N. MacLeod, 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, Volume 71, 2024, doi: 10.1109/TUFFC.2024.3408314.*

[3] Shaun McKnight, **Vedran Tunukovic**, A. Hifi, S. G. Pierce, E. Mohseni, C. N. MacLeod, T. O'Hare, "Three-dimensional ultrasonic self supervised segmentation", *Engineering Applications of Artificial Intelligence, Volume 154, 2025, doi: 10.1016/j.engappai.2025.110870*

For full research output, please see: Research output

Chapter 2: Background Research

2.1 Ultrasonic Testing

Ultrasound refers to acoustic waves whose frequency exceeds 20 kHz, placing them beyond the human hearing range [35]. Ultrasonic testing is the most used NDE method due to its flexibility, ease of use, and safety [36], [37], and it is employed in industries such as aerospace, construction, automotive, and energy for inspection of metal and composite materials [38], [39], [40], [41]. A range of UT techniques is available, such as the conventional piezoelectric transducer method, laser UT, and Phased Array Ultrasonic Testing (PAUT), each suitable to specific applications. These techniques utilise different wave propagation modes, including longitudinal, shear, surface, and guided waves, depending on the inspection requirements. This thesis focuses on the use of normal-incidence longitudinal ultrasonic waves for volumetric inspection, as this is the standard industrial practice for CFRP inspection. Longitudinal waves are preferred as most internal defects in composites tend to be oriented parallel to the surface, making normal beam inspection effective for their detection.

UT offers several advantages, including near-instant display of results, flexible operating frequencies that can be tuned based on the inspection requirements and the material composition, and the ability to examine the internal structures without the need for extensive component preparation procedures [36], [37], [42], [43]. However, some disadvantages of UT have been recognised. For thicker objects, the penetration depth of ultrasound may be insufficient to propagate effectively through the entire structure, depending on the material composition and test frequency [44]. Additionally, conventional UT is unable to detect defects smaller than half of the wavelength of the incident soundwave, due to fundamental diffraction limits [44]. While some experimental techniques have explored sub-wavelength detection using AI super-resolution approaches [45], these remain in early research stages. Similarly, defects located parallel to the ultrasound beam may go undetected [37], [44].

2.1.1 Fundamentals of Ultrasound

Ultrasonic waves are defined as self-sustaining mechanical waves that cause a series of compressions and rarefactions when propagating through a medium. Common types of ultrasound waves include longitudinal, transversal, Rayleigh, and Lamb waves. The

ultrasound wave propagation is extensively covered in existing literature [46], [47], therefore only one-dimensional wave equation in isotropic media is presented in Eq.1:

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 p}{\partial t^2} \Leftrightarrow \frac{\partial^2 u}{\partial x^2} = \frac{1}{v_0^2}\frac{\partial^2 u}{\partial t^2} \quad \text{Eq.1}$$

Where p is the pressure amplitude, x is the one-dimensional distance, v is the wave speed, t is the time, and u is the particle displacement amplitude. This form of the wave equation applies specifically to longitudinal waves propagating in a linear, isotropic medium.

Operating frequency is a crucial parameter as it directly determines both the resolution and penetration depth. Higher-frequency ultrasonic waves have shorter wavelengths, allowing for interaction with smaller features within the material at the expense of decreased penetration depth. In contrast, lower frequency operation produces waves of longer wavelengths that penetrate deeper into the material but sacrifice resolution. Therefore, achieving a suitable balance between operating frequency, penetrating depth, and the required level of detail resolved by the waves is essential when choosing UT parameters. The relationship between operating frequency and wavelengths is given in Eq.2:

$$f = \frac{v}{\lambda} \quad \text{Eq.2}$$

Where f is the operating frequency of the ultrasonic system and λ is the wavelength.

This equation assumes a constant wave velocity v, which is valid for homogeneous and isotropic materials. However, CFRPs are inhomogeneous and anisotropic, and wave velocity varies with direction and frequency (making CFRPs dispersive media). In practical UT, a constant velocity is often assumed as an approximation/simplification. However, advanced imaging techniques (discussed in section 2.1.8) require a more accurate velocity model to account for variations within the material for reliable image reconstruction.

2.1.2 Ultrasonic Beam

In its simplest form, UT waves are generated using a piezoelectric crystal that enables the conversion of input electrical voltage into mechanical vibrations and vice versa

[48]. These vibrations produce ultrasonic waves of high frequency that propagate through materials and are altered due to the scattering and attenuating mechanisms.

When an ultrasonic transducer is excited, it generates pressure waves which propagate and form ultrasonic beams described by three distinct areas. The area closest to the source of ultrasound is characterised by both constructive and destructive wave interactions, creating a non-uniform distribution of energy and intensity as the ultrasonic beam is still converging. This can cause artefacts during the scan and diminish the quality of the imaging, leading to inconsistent measurements and complex interference patterns. This zone is called the near field or Fresnel zone and is often avoided in NDE applications, particularly for single element transducers, through transducer design, stand-off wedges to offset the transducer from the inspected material, focusing techniques, or shorter ultrasonic pulse [37].

The middle portion of the beam is called the focal zone and is characterised by the smallest beam diameter, resulting in overall highest lateral resolution. Additionally, this zone contains the peak pressure point measured from the source, which can be approximated with Eq.3:

$$N \approx \frac{D^2}{4\lambda} \qquad \text{Eq.3}$$

Where N is the distance from the source to the peak pressure point and D is the dimension of the circular ultrasonic transducer. The focal zone is the most suitable part of the beam for NDE inspection as it results in a high energy measurement and a good resolution, making it particularly efficient for the inspection of smaller parts and defects [49]. In focused single-element transducers, the focal point is fixed by design, whereas in advanced techniques like PAUT, the focal zone can be dynamically adjusted through ultrasonic setup parameters. This is achieved through the application of delay laws, where each element in the array is excited at controlled time steps. By adjusting these delays, the emitted wavefront can be steered and focused at different depths or angles without moving the probe physically. This principle and its implementation are further detailed in Section 2.1.8. Lastly, the far field or Fraunhofer zone is further away from the source and can be described by a uniform ultrasonic beam with fewer energy fluctuations, and its distance from the source is described with Eq.4:

$$N \approx \frac{4D^2}{\lambda} \quad \text{Eq.4}$$

The factors mentioned above (wavelength and transducer dimensions) also influence beam spread, a phenomenon where ultrasonic wave diverges from its initial direction as it propagates through the material. Lower-frequency waves and smaller transducer sizes result in greater beam divergence, which can be estimated using Eq.5 [50]:

$$\sin \Theta = 1.22 \frac{\lambda}{D} \quad \text{Eq.5}$$

Where Θ is the angle of divergence. Excessive beam spread reduces spatial resolution but can be mitigated by using higher-frequency transducers, larger transducer apertures, or focusing techniques, as further discussed in Section 2.1.8. A simplified illustration of an ultrasonic beam with the above-described zones is shown in Figure 2.

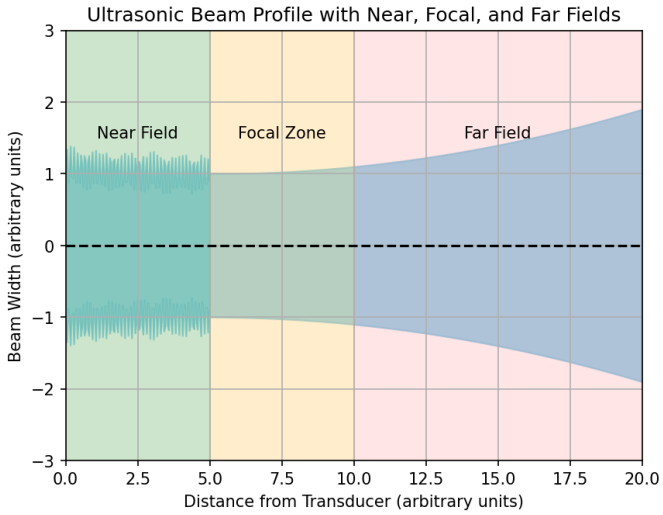


Figure 2 An example illustration of an ultrasonic beam with near field, focal zone, and far field marked

2.1.3 Wave Velocity

Wave velocity refers to the speed at which ultrasonic waves propagate through a material and is primarily influenced by the material properties. For longitudinal waves, velocity can be determined through Eq.6 [51]:

$$v = \sqrt{\frac{E}{\rho}} \quad \text{Eq.6}$$

Where E is Young's modulus, and ρ is the material's density. Wave velocity varies within the same material depending on the wave mode. For example, shear waves, which depend on the shear modulus G, typically propagate at lower speeds than longitudinal waves [52].

UT relies on measuring the time taken for a wave to travel to reflecting features within the inspected material (i.e. front wall, back wall, or reflective defects) and return to the transducer. This Time-of-Flight (ToF) measurement allows for depth estimation, and is given by Eq.7:

$$d = \frac{v * t}{2} \qquad \text{Eq.7}$$

where d is the distance travelled, and t is total travel time. Precise knowledge of wave velocity is crucial for accurate feature localisation and material thickness measurements. However, in highly anisotropic materials such as CFRPs, wave velocity estimation becomes complex due to variations in stiffness and density across different fibre orientations, which are further discussed in Section 2.1.8 and Chapter 4.

2.1.4 Attenuation

Attenuation of the ultrasonic waves refers to the loss of energy during propagation, resulting from multiple underlying phenomena such as the conversion of wave energy to kinetic energy in the form of heat and perceived attenuation due to scattering. Scattering occurs when ultrasonic waves interact with features smaller than their wavelength, causing energy dissipation. This phenomenon is particularly pronounced in CFRPs, which exhibit high levels of attenuation due to their anisotropic nature, as well as the non-linear relationship between attenuation and frequency, influenced by factors such as complex fibre orientation, layering, and heterogeneity within the material [37], [53], [54]. For comparison, metallic materials generally exhibit a more linear attenuation with respect to frequency, where signal loss increases in a relatively predictable manner as the frequency rises [55]. Attenuation can be mathematically expressed with Eq.8 [56]:

$$A = A_0 e^{-\alpha x} \qquad \text{Eq.8}$$

Where A is the pressure amplitude of the ultrasonic wave after transmission, A_0 is the pressure amplitude of the initial pulse, α is the attenuation coefficient, and x is the travelled distance. The usual practice is to present attenuation with a logarithmic scale with Decibel (dB) units:

$$Attenuation = 20log_{10}(A_0/A) \quad \text{Eq.9}$$

In other words, when features of the same size are present at varying depths within the sample, those located at larger depths with greater acoustic path from the source show lower amplitude signals compared to closer features. The effects of attenuation are demonstrated in sections 2.1.9 and 2.1.12, while methods to compensate for this phenomenon are discussed in Section 2.1.14.

2.1.5 Conventional Ultrasonic Testing

The first conventional UT method is the pulse-echo technique, where the same ultrasonic transducer produces an ultrasound pulse that propagates into the material and records the returning reflections. Depending on the mode of application, the transducer is either air-coupled, used in immersion, or directly coupled in contact to the object with the help of a coupling medium. The coupling medium (or couplant) helps to address the acoustic impedance mismatch between the mediums and eliminates air pockets which act as strong reflectors, promoting overall energy propagation into the material [57]. Typically, in manual ultrasonic inspection, a thin layer of coupling gel is used, whereas, in automated UT inspection, it is usually a thin layer of water. The constant thickness and proper application of coupling are critical to the performance of the ultrasonic setup as inconsistencies can lead to the decoupling of the ultrasonic probe and uneven energy transmission into the material [58]. From a data-driven perspective, such uneven energy transmission creates inconsistencies in the recorded amplitudes and introduces unexpected variability and noise into the ultrasonic signals, which can degrade the performance and robustness of AI models trained on such data (this phenomenon is further discussed in section 5.8).

Upon encountering object boundaries, defects, or other discontinuities, a portion of the initial wave is reflected to the transducer, where the pressure amplitude is received and converted to voltage owing to the piezoelectric effect. Recorded wave amplitudes can be visualised as an Amplitude scan (A-scan), a representation of wave amplitudes over

time. In pristine samples, areas of higher amplitude typically correspond to front and back wall reflections. However, when defects are present, part of the wave is reflected off the defect and back to the transducer, creating an additional indication in the A-scan that appears earlier in time than the back wall reflection. Pulse-echo mode of operation is presented in Figure 3.

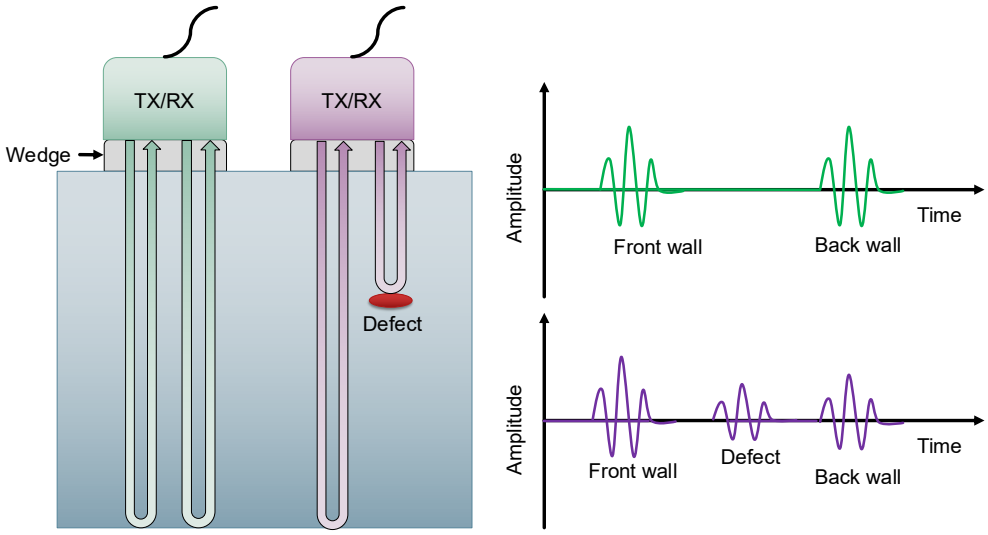


Figure 3 Illustration of pulse-echo inspection configuration with corresponding A-scan representations. Green indicates a pristine scan, while purple represents a defective scan.

The use of a single transducer for both transmission and reception induces a ringdown effect which can obscure portions of the useful signal, thus creating a dead (or blind) zone [59]. This masking effect can be mitigated by using acoustic wedges that offset the original pulse away from the surface. The use of wedges necessitates adjustment in the ultrasonic setup depending on the geometry and properties of the wedge. Additionally, wedges can be in 0° configuration for normal incidence or angled to redirect the beam at an angle.

Since both transmission and reception of UT signals are performed on the same surface, the pulse-echo technique is suitable for industrial applications where access is limited to one side. On the downside, the pulse-echo method faces challenges with rough surfaces that obstruct the wave penetration, very thin materials where a large portion of the A-scan can be masked by initial pulse reverberations, and near-surface discontinuities located in the near-field of the ultrasonic beam.

The second type of conventional ultrasonic testing is the pitch-catch mode, where two ultrasonic transducers are used. These transducers can be placed on the same or opposite surfaces of the material (in which case it is referred to as through-transmission), depending on the application. In this mode, one transducer generates ultrasonic waves, and the other receives it, and unlike pulse-echo, dead zones in the signal are eliminated [26]. This mode requires precise alignment of the transducers, and in through transmission, the depth information about features within the material is lost. Pitch-catch is often used in inspection of welds [43], inspection of polymers [44], and structural health monitoring for corrosion [45]. However, this method is generally not applicable for composite materials due to their anisotropic nature. Angled or steered beams in such materials are more susceptible to scattering and attenuation, often resulting in degraded signal quality compared to pulse-echo techniques. While some studies have explored the use of pitch-catch for static inspection of unidirectional composites [60], these approaches are limited in scope and are not well suited to the automated scanning strategies employed in this thesis. Therefore, the focus is shifted towards pulse-echo methods. The operating principle of through transmission pitch-catch ultrasonic scanning is presented in Figure 4.

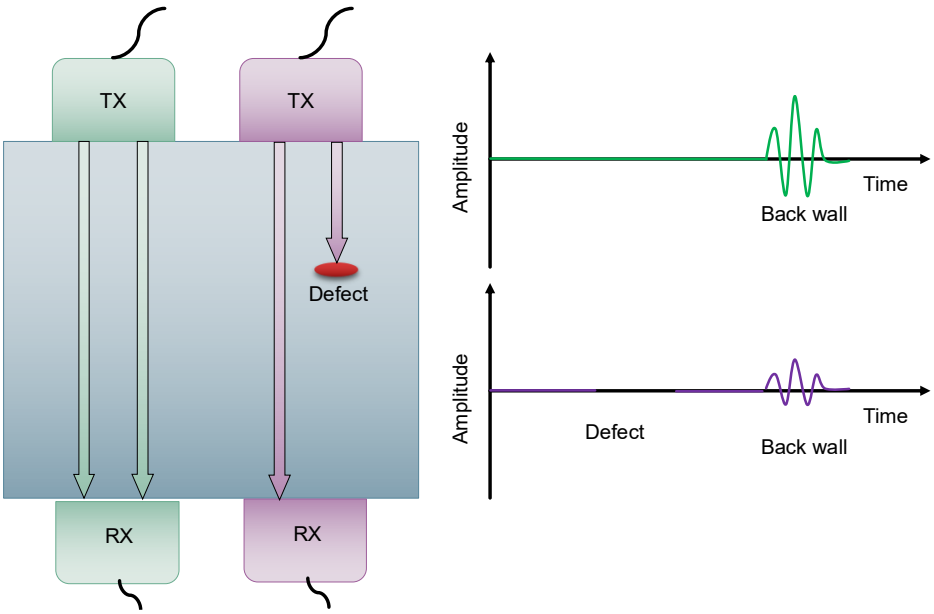


Figure 4 Illustration of through transmission pitch-catch inspection configuration with corresponding A-scan representations. Green indicates a pristine scan, while purple represents a defective scan.

2.1.6 Normal Incidence

Normal incidence occurs when an ultrasonic wave approaches an interface between two mediums perpendicularly. At an interface with another material, acoustic waves can change due to phenomena of reflection, diffraction, transmission, refraction, and mode conversion. Generally, the difference in acoustic impedance between two materials determines how much the incident wave is altered at the interface. Materials with similar acoustic impedance cause fewer changes to the wave compared to those with highly dissimilar impedance. Acoustic impedance, measured in kg/sm^2, quantifies a material's resistance to the transmission of waves and is given in Eq.10:

$$Z = \rho v \qquad \text{Eq.10}$$

Where Z is the acoustic impedance, and ρ is the density of the material. Changes in wave energy at interfaces are characterised by transmission and reflection coefficients, which define the amount of wave energy transmitted or reflected. These coefficients are determined by the acoustic impedance of the materials involved and are represented in Eq.11 and Eq.12:

$$R = \frac{Z_2 - Z_1}{Z_1 + Z_2} \qquad \text{Eq.11}$$

$$T = \frac{2 * Z_2}{Z_1 + Z_2} \qquad \text{Eq.12}$$

Where Z_1 and Z_2 are acoustic impedances of the first and second medium, respectively. Normal incidence corresponding to Eq.11 and Eq.12 is illustrated with Figure 5.

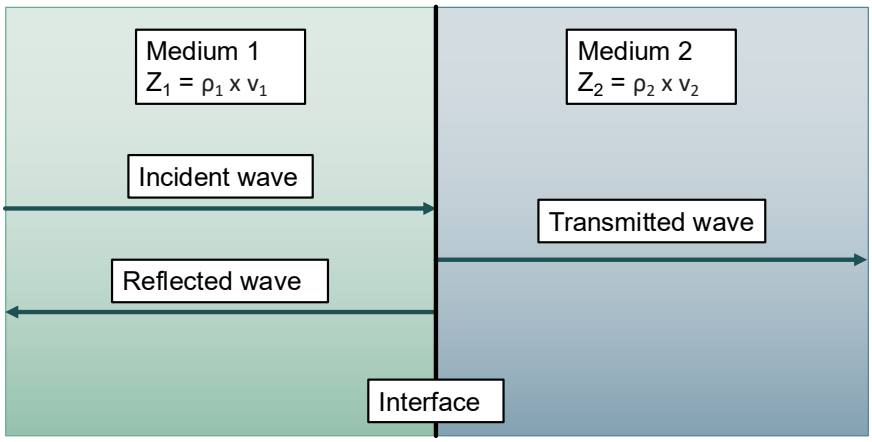


Figure 5 Illustration of transmitted and reflected waves at the interface with normal incidence between two mediums.

2.1.7 Snell's Law and Oblique Incidence

When an ultrasonic wave encounters an interface between two mediums at an angle, part of the wave is reflected into the original medium, while the other part is refracted into the subsequent medium. Refraction results in a change in waves direction (angle), and can be described by Snell's law [47], as shown in Eq.13. This phenomenon is visually represented in Figure 6.

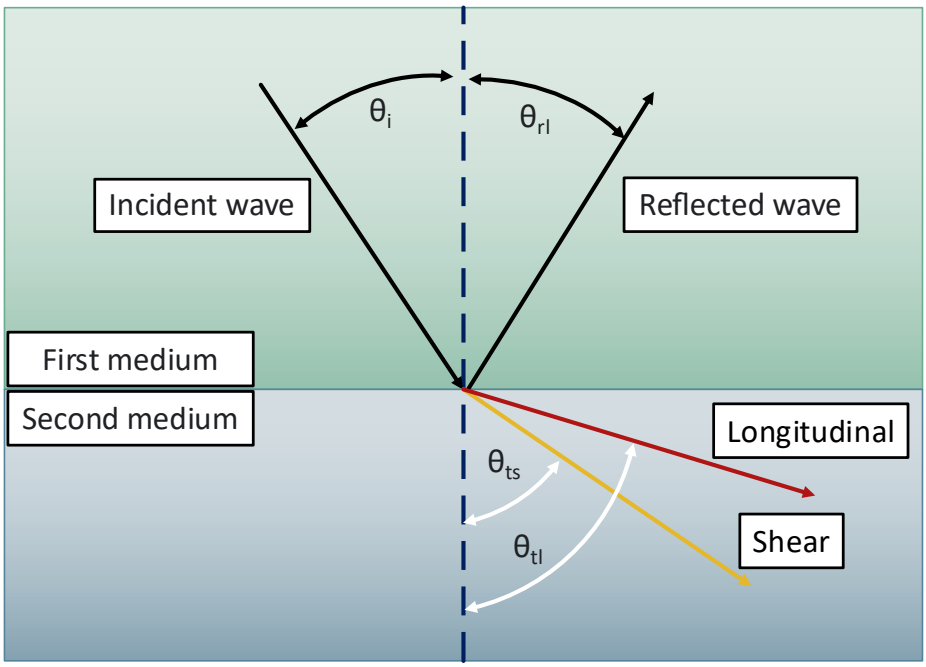


Figure 6 Illustration of reflected and transmitted waves at the interface with oblique incidence between two mediums.

$$\frac{\sin\theta_i}{v_i} = \frac{\sin\theta_{rl}}{v_{rl}} = \frac{\sin\theta_{ts}}{v_{ts}} = \frac{\sin\theta_{tl}}{v_{tl}} \qquad \text{Eq.13}$$

Where:

θ_i and v_i are the angle and wave speed of the incident wave

θ_{rl} and v_{rl} are the angle and wave speed of reflected longitudinal wave

θ_{ts} and v_{ts} are the angle and wave speed of transmitted transversal wave

θ_{tl} and v_{tl} are the angle and wave speed of transmitted longitudinal wave

In instances where the refracted longitudinal wave exceeds 90° angle (critical angle), the refraction is accompanied by mode conversion. Mode conversion is a physical phenomenon where incident wave creates waves of different types (e.g., longitudinal

waves create shear waves). This behaviour is often utilised in NDE where spatial resolution can be improved with shear waves which have shorter wavelengths [61], [62].

While this description provides insight into wave interaction at a high level, factors such as the angle of incidence, surface roughness, absorption, attenuation, and the type of bonding between interfaces influence transmission and reflection. These factors contribute to complex and nonlinear behaviour that depends on specific properties at a smaller, local scale.

These physical factors impact data-driven methods by introducing variability and stochastic effects in the ultrasonic signals that are difficult to model precisely. Finite Element Analysis (FEA) simulations attempt to capture some of these complexities, but many aspects remain challenging to predict or simulate. As a result, data-driven approaches must account for inherent noise and uncertainty in the input data to produce generalisable outcomes. While real experimental data naturally includes this variability and can help models learn to cope with it, simulated datasets often lack such stochastic characteristics, making it more difficult for models trained solely on synthetic data to generalise to real-world scenarios (this is further discussed in section 4.4).

2.1.8 Phased Array Ultrasonic Testing

In recent years, there has been a noticeable increase in the adoption of PAUT [36], [63]. Ultrasonic arrays consist of multiple individual piezoelectric transducers, arranged in an array configuration. PAUT has found wide application in various NDE tasks due to their flexibility, high resolution imaging, and reduced inspection times due to a larger coverage area [58], [64]. Each ultrasonic array can be described by several key characteristics: the number of individual elements it comprises, the elevation (the size of elements in the dimension normal to the inspection surface), the pitch (the distance between the midpoints of neighbouring elements), and the kerf (the size of the gap between two individual ultrasonic elements). Array elements and its properties are illustrated in Figure 7.

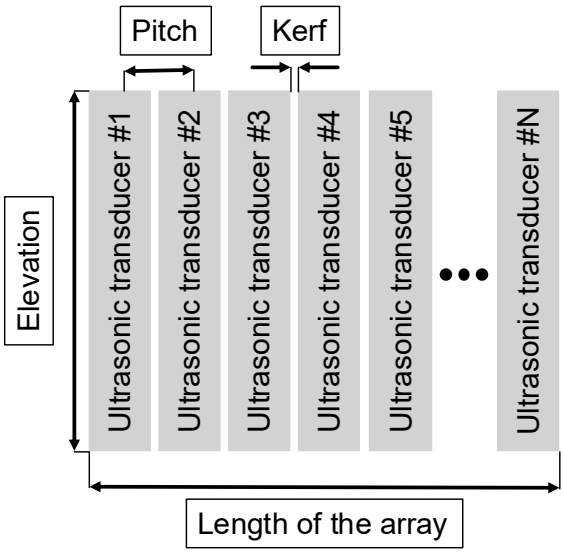


Figure 7 Ultrasonic array schematic illustrating pitch, kerf, elevation, and length of the array,

The scanning with PAUT arrays is typically performed by moving along the component in a direction perpendicular to the transducer's elevation axis. While exciting all elements at once to generate a plane wave, similar to the operation of a single transducer is possible, the array elements can be excited/activated for reception individually at different time stamps. This allows for the creation of a series of time delays, also known as delay laws, to achieve different ultrasonic beam forms and propagation orientations. Different forms of delay laws enable:

- Beam steering, where the propagation angle for the generated wavefront can be changed with respect to the arrays' normal.
- Focusing, where parabolic delay laws create a focal point for the generated beam.
- Creation of sub-apertures, where a subgroup of array elements is excited simultaneously to generate more energy, create larger wavefronts that travel into the sample, and improve the signal-to-noise ratio [65], [66], [67].

Examples of beam steering, beam focusing, and sub-apertures are presented in Figure 8.

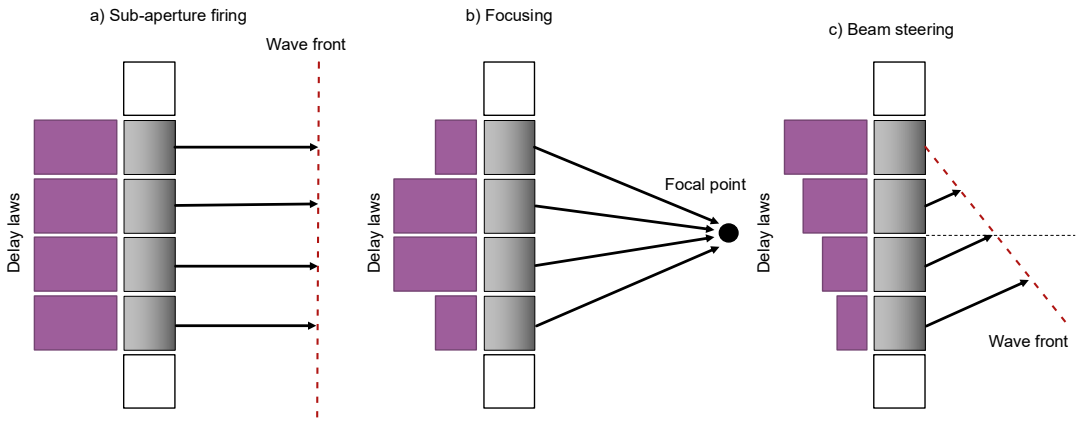


Figure 8 Illustration of phased array delay laws: a) sub-aperture firing: b) focusing, c) beam steering.

This electronic control of array elements is a significant advantage of PAUTs, reducing the need for physical alteration of the probe's position and angle. Lastly, the introduction of PAUT enabled the use of newer data acquisition methods, such as Full Matrix Capture (FMC), which records the full combination of transmit/receive possibilities in an array configuration to be leveraged to form high-resolution advanced images capturing greater detail, with algorithms such as Total Focusing Method (TFM) [68].

However, these techniques cannot be effectively used with CFRPs due to the highly anisotropic and dispersive nature, which causes the wave velocity to vary with both direction and frequency. This violates the assumptions of constant wave speed underlying the 1D wave equation (as shown in Eq.*1*). Since the wave velocity is critical for ToF calculations in TFM, this variability introduces challenges that can result in imprecise imaging. Although some adjustments have been made to algorithms to address these issues with CFRPs [69], [70], such techniques remain largely limited to laboratory settings, with minimal industrial uptake. Furthermore, ray tracing calculations used in TFM are computationally intensive and time-consuming, making real-time imaging difficult. As a result, electronic scanning methods such as linear scans remain widely accepted in the industry, providing sufficient performance for practical NDE applications.

To aid with the following chapters, Figure 9 presents the different ultrasonic views/projections used throughout this thesis.

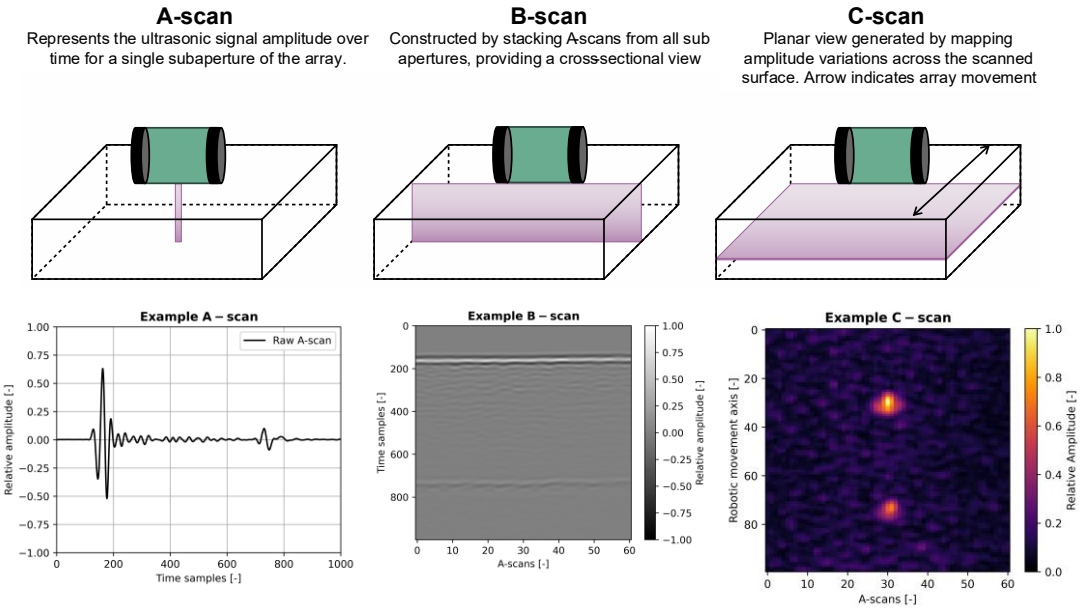


Figure 9 Demonstration of A-scan, B-scan, and C-scan ultrasonic views used throughout this thesis.

2.1.9 Ultrasonic A-scans

The A-scan representation displays the captured ultrasonic wave amplitudes in time. This format is typically used with single-element transducers, allowing an NDE operator to estimate the thickness of the inspected sample by measuring the peak-to-peak distance between the front and back wall indications. However, A-scans require expertise to interpret as they provide limited spatial information. Despite this limitation, A-scans are valuable for detecting typical CFRP composite defects (explained in Section 2.2), such as delaminations and impact damages and determining their depth due to the high temporal resolution [71]. A representation of a pristine and defective A-scan, captured from a CFRP sample containing a defect, is shown in Figure 10.

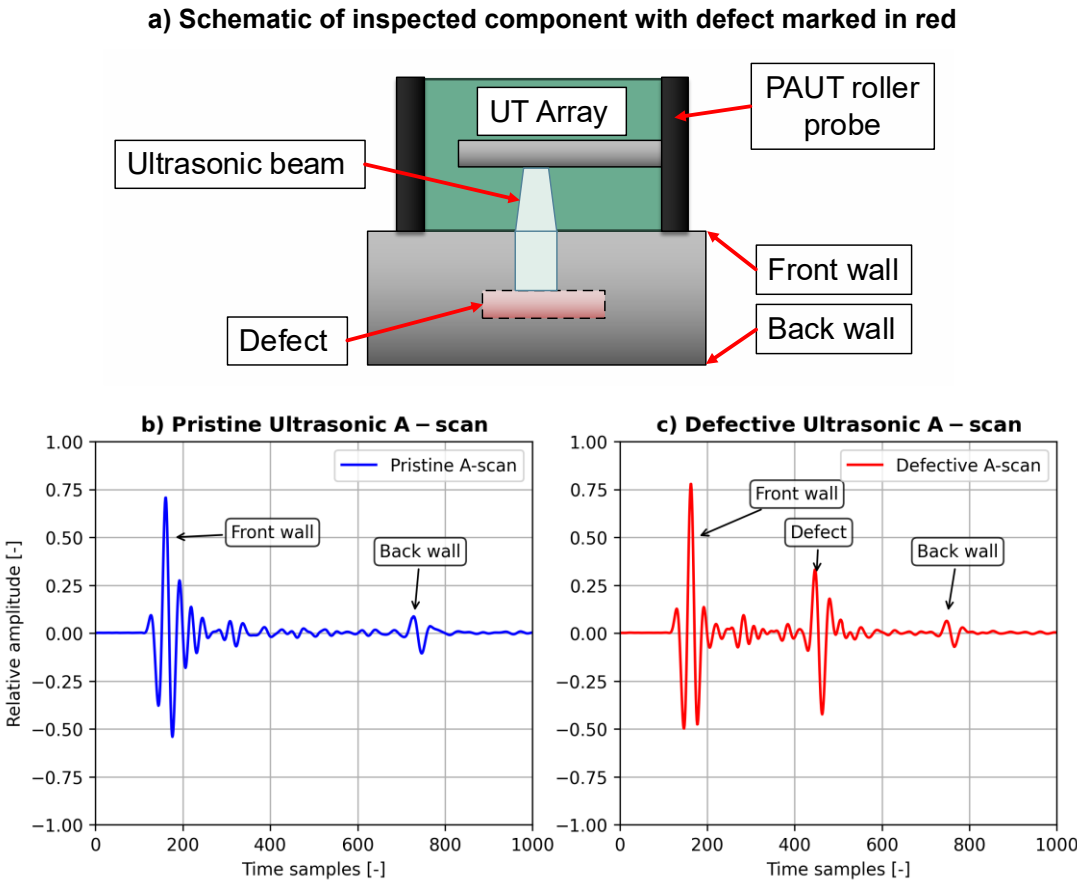


Figure 10 a) Schematic of a CFRP component containing a defect (marked with red); b) Example of a pristine ultrasonic A-scan, with labels indicating the front and back walls; and c) example of a defective ultrasonic A-scan, showing the front and back wall labels along with a labelled defect.

2.1.10 Signal Time Gating

Raw ultrasonic signals are typically acquired in the form of Radio Frequency (RF) signals, which contain both amplitude and phase information. RF data consists of both positive and negative values, representing the oscillatory nature of the ultrasonic wave. Other forms of signal representation include rectified data, where the absolute values of the signal are taken to remove negative amplitudes. Additionally, enveloped signals, often obtained using the Hilbert transform, provide a smooth representation of the signal envelope but sacrifice phase information (detailed in Section 2.1.11)

The application of several signal processing methods is used to aid the interpretation of captured ultrasonic data. Time gating involves applying a windowing technique to the time-series data to isolate areas of interest. Geometrical features at interfaces between different materials, like front and back walls, exhibit higher amplitudes compared to other structures within the sample. These indications, if not excluded, can

mask weaker ultrasonic signals of interest, such as discontinuities and defects. Therefore, windowing focuses on eliminating the contribution of such features in scan images to enhance the visibility of relevant signals and is a crucial step in the preparation of C-scan views, explained in Section 2.1.13. However, as windowing is applied manually, it introduces a degree of subjectivity and errors in the selection of gating boundaries may unintentionally exclude important features or retain unwanted signals, which can affect the data analysis steps. An example of signal gating is illustrated in Figure 11.

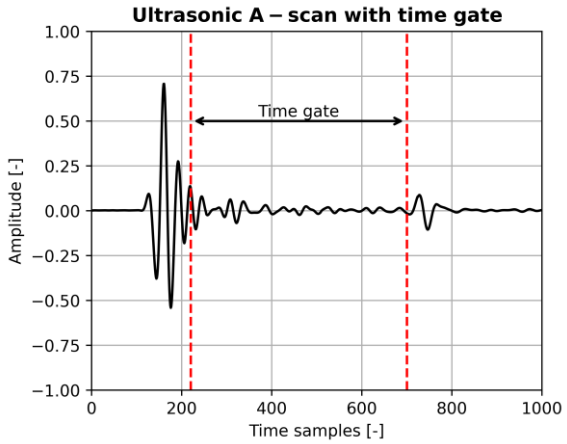


Figure 11 An example of an ultrasonic A-scan with time gating.

2.1.11 Hilbert Transform

Hilbert transform is a mathematical method used to envelop signals, commonly applied in time-series data analysis to examine instantaneous amplitude responses [72]. This transformation generates a complex signal which consists of phase shifting every Fourier transform component by 90°. Specifically in UT, applying the Hilbert transform to each A-scan signal and visualising the real component results in an enveloped signal that effectively represents only instantaneous amplitude response while eliminating phase information. This technique is particularly useful in scenarios where multiple reflections occur within a short time segment [73]. The Hilbert transform is crucial when generating C-scan views of the ultrasonic data, as discussed further in section 2.1.13. An example of RF and Hilbert-processed signal is presented in the Figure 12.

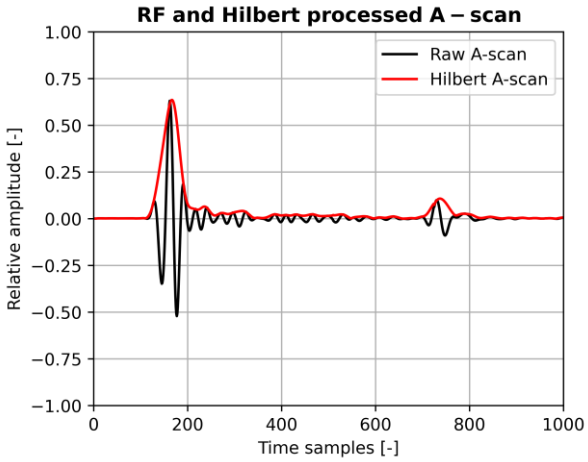


Figure 12 Comparison of RF and Hilbert-processed A-scans.

2.1.12 Ultrasonic B-scans

When using PAUT systems, each transducer (in single-element sub-aperture scanning) or group of elements (in multi-element sub-aperture scanning) generates an A-scan. By stacking the A-scans according to the location of the of the elements used for acquisition and representing the amplitudes on a colour scale, a B-scan representation can be produced. A B-scan provides a cross-sectional view of the scanned sample, allowing both the localisation of features within the active aperture of the array and the depth estimation of potential defects. B-scans are widely used in NDE and biomedical applications [67], [74]. However, B-scans lack spatial contextual information along the axis of PAUT array movement (i.e. perpendicular to the transducer's elevation axis), which can lead to difficulties when sizing defects. Furthermore, it is often impractical to examine all individual B-scans within the scans of large components, as this would be time-inefficient. An example of both pristine and defective B-scans, captured from a CFRP sample containing a defect, is presented in Figure 13.

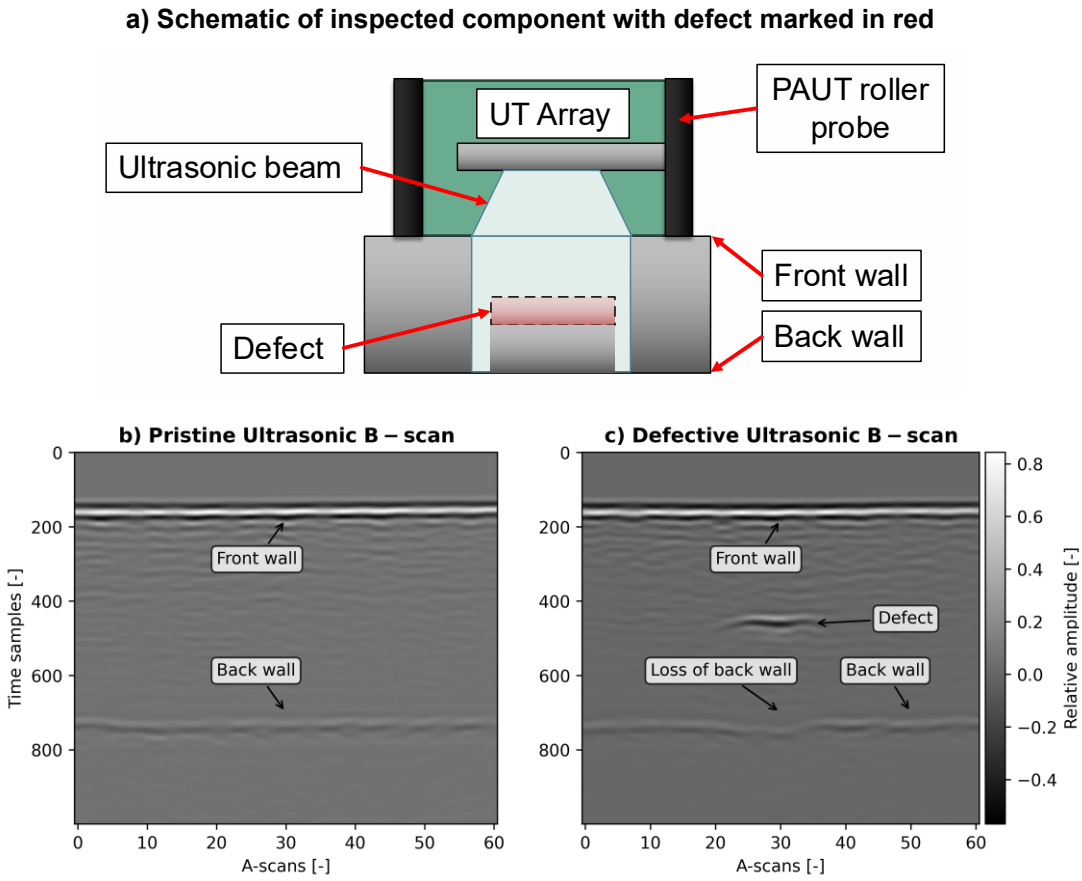


Figure 13 a) Schematic of a CFRP component containing a defect (marked with red); b) A pristine ultrasonic B-scan with labels marking the front and back walls; and c) A defective ultrasonic B-scan with labels marking the front and back walls, alongside additional annotations highlighting the defect and the associated loss of the back wall signal due to the defect.

2.1.13 Ultrasonic C-scans

Combining PAUT inspection with the mechanical movement of the array enables the capture of a series of B-scans at different positions along the component. These B-scans are stacked based on their positional encoding to form a three-dimensional (3D) volumetric scan of the component. The dimensions of the 3D scan include the robotic movement axis (defining the positions of the B-scans), the time axis (representing the length of each individual A-scan), and the array axis (representing the number of A-scans captured by the array).

A C-scan is a format that represents a two-dimensional (2D) slice parallel to the sample surface of such 3D volumetric data. It is created by applying a time gate to the A-scan signals and plotting the amplitudes versus the 2D scanning positions in form of a heat/intensity map. Hilbert transform is crucial step for C-scan generation as RF data contains a mix of positive and negative peaks that can compromise the visualisation

quality of C-scans by potentially missing maximum amplitudes. Therefore, applying the Hilbert transform ensures that C-scans accurately depict the peak amplitudes of interest without the interference of phase information. Lastly, signal gating is a critical step because it determines which portion of the A-scan data contributes to the C-scan image. If the gate is too wide or misaligned, it may include unwanted reflections or miss important signals from internal features. For example, front wall reflections typically do not contribute to defect analysis and, if included, would appear as bright spots without providing any insight into the internal structure. On the other hand, back wall reflections may or may not be relevant depending on the inspection objective (they are sometimes intentionally excluded to focus on internal features or included when monitoring full material thickness).

It is also possible to plot the time corresponding to high amplitude features versus the 2D scan position data to create ToF C-scans. C-scans preserve spatial information in three dimensions, therefore making them easier to interpret. By adjusting the time gating parameters, it is possible to focus on specific depths such as the internal structure of the material or the back wall indication [75], [76]. Representations of both amplitude and TOF C-scan are shown in Figure 14.

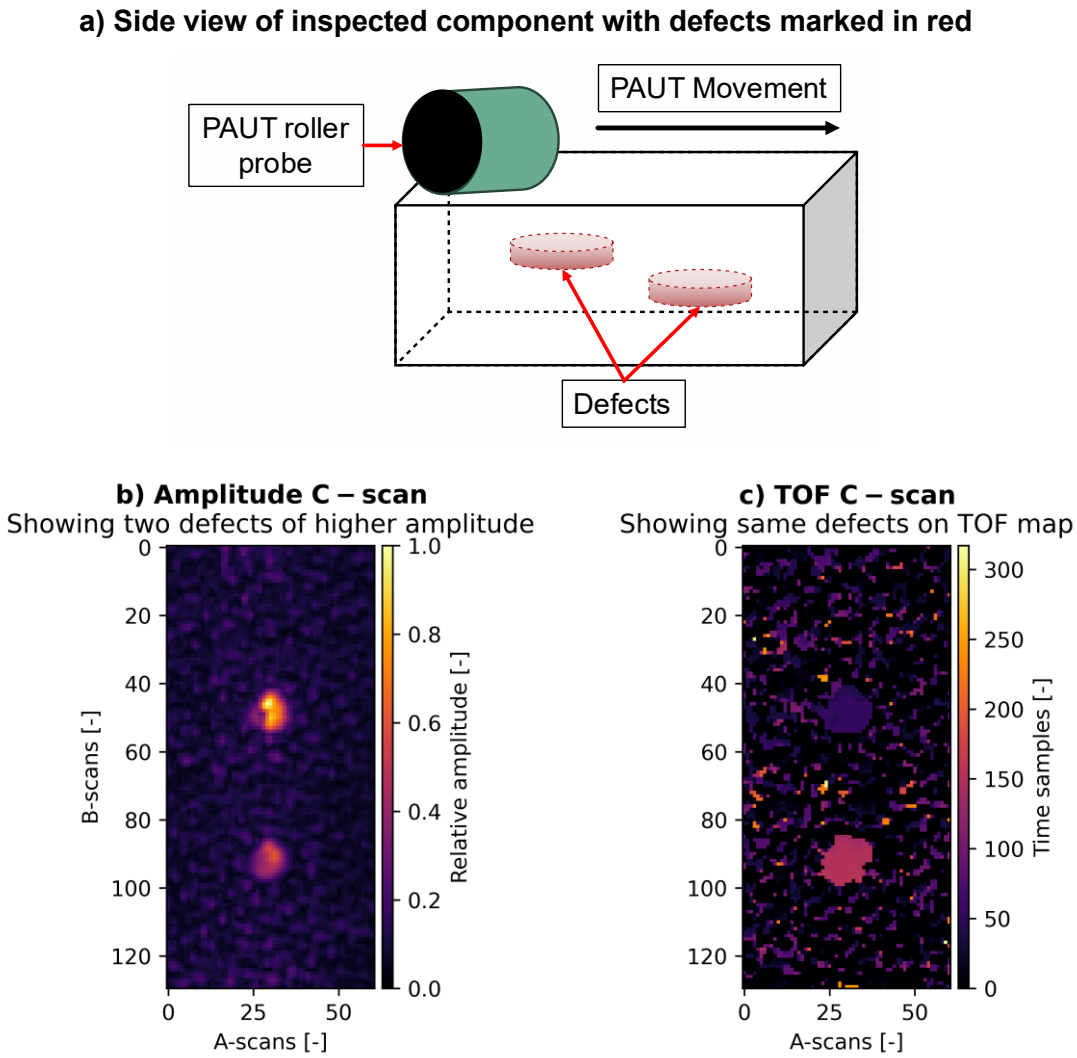


Figure 14 a) Side view schematic of a CFRP component containing two scanned defects at different depths; b) An excerpt of an amplitude C-scan showing two defects of higher amplitude; and c) An excerpt of a TOF C-scan showing the same two defects.

2.1.14 Time Varied Gain

Gain refers to the amplification of a signal to enhance its strength relative to the noise floor. Hardware gain is applied during data acquisition using analogue amplification stages, directly affecting the raw signal before digitisation, with the goal of ensuring the captured signal is strong enough. In contrast, software gain is implemented post-acquisition through digital processing. While software gain allows for flexible adjustments to the captured data, it cannot recover details that may be lost if improper hardware gain is used.

To compensate for attenuation effects, Time Varied Gain (TVG) is often used as an additional signal processing step. TVG progressively increases the gain as the ultrasonic wave travels further, compensating for the signal loss that occurs with longer propagation paths. TVG applied via the controller compensates for attenuation in real-time during the scan, while post-processing TVG allows for flexible adjustments after data acquisition. The former provides immediate correction, while the latter enables more flexible adjustments, but cannot address issues arising from inadequate gain setting during acquisition, such as insufficient signal strength at the initial capture.

In composite materials, especially for thinner samples such as those studied in this work, both TVG during capture and post-processing TVG are valid options. Since the ultrasonic pulses are relatively short and do not suffer from significant temporal broadening, the degradation of signal-to-noise ratio (SNR) commonly associated with longer pulses in TVG is minimised [77]. In practice, applying TVG during capture generally offers better SNR, but this approach is best suited to stable inspection setups scanning components of consistent geometry and material properties, where the TVG parameters can be calibrated and remain constant. When geometry or material properties vary, the TVG settings during capture may require frequent adjustments and recalibration, which can be time-consuming and impractical. In such cases, post-processing TVG provides valuable flexibility to adapt gain compensation after data acquisition without the need for repeated hardware recalibration.

Figure 15 illustrates a linear ramp TVG is applied to the B-scan from Figure 13. The shape and parameters of TVG are often determined experimentally by measuring attenuation in dB/mm and compensating for the corresponding signal drop to ensure consistent amplitude. A different approach would be the implementation of Time Compensated Gain (TCG) by using standard reference samples manufactured from the same material as the test object with a series of side-drilled/flat bottom holes at various depths to establish a TCG curve that yields consistent amplitude indications from all the holes; however, this approach was not pursued in this thesis.

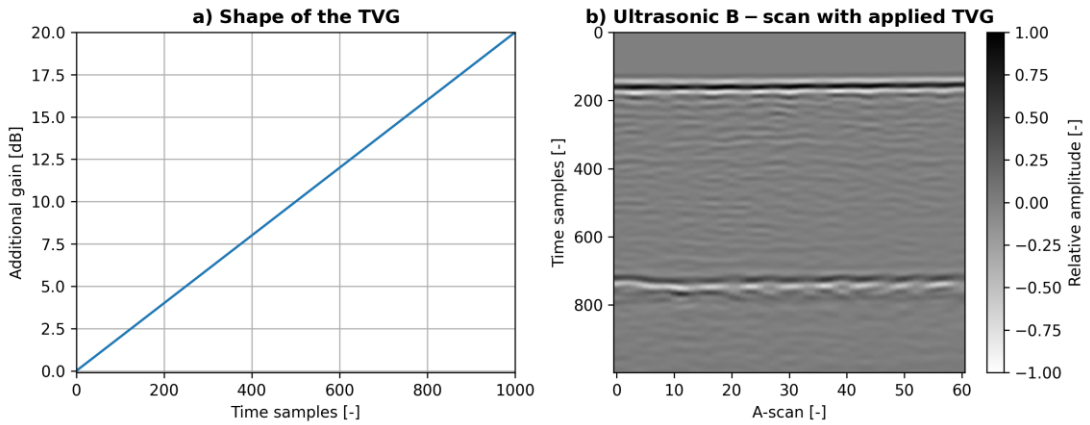


Figure 15 a) The shape of linear ramp TVG; and b) the effect it has on the B-scan representation.

2.2 Composite Materials, Ultrasonic Inspection Procedures in the Aerospace Industry, and Common Defects

2.2.1 Composite Manufacturing and Inspection in the Aerospace Industry

CFRPs are manufactured through a complex process of layering carbon fibre sheets and using thermoset polymer resin as a matrix material to bond them together. Overall, the manufacturing process can be divided into prefabrication and curing processes [78]. In the prefabrication stage, resin can be incorporated into the fibre material using either resin transfer moulding or the prepreg method. In resin transfer mould, dry fibre preforms are placed inside a mould and are infused with resin. Alternatively, in the prepreg method, fibres are injected with resin before being laid up, either manually or through automated layup technologies. While hand layup processes are still widely used due to the low initial equipment expenditures, they also remain necessary for complex layup scenarios that are challenging to automate. However, these processes are labour-intensive and demonstrate a higher chance of material variability [79], [80].

Following prefabrication, the most common method for curing CFRPs is autoclave curing, which applies both high pressure and temperature to solidify the CFRP into its final form. However, this method involves high energy expenditure and upfront costs, particularly larger autoclaves used for large components often used in aerospace manufacturing. In contrast, out-of-autoclave processes offer a more cost-effective alternative by using lower pressure levels, reducing the energy required for curing. Additionally, out-of-autoclave method utilises alternative heating mechanisms, such

as ovens or heat blankets [81], which significantly reduce operational costs while still achieving high material quality [82].

The layered construction of CFRPs, with varying fibre orientations, enhances mechanical properties in multiple directions. However, it also introduces anisotropy, which impacts materials behaviour under different loading conditions. During the manufacturing process, defects in various forms, such as porosity, fibre misalignment, and delaminations, may emerge, posing risks to safety and increasing the likelihood of critical material failure [42], [83]. If left unaddressed, these flaws can significantly influence material properties, such as strength [84], and lead to material failure when exposed to cyclic stresses during service [83].

PAUT has emerged as the preferred NDE modality in the aerospace industry due to its flexibility, safety, ease of integration with robotic setups, and the capability to detect various types of critical defects [85], [86], [87]. While the central operating frequency for UT typically ranges between 20 kHz and 25 MHz, frequencies in the range of 1 to 5 MHz are most commonly used for the inspection of CFRPs [37]. This range offers an effective trade-off between penetration depth and spatial resolution. Lower frequencies (e.g., 1 to 2.25 MHz) are suitable for inspecting thicker components due to better penetration, while 5 MHz probes are generally used for standard aerospace applications. Although higher frequency probes can be employed, there is a risk of resonance occurring. While resonance is intentionally used in some ultrasonic NDE applications [88], its effects in CFRPs can be problematic, as operating frequencies around 12 MHz increase reflections between ply interfaces, leading to greater masking of useful ultrasonic responses [89], [90]. Furthermore, higher frequencies result in reduced penetration depth, making them unsuitable for scanning thicker material samples, and increased attenuation, which is directly proportional to the operating frequency.

As CFRPs are used in safety-critical components such as wing covers and aircraft fuselages, extensive post-manufacturing NDE inspection is conducted [87]. However, the anisotropic nature of CFRPs poses challenges due to the complex scattering and high attenuation, resulting in a low signal-to-noise ratio in the captured data, which reduces the probability of detection [91], [92]. Additionally, the ultrasonic wave

velocity in CFRPs is angle-dependent, varying with fibre orientation, complicating accurate ToF measurements. Combined with multiple stacked layers and different types of thermoset polymers, this makes multi-layer wave refraction calculations for advanced imaging techniques such as TFM computationally demanding. Lastly, the complex geometries and varying thicknesses of CFRP components further complicate inspection.

As a result, electronic scanning methods such as linear scans remain widely accepted in the industry, providing sufficient performance for practical NDE applications. Regardless of the imaging approach used, NDE operators rely on analysing multiple ultrasonic views, primarily B-scans and C-scans, simultaneously to evaluate scanned components. The overall NDE workflow for aerospace industry is further discussed in Chapter 6. However, it's important to note that not one ultrasonic view is suitable for detecting all types of defects. Figure 16 illustrates the applicability of different ultrasonic views for different defect types.

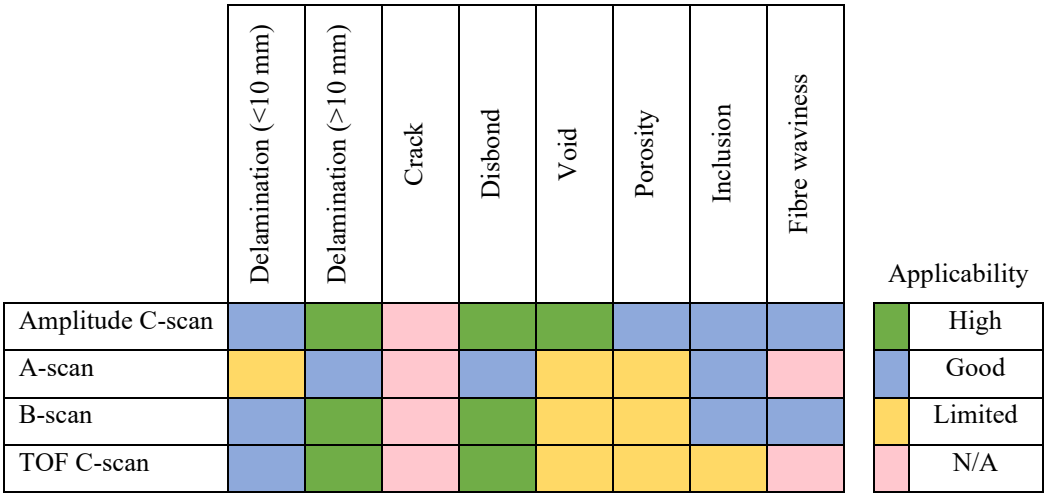


Figure 16 Applicability of different UT modalities for different defect types. Green indicates high applicability, blue indicates good applicability, yellow limited applicability, and pink no applicability. (adapted from [71])

Advancements in PAUT have been paralleled by significant improvements in the field of robotics, such as advancements in robotic path planning and the use of modern robotic manipulators for sensor delivery, thus enabling automated scanning [19], [58], [93], [94], [95]. The integration of PAUT and industrial manipulators has greatly improved the inspection processes of large and complex components, enabling high

precision, reduced scan times, and overall enhancing the NDE process in terms of reliability and safety for the inspection of high-value aerospace components [96].

2.2.2 Delaminations

Delaminations, one of the most common post-manufacturing defects [97], refer to the detachment of individual layers of carbon fibre sheets, creating disbond between the layers [98], [99]. Delaminations compromise the structural integrity of components by reducing their load-bearing capacity, particularly in applications where high bending, or shear forces are present as CFRPs lack strength in the direction perpendicular to the layered sheets [100].

Delaminations can originate from various sources and are broadly categorised into manufacturing and in-service defects. Manufacturing induced delaminations often occur due to improper curing, residual stresses from differential thermal expansion between fibre and matrix materials, trapped air, or moisture absorption, all of which lead to weak interfacial bonding between composite layers [101], [102]. Additional factors such as poor resin flow during vacuum infusion or inconsistencies in layup procedures can also contribute to delamination formation. In-service delaminations are predominantly formed through impact damage and extensive thermal and cyclic loading, where repeated expansion and contraction introduce stress, leading to the progressive separation of layers at weak interfaces [103].

Delaminations vary in size and can propagate under operational loads, which is often described by three failure modes: Mode I (Opening mode) where tensile stresses pull the layers apart; Mode II (Sliding mode) where shear forces cause slipping between the layers; and Mode III (Tearing mode) where out-of-plane shear and twisting forces cause propagation [104]. Delamination growth poses a significant risk to materials' structural integrity, as it directly reduces compressive, tensile, and shear strength [102], increasing the likelihood of catastrophic failure. Given the use of CFRP's in high-value safety-critical applications, early identification and characterisation of these defects is essential.

Due to their orientation parallel to the carbon sheet layers, delaminations are good reflectors of ultrasound normal beam inspection and are characterised by high

amplitude areas when visualised. They also obstruct the acoustic path to the back wall, causing a loss/reduction of back wall echo which can be observed in Figure 13. The thesis will focus on delamination defects in CFRPs. Delamination example caused by low velocity impact is shown in Figure 17.

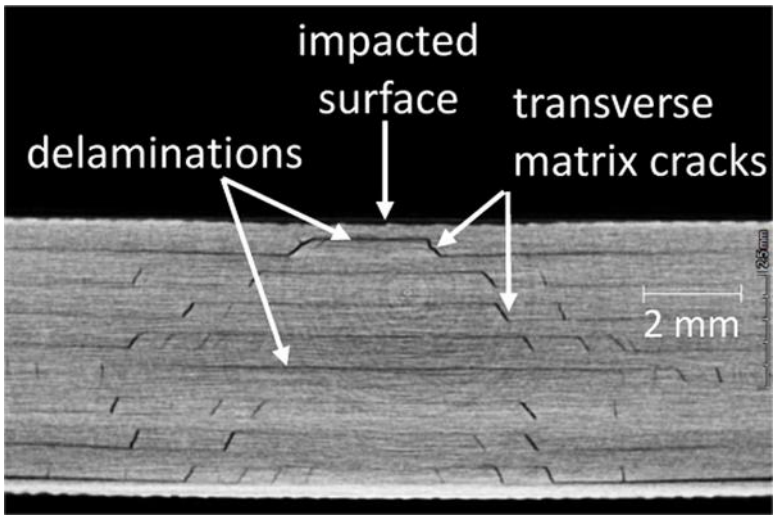


Figure 17 Example CFRP delamination caused by a low velocity impact. Reproduced without modification from [105]. CC BY-NC-ND 4.0 license (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en)

2.2.3 Voids and Porosities

Voids are trapped air or gas bubbles within CFRPs, whereas porosities refer to smaller, dispersed voids throughout the material. These defects primarily originate during manufacturing and can significantly impact the mechanical performance and long-term durability of CFRPs. Voids and porosities introduce localised material weaknesses that serve as stress concentrators, promoting material degradation under cyclic loading.

Their formation is often attributed to incomplete resin impregnation, suboptimal curing conditions, improper control of processing parameters (resin viscosity, pressure, and temperature), or external factors such as increased moisture levels or contamination [106], [107]. Their presence directly affects mechanical properties by reducing strength, fatigue life, and impact toughness [108], [109], [110], thereby increasing the risk of material failure. All CFRPs contain porosities, expressed as a percentage of the volume that must fall within an allowable range.

Voids can be identified using UT due to the high acoustic impedance mismatch between the air pockets and composite materials, resulting in an ultrasonic response comparable to delaminations. However, identifying porosities is more challenging due to their smaller size and weaker acoustic response. Porosities typically appear in clusters rather than as isolated defects, and when clustered, they can create a shadowing effect on the back wall echo. This shadowing, along with increased attenuation in the captured signals, can serve as an indicator of porosity presence. An example porosity in CFRPs is shown in Figure 18.

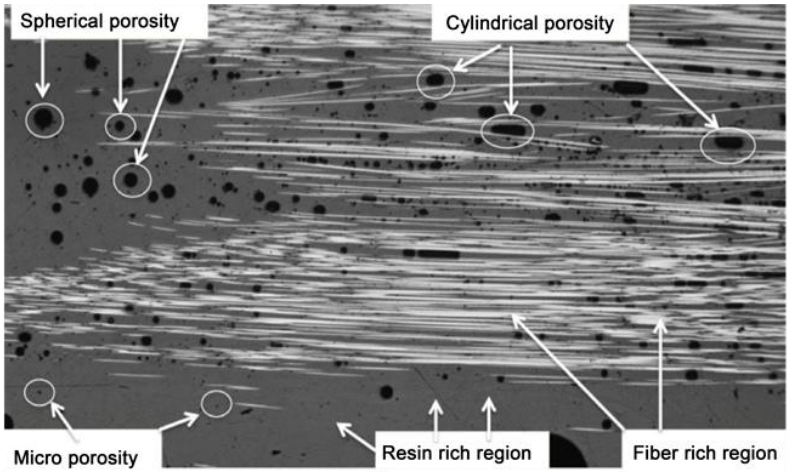


Figure 18 Example CFRP porosity. Reproduced without modification from [111]. CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/deed.en)

2.2.4 Inclusions

Inclusions refer to any foreign materials incorporated into the composite structure during the manufacturing process. These can include various particles or materials from the environment, used equipment, or even residual contaminants from processing steps, such as resin or carbon fibres [112]. Inclusions can vary in size, shape, and material composition, and their presence can significantly impact the mechanical properties of CFRPs. These inclusions act as stress concentration points, which can serve as initiation sites for more critical defects (such as delaminations) while also deteriorating the material's strength and fatigue resistance [112] [113]. The aerospace industry aims to minimise the occurrence of such inclusions by conducting manufacturing processes in controlled and clean environments.

Detection using UT can vary significantly; unintended inclusions, such as carbon fibres, may produce an acoustic response similar to the surrounding material, posing challenges for detection. However, inclusions with a different acoustic impedance are easier to detect.

2.2.5 Fibre Waviness

Fibre waviness in composite materials refers to deviations in the orientation of carbon fibres from their intended alignment within the matrix. This phenomenon can manifest as either in-plane waviness, where variations occur within the composite layer, or out-of-plane waviness, where variations occur perpendicular to the composite layer.

Waviness can occur due to various factors during the manufacturing process, such as the placement errors in automated or manual layup, uneven thermal gradients during curing, improper consolidation for complex geometry parts, fibre misalignment, or the presence of inclusions [114], [115]. Fibre waviness is considered a critical defect because it disrupts the intended fibre reinforcement structure, weakening the composite's ability to bear load efficiently, and decreasing static strength, stiffness, and fatigue resistance [116], [117].

Using UT, fibre waviness can be identified primarily from ultrasonic B-scans, where it may cause distortion of the incident waves. This distortion can mask other features present within the material, making detection and sizing of other features more challenging [90]. Additionally, the presence of waviness can affect the accuracy of thickness measurements. An example micrograph of fibre waviness in CFRPs is shown in Figure 19.

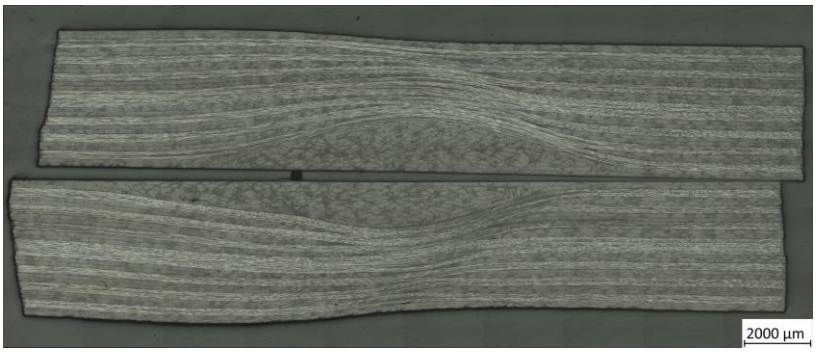


Figure 19 An example of fibre waviness in CFRP components (adapted from [118])

2.3 Artificial Intelligence

There exist numerous definitions of Artificial Intelligence. For the purposes of this thesis, the definition provided by the UK National Cyber Security Centre will be followed: "*Artificial Intelligence describes computer systems which can perform tasks usually requiring human intelligence*" [119]. Machine Learning (ML), a subset of AI, is defined as "*a method by which computers find patterns in data or solve problems automatically without being explicitly programmed*" [120]. Lastly, for Deep Learning (DL), a subcategory of ML, the following definition will be used the definition proposed by Yann LeCun et al. will be followed: "*Deep learning allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction*" [121].

Tasks tackled by ML can be described as classification tasks, where the aim is to categorise data in one of the predefined classes, or regression tasks where the aim is to predict some continuous variable. An illustration of the categorisation of AI is presented in Figure 20.

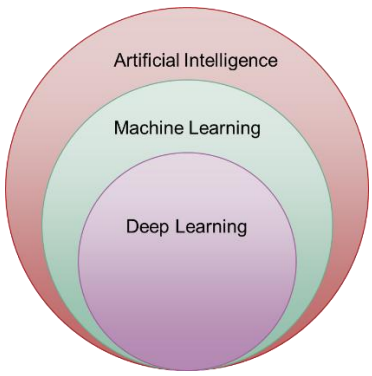


Figure 20 Illustrative representation of Artificial Intelligence (red), Machine Learning (green) and Deep Learning (purple)

The increase in the adoption of AI during the past decade can be attributed to several factors. Firstly, the creation of large readily available datasets for AI training has led to the development of more accurate and powerful models. Secondly, advancements in computing power, especially the availability of GPUs, have significantly reduced the time needed to train such models. Additionally, the availability of open-source

tools like PyTorch [122] and TensorFlow [123] has lowered the entry barrier for new research and applications, which paired with increased investment and interest from industry and academic institutions resulted in a massive influx of new publications.

One of the AI categorisations is based on the type of training, which includes supervised, unsupervised, self-supervised, and reinforcement learning approaches.

Supervised learning involves training models on labelled datasets, pairing each data point with ground truth. During training, both the input data and correct output predictions are presented to the model, to learn patterns and features that will enable the operation on new unseen data. This approach is widely used for classification and regression tasks, as well as for more complex applications like object detection, which combines the two to localise objects within images [121]. Unsupervised learning leverages unlabelled datasets, where no ground truth is provided during training, with the objective of identifying underlying patterns or distributions within the data. This approach is commonly used for tasks such as clustering, dimensionality reduction, and anomaly detection. Self-supervised learning lies between supervised and unsupervised learning, utilising automatically generated pseudo-labels directly derived from the data, which reduces the need for manually generated ground truth [124]. This approach has shown potential for tasks where large datasets are available, particularly in speech recognition [125], [126] and computer vision [127], [128]. Reinforcement learning [129] involves a model interacting with its environment to learn actions that maximise behaviours influenced by a reward function. During training, the model selects actions within the environment and receives feedback from a user-defined reward function. This feedback adjusts the model and influences its subsequent actions. Reinforcement learning is often applied in robotics [130], self-driving cars [131], and games [30], [31], where models use a trial-and-error approach to optimise their behaviour.

2.3.1 Basic Deep Learning Neural Network

The most basic form of deep learning (DL) networks is a neural network (NN) model consisting of an input layer, multiple hidden layers, and an output layer. In contrast, simpler NNs may have no hidden layers (single layer perceptrons) or just one hidden layer (shallow NNs). Each layer comprises interconnected nodes known as neurons or perceptrons, which were first introduced in the 1940s [132] and applied in the 1950s

[133]. In the example presented in Figure 21, the classification Neural Network (NN) consists of an input layer, two hidden layers, and an output layer.

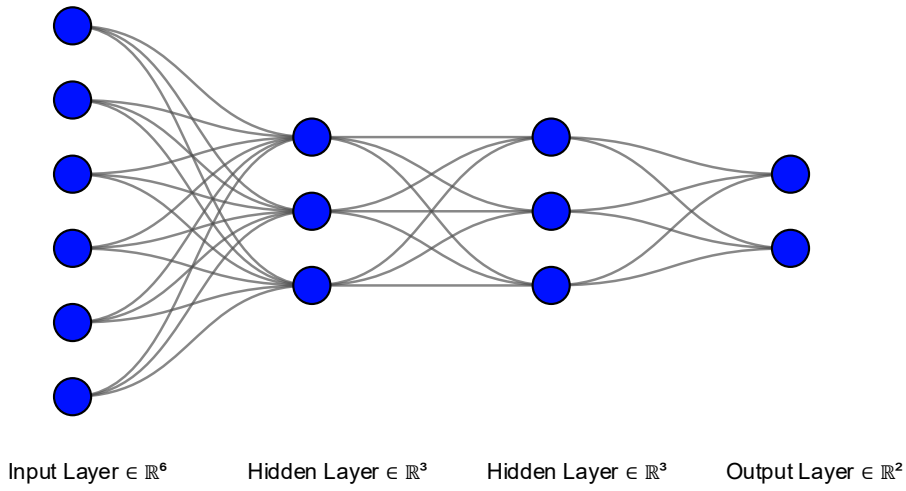


Figure 21 The basic architecture of a deep neural network. Blue represents the neurons, while black lines indicate the connections between them. The input, hidden, and output layers are labelled on the plot.

During the forward pass, data undergoes transformations that typically reduce its dimensions within hidden layers. The interactions between neurons involve three main components:

- **Weights**, which determine the strength of connections between neurons.
- **Biases**, which are added to the output of neurons to introduce flexibility.
- **Activation functions**, which introduce non-linearities, allowing the model to capture complex patterns in the data.

Mathematically, the output of the neuron can be represented with Eq.14:

$$Output = Activation\left(\sum_{i=1}^{n}(Input_i * Weight_i) + Bias\right) \qquad \text{Eq.14}$$

There are many activation functions, with the most commonly used being the Sigmoid, hyperbolic tangent (Tanh), and Rectified Linear Unit (ReLU). These can mathematically be presented with Eq.15, Eq.16, and Eq.17:

$$ReLU = f(x) = \max(0, x) \qquad \text{Eq.15}$$

$$Hyperbolic\ tangent = f(x) = \tanh(x) \quad \text{Eq.16}$$

$$Sigmoid = f(x) = \frac{1}{1 + e^{-x}} \quad \text{Eq.17}$$

The choice of activation functions is strongly influenced by the application and task requirements. The sigmoid activation function outputs values bounded between 0 and 1, making it commonly used in classification tasks. The tanh activation function, which outputs values between -1 and 1, was the preferred activation function in early ML research. These activation functions do not just shape individual neuron behaviour but also influence the types of functions the network can learn. Sigmoid and tanh encourage smooth and bounded transitions, which are useful learning gradual relationships but may suffer from vanishing gradients in the deeper networks. Today the most popular activation function, especially in DL is ReLU. ReLU was introduced in [134], and unlike previous functions is unbounded for the positive input values, meaning its output ranges from 0 to infinity. ReLU encourages sparse and piecewise linear activations, allowing the network to focus on more complex non-linear functions.

Several iterations of ReLU were introduced, like Gaussian Linear Unit (GeLU) and Exponential Linear Unit (ELU) [135], [136] which are unbounded on both positive and negative sides. However, unlike ReLU, ELU and GeLU add smoothness around zero, helping to maintain stable gradient flow. By stacking multiple layers of the described computations, the models gain the ability to tackle complex non-linear tasks. Discussed activation functions are illustrated in Figure 22.

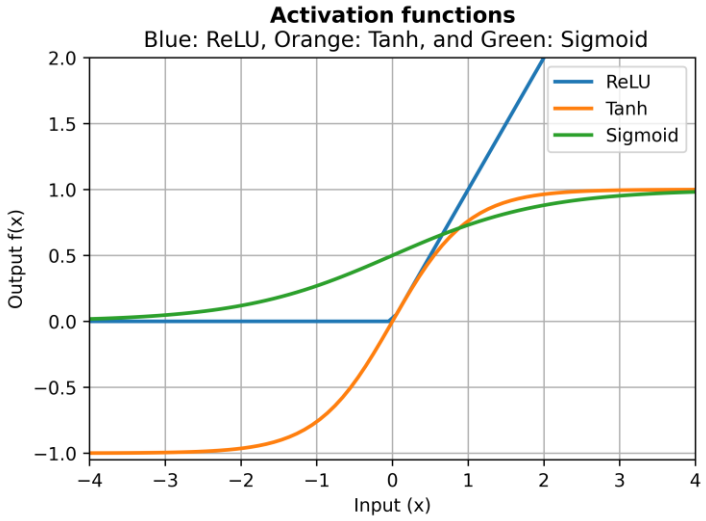


Figure 22 Illustration of ReLU, Tanh, and Sigmoid activation functions (blue, orange, and green)

The backward pass involves calculating the gradient of the loss function with respect to the network parameters using backpropagation. During this process, the gradients are propagated backward through the layers of the network, adjusting the weights and biases. However, in networks with many layers (deep networks), the gradients calculated during backpropagation can diminish exponentially, leading to what is known as the vanishing gradient problem. This phenomenon occurs when the gradients become very small, preventing the network from learning effectively in earlier layers, especially when certain activation functions (like sigmoid or tanh) are used.

When designing a model, trainable parameters (weights and biases) are initialised randomly, meaning the model does not contain any "knowledge" about how to perform a specific task. The process of adjusting these parameters iteratively is called training, and each iteration is commonly referred to as a training epoch. A training epoch refers to the number of times the entire training dataset is passed forward through the network. During the forward pass, data is processed, and in the backward pass, gradients are computed and propagated backwards. Model parameters are then updated based on these gradients.

Depending on the network task and output, a defined loss function between the expected and received output influences the updating of the model's parameters. The loss function serves as a measure of the network's performance and can take various forms. For regression tasks, the loss function is often represented as Mean Squared Error (MSE), while in classification tasks, cross-entropy loss is commonly used. However, there are numerous other types of loss functions such as mean absolute error, Huber loss, and hinge loss, among others. To provide an illustrative understanding of how a loss function behaves during training, Figure 23 illustrates a simple 3D loss landscape, where the vertical axis represents the loss value for different combinations of model parameters. In such a landscape, optimisation algorithms attempt to find the lowest point (i.e., minimal loss), which corresponds to the best model parameters.

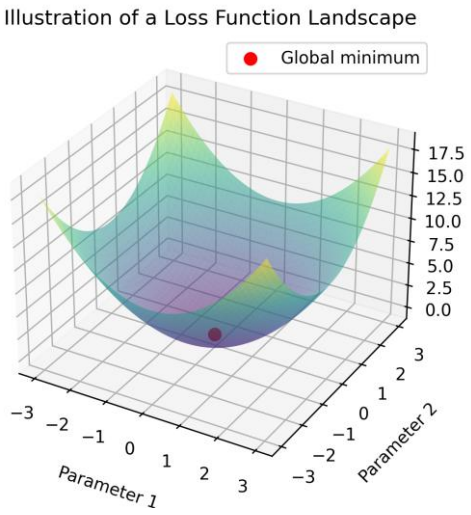


Figure 23 Illustration of an example loss landscape, with global minimum marked in red

The optimisation process is performed by an optimiser, which iteratively makes incremental changes to the network parameters to minimise the defined loss function $f(\Theta)$, where Θ represents trainable model parameters. One of the foundational optimisation algorithms is gradient descent, which updates the parameters in the opposite direction of the gradient of the calculated loss, and is show with Eq.18:

$$\theta = \theta - \eta \times \nabla_\theta f(\theta) \qquad \text{Eq.18}$$

where η is the learning rate, and $\nabla_\Theta f(\Theta)$ is the gradient of the loss with respect to the model parameters [137]. In Stochastic Gradient Descent (SGD) [138], this update is performed using a mini-batch of the training data rather than entire dataset, which introduces noise into the process but allows for scalability to large datasets and helps escape shallow local minima by encouraging wider exploration of the parameter space.

However, in basic SGD, selecting an appropriate learning rate can be challenging. A low learning rate may lead to slow convergence, while a high learning rate can cause unstable oscillations in parameter updates or divergence. This sensitivity often necessitates the use of learning rate schedulers to adapt the learning rate during training (though this introduces another layer of complexity, as these schedulers must be defined in advance). Lastly, SGD can struggle with minimising highly non-convex loss functions, which are common in DL tasks.

To address these challenges, momentum-based methods were developed, which accumulate a moving average of past gradients to smooth the updates and accelerate convergence in relevant directions. This process can be described with Eq.19 and Eq.20:

$$v_t = \gamma v_{t-1} + \eta \times \nabla_\theta f(\theta) \quad \text{Eq.19}$$

$$\theta = \theta - v_t \quad \text{Eq.20}$$

where γ is momentum term and v_t is the velocity vector that builds over time. Modern iterations of this algorithm such as adaptive gradient algorithm [139] and Adaptive Moment Estimation (ADAM) [140] stand out as popular choices. The model's trainable parameters updating is performed in reverse order, starting from the final network layer, and moving backwards through the network.

Hyperparameters are parameters that control the model architecture and training process. While it would be impractical to list all possible hyperparameters, some common examples include the number of neurons in each layer (refer to Figure 21), the number of training iterations, the choice of activation functions (refer to Figure 22), the selection of optimisation algorithms, and the number of hidden layers. One of the most important hyperparameters is the learning rate, which determines the magnitude of updates to the trainable parameters. Small learning rates lead to minor network updates, increasing the risk of getting stuck in the local minima of the loss function without reaching the global minimum, leading to suboptimal model parameters. On the other hand, large learning rates can destabilise the training process by overshooting optimal weight and bias values with new updates. To strike the right balance, some optimisers use adaptive learning rates, while learning rate schedulers can be employed to make predefined changes in the learning rate after certain training iterations.

The batch size hyperparameter determines the number of data samples processed in each iteration of the training loop. In training, the model iterates over the entire dataset during each training epoch. However, processing the entire dataset at once is often impractical due to memory constraints, while processing one data point at a time would be inefficient, especially for large datasets. To address this, the dataset is divided into

smaller batches and is processed concurrently during training to utilise parallel computing capabilities offered by modern GPUs. Choosing an appropriate batch size involves trade-offs. Larger batch sizes can lead to faster computation times per epoch but require more memory, potentially limiting the size of the model that can be trained. On the other hand, smaller batch sizes consume less memory per batch but may increase the variance in gradient estimates due to smaller sample sizes, which can impact the stability of training.

The model's performance during training is typically monitored by tracking the trend of loss function across training epochs. Ideally, this evaluation is performed on a separate dataset originating from the same domain and distribution as the training dataset, called the validation dataset. As the validation dataset is not directly used during training, it plays a crucial role in preventing overfitting, a phenomenon where the model fails to generalise well to new inputs. To further mitigate the risks of overfitting, additional regularisation techniques during training can be used, such as weight decay [141], batch normalisation [142], or early stopping. Validation data can also be used to guide model architecture selection and hyperparameter tuning.

Overall, choosing the right combination of model architecture and hyperparameters depends on the specific task and application and presents a challenging task. This thesis primarily focuses on the use of Convolutional Neural Networks (CNN), due to their proven effectiveness in processing spatially structured data such as ultrasonic C-scan and B-scan images. However, the basic principles, training methods, and reasoning behind neural networks introduced in this section remain consistent across various architectures. A training loop described in this section is illustrated in Figure 24.

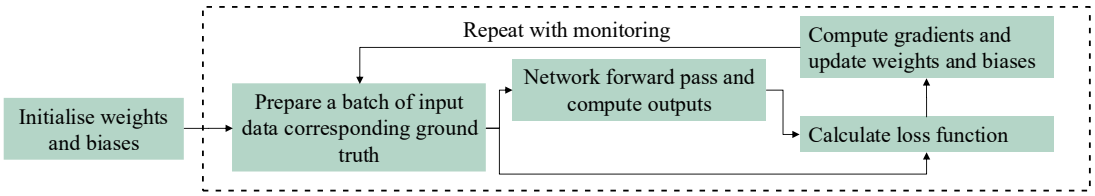


Figure 24 Flowchart of a supervised training loop.

2.3.2 Convolutional Neural Networks

Convolution layers offer an alternative to connected neurons and are building blocks of CNNs. CNN revolutionised the field of computer vision, starting with the LeNet network for handwritten digit classification in the 1990s [143] and later with AlexNet [27], the seminal model that started the current wave of AI research. Following the introduction of AlexNet, many models improved upon the CNN structure aimed towards vision tasks on large datasets. Notable examples include the Visual Geometry Group (VGG) network, which advanced the field by incorporating a significantly deeper architecture [144]; Residual Networks (ResNet) where skip connections effectively reduced the effect of vanishing gradients, thus enabling construction and training models with many hidden layers [28]; and, more recently, the integration of transformer structures with CNNs [145].

CNNs are often used in classification, object detection, and segmentation tasks. Object detection involves both localisation and classification of objects within an image by generating bounding boxes around them. Image segmentation is a more complex task, as it assigns a class to each pixel in the input image, producing a detailed and precise segmentation map. However, convolutional layers are not limited to image data - they can also be applied to time-series data and other data formats. Despite their advantages, CNNs are computationally intensive during training and require large datasets to achieve optimal performance.

CNNs process inputs in a grid-like format using filters (kernels) to create feature maps. Convolutional layers are often followed by activation functions to introduce non-linearity into the overall model. This dimensionality reduction extracts only the most relevant features for specific tasks and lowers the overall number of trainable parameters. Furthermore, pooling layers, such as max pooling or average pooling, summarise regions of the feature maps by retaining essential information while discarding less relevant details. This reduces the spatial dimensions of feature maps, decreasing computational complexity and improving model efficiency. An example of a convolution operation is illustrated in Figure 25, pooling operations in Figure 26, while the basic CNN architecture is presented in Figure 27.

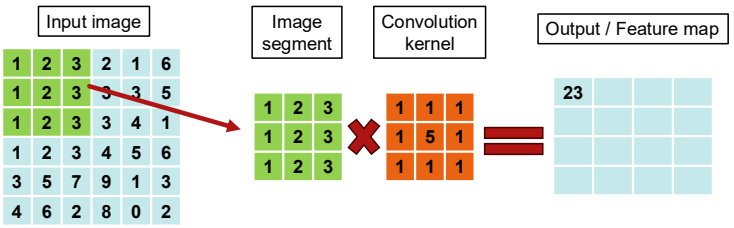


Figure 25 Illustration of the convolution operation in a CNN. The input image is convolved with a kernel to produce a feature map, where each value represents the sum of element-wise multiplications between the kernel and image segments.

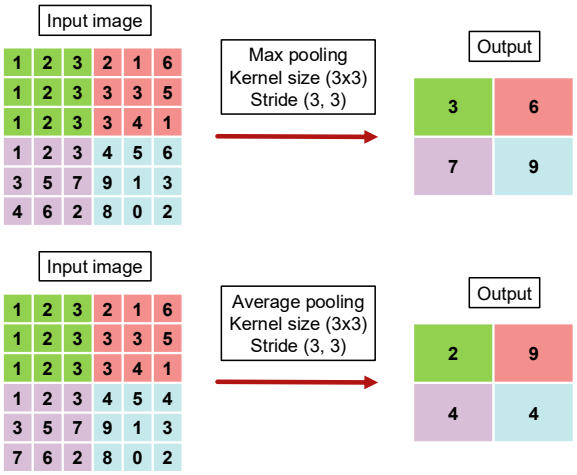


Figure 26 Illustration of how pooling operation in a CNN. Max pooling retains maximum value observed within a specified kernel, while average pooling provides an average value in the observed kernel.

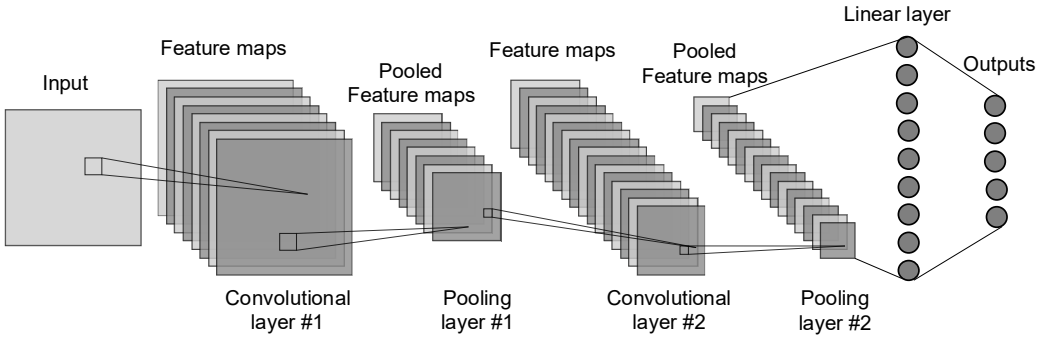


Figure 27 Basic convolutional neural network architecture, containing max pooling and linear layers towards the end.

2.4 Machine Learning in Ultrasonic Testing: A-scans

While a broad range of ML techniques have been applied to UT data, certain trends and gaps are evident across the literature. Research has predominantly focused on A-scan data, with limited exploration of B-scans and C-scans due to their acquisition complexity. There is also a noticeable lack of studies combining multiple scan types. Moreover, data scarcity remains a recurring challenge, often addressed using synthetic data and transfer learning. These themes are discussed in more detail throughout sections 2.4 - 2.8 and are revisited in the concluding summary 2.9.

This section serves as an overview of recent studies that utilise A-scan data as input for ML methods. In [146], [147], the authors focused on UT A-scans of welds, introducing an automated defect classification method based on a PNN. This approach utilised handcrafted features from time and frequency domains to determine the class of individual A-scans. The developed model demonstrated good results in detecting cracks, porosities, and slags, with an advantage in terms of improved computational efficiency.

In [148], the authors investigated the application of NNs in assessing fatigue life and tensile strength of welds produced through resistance spot welding. The network input was the number and amplitudes of ultrasound echoes, while the fatigue life and tensile strength were set as outputs. Approximately 200 weld samples were created, with 60 of them paired based on matching UT A-scans. This pairing strategy proved essential due to the destructive nature of both fatigue and tensile strength tests, making it impossible to obtain both values from the same sample. By pairing samples, researchers were able to conduct both tests on identical specimens, creating a model capable of predicting the physical properties of welds from UT A-scan data.

In the subsequent study [149], UT was used to evaluate the quality of resistance spot welds. A mathematical model of wave propagation in such joints was developed, followed by Finite Element Analysis (FEA) simulation. A 14% error in theoretical calculations, attributed to wave attenuation, was corrected via trial-and-error before being used as input for a NN. The trained model, tested on both simulated and real data, classified welds as acceptable or undersized, achieving over 90% accuracy across 12 test samples.

The authors of [150] developed a CNN capable of accurately recognising crack dimensions, location, and orientation in load-bearing structures. Training data was simulated using FEA software, incorporating virtual models with artificial defects. Modifications to the dimensions, depth, and location of the voids resulted in a dataset comprising 900 samples. Initially, the CNN performed poorly, which the authors attributed to the small training dataset size. However, after augmenting the dataset using the parameter-space augmentation (where new data points were simulated by randomly sampling within the defined parameter ranges), the CNN's performance improved significantly. This augmentation strategy complemented the 900 structured samples with an additional 300 randomly generated cases, helping to mimic the variability expected in real experimental scans. To validate the model, 21 UT experiments were conducted on 3D-printed samples, achieving an accuracy of over 90%.

In [151], the application of various ML models on NDE data collected from butt-fused joints in polyethene gas pipes was investigated. The authors created 20 pristine and 30 flawed sample welding joints, with 100 A-scans collected per joint, resulting in a total of 5000 A-scans serving as inputs. A comparative analysis of various ML and DL models showed that CNNs delivered the best overall performance for classification tasks.

Researchers in [152] employed different feature extraction methods to generate inputs for an NN classifier. A total of 400 A-scan signals were obtained from carbon steel plate test samples containing welding defects. The authors reported poor performance when raw UT data was being used, which paired with a substantial need for high computing power, underscored the need for a dimensionality reduction. To address this challenge, they applied the discrete Fourier, discrete cosine, and discrete wavelet transforms, resulting in substantial improvements in training and model performance.

In [153], the authors compared weld flaw classification performance between NN and DL models using 720 ultrasound signals from two ultrasonic transducers with varying operating frequencies. They specifically investigated the impact of the applied drop-out regularisation method to prevent overfitting. DL demonstrated higher overall accuracy.

In their follow-up study [154], the authors expanded their work by comparing the performance of CNN and NN models in noisy environments. They augmented the dataset from their earlier study by introducing time shifts and varying levels of Gaussian white noise, resulting in five datasets with different noise levels. The findings showed that while NN performed well under low noise conditions, its performance declined at higher noise levels. In contrast, CNN consistently outperformed NN, particularly in high-noise classification tasks. The study suggests that CNN is a more robust method for industrial applications, where UT data is potentially subject to varying levels of noise.

In their last study [155], the authors extended their research by integrating a denoising AE into their workflow. The denoising AE demonstrated significant benefits by enhancing classification performance across all noise levels while minimising signal degradation. The authors concluded that incorporating the AE improved CNN performance, making it more viable for real-world applications.

In [156] the authors explored the classification of thermal ageing defects in cast austenitic stainless steel, commonly used in nuclear power plants. Their study involved recording and processing 125 ultrasonic A-scans used to train three ML models. All models achieved high accuracy when processing data with high peak-to-peak amplitudes. However, KNNs and SVMs exhibited decreased performance with data featuring lower peak-to-peak values, whereas NN maintained consistently high accuracy across varying data conditions. This study underscores the NN approach as a superior choice for this application compared to older ML approaches.

Authors of [157] focused their research on laser UT for selective laser melting of steel in additive manufacturing. They compared a feature extraction using principal component analysis with directly handcrafted features from A-scans in training of a NN. The study found that the principal component analysis feature extraction approach outperformed handcrafted features, underscoring the importance of effective feature extraction for improving classification tasks.

In [158] researchers utilised 6000 UT A-scans from CFRPs, including both flawed and pristine samples. The study aimed to compare various classifiers and hyperparameters to identify the optimal ML algorithm. CNN emerged as the best classifier, surpassing

NN and SVM in performance. Furthermore, the authors acknowledged the potential noise and inaccuracies associated with the direct generation of C-scans from unprocessed A-scans. To mitigate these issues, they proposed reconstructing C-scans using already classified A-scan data.

In the study detailed in [159], the authors employed a DBSCAN algorithm to classify defects in pressure tubes used in nuclear reactors. This unsupervised method clusters statistical signal features without requiring predefined class labels, demonstrating its potential for defect detection in UT signals.

Researchers in [160] developed a CNN integrated with a gate recurrent unit (A type of recurrent neural network that captures sequential patterns in data by retaining important past information and filtering out less relevant details [173]) to classify defect and non-defect areas of a braided composite material from A-scans. The dataset consisted of 3600 A-scans from a reference sample with three types of debonding defects. Comparative analysis against various ML methods demonstrated that the addition of a gate recurrent unit increased the performance.

In [161], A-scan UT data was used alongside an SVM classifier with various feature extraction methods to classify defects in gas pipe girth welds. Feature extraction methods including CNN, discrete wavelet transform, wavelet packet transform, Shannon entropy, and statistical features were evaluated using a dataset of 2160 A-scans obtained via an electromagnetic acoustic transducer. CNN feature extraction method outperformed all other techniques. The authors hypothesised that this superior performance stems from its ability to extract a larger number of high-level features compared to the other methods.

A study presented in [162], compared several ML methods for classifying carburisation levels in industrial pipes using 200 A-scans. Feature extraction was performed using discrete Fourier transform, which served as input for NN, KNN, and decision tree algorithms. The study included a small hyperparameter investigation focusing on different activation functions for NN. KNN achieved 100% accuracy, NN scored 99.1%, and the decision tree reached 87.6% accuracy.

In [163], the authors developed an NN classifier for additive manufacturing components, using UT to evaluate porosity levels. The study aimed to classify porosity levels into 6 classes, trained on 24 samples manufactured with varying selective laser melting parameters. Reported results showed an overall accuracy of 93%, which was validated by testing with three new samples produced using a different set of parameters, where the model successfully classified the new data.

In the follow-up study [164], researchers explored the impact of different surface qualities on ultrasonic signals used as input for CNN, deep NN, and NN classifiers in porosity level classification. CNN demonstrated the best performance, even with UT data of low signal-to-noise ratio resulting from interaction with rough surfaces. However, the training and testing datasets were derived from the same samples, introducing a significant risk of data leakage that potentially masks the true performance of the developed models.

Authors of [165] conducted a proof-of-concept study to assess the tool degradation in friction stir welding using simulated UT data. The study aimed to determine tool length and compared 16 traditional ML models. Experimental verification indicated that the random forest algorithm performed best, achieving an average error of 2.1% compared to ground truth measurements.

In [166] researchers addressed the challenge of detecting defects obscured by larger geometrical features in composite materials using UT. Gaussian chirplet decomposition was applied to 14763 experimentally acquired A-scans to extract features for the training of a DL algorithm. Despite extended training times, the developed model successfully differentiated small flaws that were previously undetectable.

In [167] researchers utilised 75 A-scan signals from carbon and stainless-steel welds with defects. They trained a LSTM network on both raw signals and statistical features extracted from them. Training on raw signals resulted in poor performance, whereas feature extraction significantly improved overall performance.

Study detailed in [168] used DL to determine defect depth in CFRPs. The authors have compared the performance of the CNN-LSTM network to the CNN network on gated

A-scan signals. Optical microscope measurements were used as ground truth, with results showing that the CNN-LSTM was the best overall approach, achieving predictions with an 8% relative depth error.

Authors of [169] explored aluminium-epoxy joint adhesive bond quality classification. Both the dataset and extracted features were explored with extensive statistical testing, with promising classification results. Despite the challenge of manually extracting features for classification, researchers demonstrated potential in enhancing model transparency and understanding the impact of signal features on the results.

Authors of [170] employed air-coupled UT for inspection of impact damage in several types of composites. Several DL networks were developed and tested, with authors reporting promising results on a relatively simple dataset.

Study detailed in [171] developed a 1D-CNN model for reconstruction of rough surface morphology. PAUT systems struggle with rough surfaces as they influence the energy propagation into the sample due to wave scattering. The proposed network was trained on FEA-simulated data using aluminium blocks with known surface roughness. The goal of the network was to accurately reconstruct the surface profile from reflected ultrasonic signals, even when using a reduced number of sensors. The performance of the model was compared to that of a TFM algorithm, and it was shown that the model outperforms it, especially in inspection scenarios where fewer transducers are used. This work highlights the potential of ML models to handle imperfect signals and challenging inspection scenarios, and it may serve as an additional step in the data acquisition and preprocessing pipeline to improve data quality.

In [172], the authors inspected adhesive joints between aluminium and CFRPs using two samples containing a total of nine defects, inspected with an immersion PAUT setup. After acquisition, the A-scan signals were gated, bandpass filtered, and aligned to compensate for sample curvature. From the processed A-scans, 32 ultrasonic features were extracted in both time and frequency domains (e.g., as peak-to-peak amplitude, zero-crossing rate, and harmonic noise ratio). These features underwent an outlier removal process before being analysed for statistical significance. The most relevant features were then used to train a SVM classifier, which achieved 83% accuracy in defect classification and 97% accuracy in depth estimation. However, the

use of discrete categories for depth estimation is somewhat counterintuitive, as accurately localising the exact depth of defects is critical in practical inspection scenarios. Despite this, the study presents a promising approach by leveraging explainable and physically interpretable features.

While A-scans provide high-resolution temporal information from individual transducer elements, they lack the spatial context that B-scans or C-scans offer. This presents a key challenge for both human interpretation and ML models. In practice, NDE operators often rely not only on the waveform shape of a single A-scan, but also on its location within the scan area and its relationship to neighbouring A-scans. Without this contextual information, the network may struggle or fail to identify patterns that are only apparent across multiple A-scans.

2.5 Machine Learning in Ultrasonic Testing: B-scans

An alternative approach to training ML models involves using B-scans. The authors of [21] highlighted the benefits of using augmented data to train ML algorithms. Data from PAUT using the transmit-receive shear beam technique on metal pipe welds was augmented by removing and then reintroducing defects with varying locations and dimensions using virtual flaw software (a tool that generates synthetic defects on pristine data, serving as a form of data augmentation for ML training). A VGG-based model trained on this augmented data correctly classified all defects, outperforming human inspectors who had more false calls, indicating that modern ML models are capable of matching or exceeding NDE operator performance.

In the follow-up study [175], the authors applied a similar augmentation method on UT scans of austenite welds. A VGG-like architecture was trained on a simulated defective/undefective dataset and achieved the performance level of NDE operators. However, the model initially had a 14% false call rate when trained with all defect sizes. To improve generalisability, the authors refined the model by training it exclusively on larger defects, significantly reducing the false call rate to 2.3%.

In continuation of previous work with virtual flaw software, study [174] examines the impact of different defect types on an ML model's performance. The developed model excelled at detecting larger defects when trained on small ones but struggled with smaller defects when trained on large ones. This highlights a clear correlation between

model performance and the level of generalisation to the diversity of defect sizes in the training data.

The authors of [176] used 4000 UT B-scans from six stainless steel blocks with 68 unique defects. A comparative study of various object detection models was conducted, with the EfficientDet-D0 model achieving the results due to the custom anchor box design proposed by the authors. The EfficientDet-D0 outperformed the previous state of the art algorithm YOLO by 9%, demonstrating the potential of object detection models in NDE workflows.

In subsequent work [177], the authors explored integrating sequences of ultrasound B-scans into object detection models. This aimed to mimic human inspectors' visual inspection process, where assessing the surrounding area provides context for evaluating defects. Recognising that defects can span multiple B-scans, they experimented with two approaches: a) adding a sequence of 3 B-scans to the input; and b) extracting features from 3 sequential B-scans and merging them before the detection phase. The authors found that while the former approach showed no improvement, the latter method enhanced accuracy by up to 3.4%.

The final work by the same authors [178] aimed to refine detection methods for B-scans, focusing on addressing extreme aspect ratios common in UT. These extreme ratios arise due to differences in resolution between the spatial domain, limited by ultrasonic probe geometry, and the time domain, captured with higher resolution. The authors introduced a modified detection model, drawing inspiration from U-net models. The proposed changes reduced the number of trainable parameters, leading to great improvements in inference time while enhancing mean average precision by up to 2.7%.

In [179], defect detection in UT B-scan data was tackled using YOLO and SSD object detection models. The dataset comprised 490 images, featuring 157 challenging instances where NDE operators struggled with proper detection. The study found that the YOLO model achieved an average precision of 89.7%, outperforming SSD which achieved 84.5%, albeit with slower inference speeds.

In their follow-up research [180], the authors explored B-scan analysis using state-of-the-art anomaly detection algorithms. Among the tested methods, the PaDiM model showed the best performance, suggesting its potential application for anomaly detection in NDE.

In another study on the same dataset [181], researchers explored anomaly detection methods using a modified VAE architecture inspired by the GANomaly model [184]. In this approach, an additional encoder is added after the decoder to generate latent representations of the reconstructed input. This second encoder is trained separately from the rest of the network, as reusing the original encoder for reconstructions proved ineffective. During inference, differences between the latent representations of the original input and its reconstruction are used as the anomaly score. While this method yielded promising results, it struggled to accurately identify smaller defects. It was also observed that large geometric features in the data sometimes produced higher reconstruction errors than actual anomalies, leading to false positives.

In [182], the authors modified a YOLO-based architecture for defect detection in B-scan data acquired from aluminium blocks. The dataset included both simulated and experimental scans containing artificial defects. Several architectural changes were proposed, including the replacement of strided convolutions with SPD-conv [185], the integration of attention mechanisms, and the use of an adaptive feature pyramid network [186]. These modifications significantly improved model performance, achieving an F1 score of 75.68% for defect detection, outperforming both the Faster R-CNN baseline (62.50%) and the standard YOLO model (66.67%). This study demonstrates that targeted architectural modifications can enhance the applicability of general-purpose object detection models to UT inspection data.

The study detailed in [183] investigated the classification of B-scans acquired from wind turbine blades using a 4 element ultrasonic transducer mounted on a robotic arm. The objective was to categorise scans into one of four classes: defective and non-defective in the cap zone, and defective and non-defective in the cap-web zone. The data acquisition setup differed from that used in this thesis, as a water pumping system was used for coupling and generating B-scans was achieved through post-processing of incrementally acquired signals as the transducers moved across the sample. A

conventional CNN classification model was applied, achieving strong performance with classification accuracies exceeding 89% in various scenarios. To address class imbalance in the available datasets, the authors used a weighted loss approach that assigned higher importance to underrepresented classes during training. However, the use of classification outputs as a form of pseudo-localisation of defects within the scan is an unconventional approach that departs from typical strategies in NDE (i.e. the use of bounding boxes).

2.6 Machine Learning in Ultrasonic Testing: C-scans

Similar to B-scans, C-scans have also been infrequently utilised as inputs for ML models in academic research. In [187], researchers modified the YOLO family of models to tackle defect classification tasks in ultrasonic C-scans of aircraft components. Several changes were introduced to the models, the use of dilated convolutions and an additional A-scan signal classification network. The authors reported promising results and have highlighted the potential for object detection models to effectively analyse C-scan data.

Study [188] utilised a dataset previously explored in [178], this time visualising the data as C-scans. The used DL model was a CNN that classified each row of the ultrasonic C-scan as either defective or non-defective. Although positive results were reported, this approach differs from the typical analysis method used for such data, including those found in other research works or how a human NDE operator would interpret an ultrasonic C-scan. In standard NDE practice, C-scans are interpreted as 2D spatial maps representing variations in signal amplitude or time-of-flight, allowing operators to visually identify defect signals in relation to surrounding regions. Human operators typically analyse the spatial relationships across the entire scan area (not just isolated rows) to detect and characterise defects. By reducing the analysis to per row classification, the proposed method disregards this spatial context, which is especially important when defects span multiple rows or are presented as non-linear geometries.

PAUT was used for evaluating adhesive bonds in [77], where researchers used a DL model as a classification tool. The proposed workflow involved extraction of 18 features that serve as an input to the model. The study concluded that this method

enables real-time inference on thermoplastic composites, offering flexibility to handle the significant heterogeneity typical of composite materials.

In [189] the authors showcased the capability of transformer enhanced network to classify material texture from UT backscattering C-scans. The authors reported that transformer-based networks outperformed CNN-based ones.

Classification of out-of-plane fibre waviness was explored in [190]. Experimental samples were prepared by adding material strips during the layup stage to induce this waviness. Several preprocessing steps were used, such as alignment of the scans to account for imperfections during the immersion scanning and contrast enhancement. However, a potential issue of data leakage arose because multiple C-scans were generated from the same original scan used for training data extraction.

Authors of [191] conducted a comparative study on various ML models for binary classification of defects in CFRP components used in aircraft. The task was framed as a segmentation problem, where each pixel is assigned a class. Several models were tested, using both raw signals and signals transformed with discrete Fourier transform. Overall, the U-net model achieved the best performance.

The study detailed in [192] used an immersion UT setup to inspect composite panels containing artificial impact damage. A total of 60 C-scans were captured, and data augmentation techniques such as random flipping, rotation, and cropping were applied to generate a dataset of 1,150 images (300 without defects and 850 with defects). A range of classification models were evaluated (including ResNet variants, VGG, and MobileNet), with transfer learning used to improve performance. However, a potential concern arises regarding the generation and use of training data. Specifically, creating a large number of augmented samples from a small number of original scans and then randomly splitting them into training and validation sets poses a risk of data leakage, as structurally similar images may be present in both datasets. Nevertheless, the study demonstrated that transfer learning is a valuable tool, as DenseNet121 model achieved 98.8% accuracy during evaluation.

The authors of [193] examined impact damage in composite materials using ToF C-scans acquired manually with a 5 MHz PAUT transducer. The experimental dataset

consisted of 19 images, which were augmented through scaling, rotation, and elastic deformation. By analysing the damage patterns, the authors identified and measured "petal-like" impact delamination features, which were then statistically modelled to generate additional synthetic data using a custom Python script using three data configurations: (1) experimental data only, (2) experimental data with augmentation, and (3) experimental data combined with both augmented and synthetic data. The best segmentation performance was achieved when both augmented and synthetic data were included, reaching an average IoU of 88.2% with a 4.7% deviation, compared to 66.9% IoU with a 10.3% deviation when using only experimental data.

2.7 Machine Learning in Ultrasonic Testing: Alternative Works

This section provides details on works that focused on alternative input data or the use of generative models. In [194], the authors compared various feature extraction methods and ML models based on their classification performance using wave propagation images, which were generated by a laser ultrasonic imaging system. This system captures ultrasonic waves as they travel through the material, with a receiving transducer detecting the waves and producing time-series snapshots that visualise the amplitude of the waves at different points, essentially creating a dynamic image of wave propagation across the inspected specimen. The data used in this study was acquired from flawed fabricated steel plates and augmented during the training phase to enhance model performance. The authors concluded that even relatively simple DL methods achieve accuracy levels comparable to those of handcrafted visual feature extraction and traditional ML approaches. Lastly, authors have identified the scarcity of publicly available data as the primary challenge in the development of ML in the NDE field.

In the follow-up work [195], the authors introduced an open-source dataset of images, comprising over 7000 images of pristine and defective steel plates. In addition to this, the performances of various DL models were compared. The authors concluded that deeper networks with improved cross-layer connections perform better on image data from smaller datasets. Cross-layer connections (e.g. skip connections) allow information to flow more easily between non-adjacent layers, helping to preserve important features and gradients during training. These connections mitigate problems

like vanishing gradients and enable more effective learning in deeper architectures [28].

Lastly, in study [196] by the same authors, a spatial-temporal CNN for the analysis of video data generated in their previous work. Video data in this context is presented as a 3D representation of a signal, incorporating dimensions of height, width, and time. The dataset consisted of 50 videos comprising a total of 7004 individual frames. To integrate temporal information into the model, the authors explored three distinct approaches. Across their experiments, several ML models were evaluated, again confirming the superior performance of deeper DL networks.

Authors of [197] used a modified VGG-16 model to classify defects in concrete blocks using ultrasonic tomography. Due to the limited dataset comprising only 246 B-scans, the researchers employed the dropout method and data augmentation techniques to mitigate potential model overfitting. Despite the dataset's constraints, the model demonstrated high accuracy and proved effective for this application. However, the authors acknowledged limitations, such as the dataset containing only one type of easily visible defects.

In [198], the authors investigated the determination of crack length and orientation in metal workflows with plane wave imaging using multiple ultrasonic transducers. The study also addressed the challenges in NDE research related to limited datasets by creating simulated data using FEA and ray-based models. The model outperformed the standard 6 dB drop method, achieving high accuracy in predicting crack lengths and orientations, and correctly sizing 97% of the tested dataset.

In the follow-up study [199], the authors investigated various domain adaptation methods aimed at enhancing simulated data for training purposes. These included training on a weighted combination of experimental and simulated data, Regression and Contrastive Semantic Alignment (RCSA), which encourages the model to learn similar features for data points with similar labels regardless of whether they come from simulated or real data, and adversarial domain adaptation, which trains the model to confuse a separate domain classifier so that it cannot distinguish between simulated and real data features. These methods were evaluated based on their impact on the performance of a CNN model. Overall, adversarial domain adaptation proved to be

computationally intensive with many tuneable parameters, while RCSA adaptation was chosen as the best approach due to its easier implementation (only one tuneable parameter) and improved performance of the final CNN model, even when utilising limited training datasets.

In the subsequent work [200], the quantification of uncertainty using ML models was evaluated using the deep ensembles and Monte Carlo dropout methods. The authors suggest that such analyses will be crucial for future developments, particularly as data-driven NDE modalities which require rigorous regulatory qualification.

In [201], a method of explainable dimensionality reduction for crack characterisation was introduced, comparing 2D Gaussian elliptical function fitting with the traditional 6 dB method and principal component analysis. The proposed method achieves high accuracy in sizing and orientation identification for cracks while significantly reducing the dimensionality of input data. Its advantage lies in the transparent nature of the parameters derived from the fitted Gaussian elliptical function, contrasting the "black box" nature of ML models.

Domain adaptation was explored in [202], where the authors compared A-scan noise addition, C-scan noise addition, superposition of real noise, and GAN methods to enhance UT datasets captured from CFRP samples. The GAN approach demonstrated superior performance in classification tasks by using GAN-generated synthetic data to train a CNN. However, the complexity involved in training and implementing GANs posed challenges. As a result, the more interpretable method of A-scan noise addition emerged as a promising alternative for achieving effective domain adaptation.

The study detailed in [203] investigated defect detection in seven steel weld samples using a PAUT setup with an angled wedge, FMC, and different probe frequencies. From the acquired FMC data, the authors generated a dataset of over 136,000 sectorial scans by varying reconstruction parameters, with approximately one-third containing defects. Labelling was performed using a custom-built tool directly on the TFM reconstructions, under the assumption that defects visible in TFM images would also be present in the corresponding PAUT views. Each scan was classified as either defective or defect-free. The authors noted a key challenge in interpreting complex sectorial scans: they are not self-contained and typically require contextual and/or

geometric knowledge during manual analysis. To address this additional geometric information about the weld was provided to the model. A CNN classifier achieved an F-score of 93%.

Authors of [204] also investigated PAUT data from steel welds, this time using a U-Net-based architecture for defect segmentation on individual sectorial scans. The segmented outputs were then concatenated to reconstruct a 3D representation of the defects. A total of 196 ultrasonic volumes were acquired, and the 6 dB drop method was employed for labelling. However, the paper does not clearly specify the number of sectorial scans used or the details of the train/validation split. Moreover, it appears that no separate test set was used, which raises concerns about the reported results.

In [205], three thick steel welds were inspected using PAUT, producing 677 sectorial scans, each paired with an A-scan extracted from the main beamline (i.e., the central beam in the sectorial scan). Ground truth labels (non-defective, true defective, and pseudo defective) were established using X-ray, which helped address the common challenge in weld inspection where ultrasonic artefacts can mimic real defects. To process this multimodal data, a ResNet50 network was employed for sectorial scan feature extraction, while a gated recurrent unit was used to process A-scan data. The outputs were concatenated and passed through fully connected layers for final classification. This fusion approach achieved great performance, with an F1 score of 98.19%, outperforming alternative ML models such as SVM, LSTM, and AlexNet.

2.8 Transfer Learning

Transfer learning is the process of repurposing already trained ML models and adjusting them to tackle different tasks. In this approach, models are first trained on a source domain dataset, and for a new task, some retraining of the model parameters is facilitated through a new target domain dataset. This method allows for achieving improved model performance without the need to have a large, labelled target dataset, due to the efficient use of the model's features that were previously learned. An illustration of transfer learning principle is presented in Figure 28.

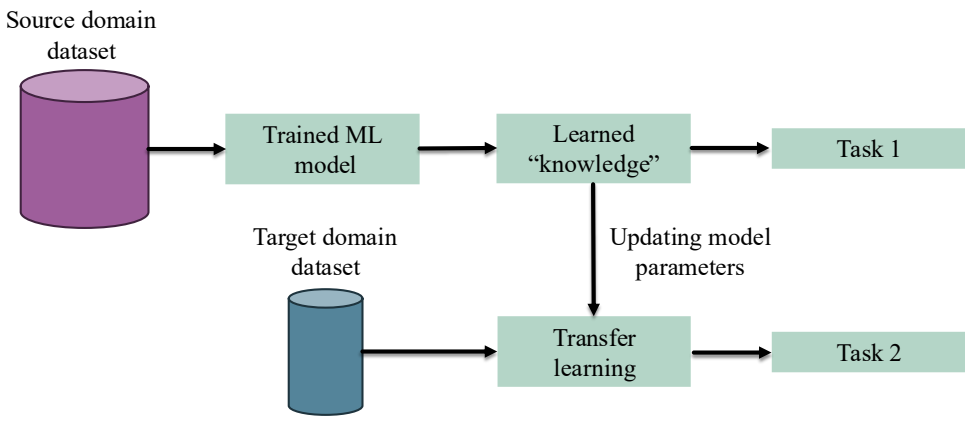


Figure 28 Illustration of the transfer learning process.

Domain adaptation is a subset of transfer learning that focuses on situations where the source and target datasets different in their distributions (domain shift) but the task remains the same [206]. These methods aim to reduce the discrepancy between source and target data representations so that model trained on one domain performs well on another.

Several academic works successfully used transfer learning, including domain adaptation, in the field of NDE. In [207], the challenge of limited X-ray imaging of CFRP material defect datasets was addressed using transfer learning and feature distribution alignment (a domain adaptation method) to classify defect, improving model accuracy by 24% with just 40 target domain images. This study demonstrates that transfer learning in combination with domain adaptation is a valuable tool for overcoming data scarcity, reducing the need for extensive dataset labelling without sacrificing model accuracy.

In another study by the same authors [208], the focus was on cylindrical metal shells commonly used in the automobile and military industries. The authors have developed a detection model pre-trained on ImageNet and successfully deployed it on a target domain of 2045 images, achieving a high model accuracy.

In [209], object detection models for thermographic images of CFRP using transfer learning were evaluated. By adapting models pre-trained on Canadian Institute For Advanced Research (CIFAR-10) and ImageNet [29], [210] the study achieved promising results despite the source datasets greatly differing from the target domain.

Studies [211] and [212] combined transfer learning and radiographic imaging of welds. Several pre-trained classification networks were adapted to X-ray data, demonstrating that the transfer learning approach achieved the best results.

Transfer learning has proven effective with ultrasonic data in medical applications. Work detailed in [213] focused on using ultrasonic data from 185 clinical studies, where DL models based on AlexNet and VGG-16 achieved superior performance compared to an abdominal radiologist, who had an accuracy of 71.7%. The DL models achieved 77.3% and 77.9% accuracy, respectively, showing the ability of ML models to surpass human operators in accuracy, even with smaller datasets aided by transfer learning.

Another study [214] applied transfer learning to develop an automated system for classifying abdominal ultrasonic images, achieving an accuracy of over 90%. The classification of breast cancer from ultrasonic images was explored in [215]. Transfer learning enabled improvements of over 15% compared to training from scratch.

Overall, these studies highlight the advantages of transfer learning and domain adaptation, leveraging models pre-trained on different datasets of transformed data for improved performance on specific tasks from a new target domain. This thesis leverages both transfer learning and domain adaptation to improve model performance. Domain adaptation (refer to Section 4.6) is used to bridge the gap between real and synthetic data distributions. Meanwhile, transfer learning (refer to Section 4.10) is used to enhance training stability and efficiency, as randomly initialised weights in object detection models often lead to suboptimal convergence. Combining these approaches allows for better handling of limited real data and more stability during model training.

2.9 Closing Remarks

The background research resulted in several key findings:

- ML research in the field of NDE has experienced substantial growth in recent years.
- Research in ML for UT primarily focuses on welds, with a smaller body of works focused on CFRP materials.

- Most studies concentrate on processing A-scan signals, with B- and C-scans rarely utilised as inputs for ML networks. This is largely due to the higher cost and complexity of phased array equipment required for B- and C-scans, as well as the need for automated scanning systems to reliably generate C-scans. Consequently, many researchers rely on the more accessible A-scan data.
- Existing studies typically rely on a single type of input (A-, B-, or C-scan), without combining multiple ultrasonic views.
- There is a clear trend towards using DL approaches, which frequently outperform traditional ML with hand-crafted features.
- Data scarcity poses a challenge and barrier to the development of ML models. Many studies acknowledge this issue and propose solutions in the form of data augmentation, synthetic data generation, and transfer learning methods.

These findings highlight key gaps in the current state of ML for UT, particularly the underutilisation of B-scan and C-scan data and the ongoing challenge of data scarcity. To address these limitations, the remainder of this thesis explores DL approaches for defect detection using B-scan and C-scan inputs, an area that remains underexplored in the literature. Chapter 4 investigates a supervised defect detection approach incorporating transfer learning, synthetic data for model training, and augmentation techniques to mitigate data scarcity, while Chapter 5 explores an unsupervised anomaly detection approach where only pristine data is used for training. Finally, Chapter 6 examines the integration of multiple ultrasonic views in a multi-modal workflow, aligning with how NDE operators interpret inspection data in real-world scenarios and proposing strategies for collaborative data analysis between human experts and DL models.

For a comprehensive review of the state of the art of ML research in NDE please refer to [22], [216], [217]. The field of ML is advancing rapidly, and a comprehensive discussion of network structures, hyperparameters, and other technical details is beyond the scope of this thesis. For an extensive overview of vocabulary and terms used in the field of ML, please refer to [218].

2.10 On the Use of Performance Metrics

The performance of ML models is typically evaluated using performance metrics such as accuracy, precision, recall, F1 score, and Area Under the Curve (AUC), among others. Each metric offers a different perspective on how well the model performs for a given task. However, in the field of NDE (especially in defect detection) performance is often reported in the form of a Probability of Detection (POD) curve. A POD curve reflects how confidently defects of varying sizes can be identified by a given method. A commonly used indicator derived from this curve is *a90/95*, which denotes the defect size that can be detected with 90% probability at 95% confidence [219].

In this thesis, standard ML performance metrics were used instead of POD curves, and different metrics were selected based on the nature of each task:

- Chapter 4: Object detection methods were evaluated using precision, recall, F1 score, precision-recall curves, AUC, and Intersection over Union (IoU).
- Chapter 5: The anomaly detection model was assessed using False Positive Rate (FPR) and True Positive Rate (TPR), which were used to generate Receiver Operating Characteristic (ROC) curves and compute the AUC.
- Chapter 6: The models were again evaluated with precision, recall, and F1 score.

Each respective chapter includes definitions and explanations of the selected metrics.

The reason POD curves were not used in this work is due to the limited amount of available testing data. Generating a reliable POD curve requires repeated testing across a wide range of defect sizes, with multiple samples per defect size (often tens or more) to establish statistically meaningful confidence bounds. Given the data constraints, using POD analysis would have resulted in unstable, noisy, and unreliable estimates. Therefore, traditional ML metrics were adopted to provide more consistent and interpretable performance evaluation within the context of this thesis.

It is also important to note that the quality and consistency of human labelling has a significant impact on the reported performance metrics. For example, in object detection tasks, a one-pixel offset in annotation corresponds to approximately 0.8 mm in real-world coordinates, which can substantially affect metrics such as IoU. Perfect

labelling is challenging, so pragmatic approaches were adopted. In Chapter 4, the IoU threshold for a correct detection was lowered to 0.25, similar to other NDE studies [220], acknowledging that predictions with higher IoUs were sometimes flagged as incorrect due to minor annotation offsets, even though they were visually accurate and practically useful. For the anomaly detection task in Chapter 5, where three observers labelled the data, inter-observer variability was addressed by averaging the labelled regions. These practical compromises had a considerable influence on the final performance metrics and may underrepresent the model's real-world effectiveness. These issues, and their implications, are discussed in more depth in section 5.9.

Chapter 3: Experimental Setup and Materials

3.1 Ultrasonic Setup

The ultrasonic NDE inspection setup used in this thesis is based on the previous work presented in [93]. It combines PAUT sensor, a Force-Torque (FT) sensor, an industrial robotic manipulator, and a Personal Computer (PC) desktop unit for automated data acquisition from different CFRP panels.

The focal point of the ultrasonic setup was the 5 MHz PAUT roller probe Olympus/Evident Inspection Solutions RollerFORM-5L64 [221] and Peak NDT Ltd. MicroPulse 6 [222] ultrasonic controller. The probe, consisting of 64 individual elements with a pitch of 0.8 mm and a total active aperture of 51.2 mm, was selected for its geometry and rolling capability, making it ideal for integration with robotic manipulators. The roller probe tyre was made from low-attenuation material with a similar acoustic impedance to water/glycol to improve coupling and wave propagation. The interior of the tyre was filled and pressurised with glycol to prevent the formation of air bubbles. The choice of glycol over deionised water was made as per the manufacturer's instructions the water needs to be replaced often in order not to damage metal parts of the assembly and to diminish the probability of air bubble formation. For this project, frequent changing of the water would be a time-consuming task, as the roller probe would have to be removed from the robotic setup and dismantled weekly. To this end, glycol was used as it does not require frequent refills while having similar acoustic properties to water. Detailed specifications of the roller probe are shown in Table 2, while the cross-sectional schematic is shown in Figure 29.

Table 2 Technical characteristics of the phased array ultrasonic roller probe used in this thesis.

Manufacturer	Olympus/Evident Inspection Solutions
Model	RollerFORM-5L-64
Central operating frequency	5 MHz
Number of elements	64
Delay line height	25 mm
Pitch	0.8 mm
Elevation	6.4 mm
Active aperture	51.2 mm

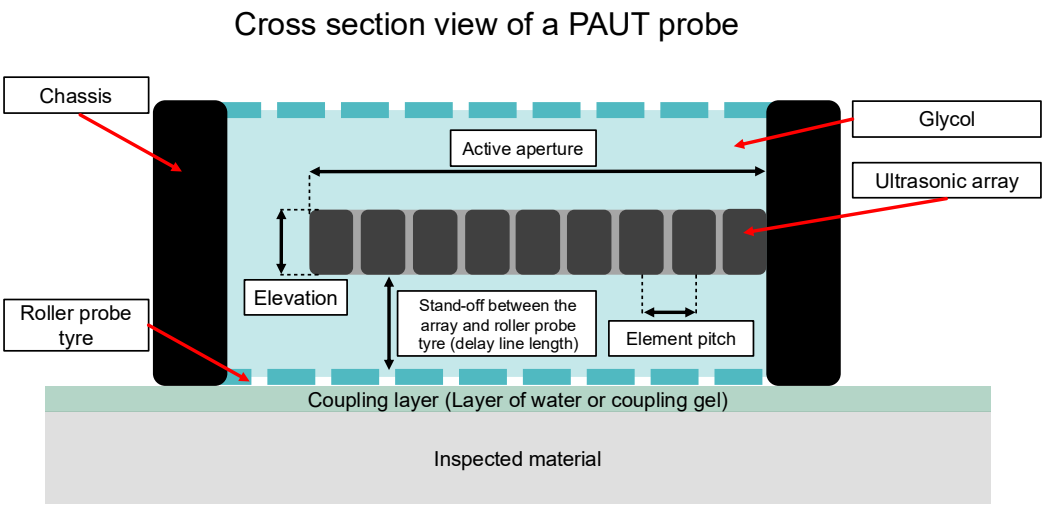


Figure 29 Cross-section schematic of phased array ultrasonic roller probe.

The ultrasonic controller was used to create custom delay laws and to drive the PAUT. With 128 transmission and reception channels, the ultrasonic controller offered flexibility in delay law design with the custom coding instruction language, edited with the central desktop PC unit. The sampling rate of 100 MHz and the 16-bit digitiser for amplitude values were employed. Due to the relatively high attenuation in CFRPs [223], a linear unfocused scanning mode was used with the sub-aperture of 4 array elements. This effectively lowered the amount of total recorded A-scans to 61 and the active aperture to 48.8 mm, while increasing the amount of energy being transferred into the material. An overall hardware gain of 22.5 dB was applied upon reception of the signal, in addition to TVG added during post-processing. The use of TVG enhances the signal amplitudes in the later stages of the ultrasonic propagation, compensating for the highly attenuative nature of the inspected CFRP material, as is set to 1.5 dB/mm, which was determined experimentally by matching amplitude responses of front and back wall. The voltage of the pulse was set at 80 V while the pulse length was 100 ns. A time delay of 11.7ms before data acquisition was added to avoid recording the initial ultrasonic pulse. A scanning speed of 10 mm/s was paired with a pulse repetition rate of 760 Hz. Digital 6 MHz low-pass and 2 MHz high-pass filters were used to filter out unwanted higher frequency signals that might induce resonance of near-surface carbon fibre layers [90].

3.2 Robotic Setup

The industrial robotic manipulator KUKA KR90 R3100 Extra HA was used for the delivery of the PAUT assembly [224]. This manipulator is the key enabler of the automated NDE system, as it allows for controlled automated sensor delivery. Its potential has been recognised in various academic works focused on robotics, such as [58], [64], [93], [94], [95], [225]. This model allows for a maximum payload of 90 kg with a reach of 3095 mm, but instead of utilising its high payload capacity, the focus in this work is on its high reach for scanning larger composite samples. The 6 degrees of freedom with the combination of an extra translation axis provided by tracks mounted in the lab allowed for coverage of a large area for flexible scanning. Most importantly, the pose repeatability of ± 0.04 mm allows for repeatable experiments, making sure the PAUT sensor is positioned at the same location between the different scans. The described robotic setup also enabled programmatic movement of the roller probe at a constant speed.

Even though the probe is used on the surface with sprayed water coupling between its tyre and the component's surface, achieving stable and constant contact force is crucial for sustained image quality during the mechanical scan. Therefore, real-time corrections and control were implemented for the PAUT probe's orientation normal to the component's surface and translation along the surface to maintain a constant coupling force throughout the surface raster scan. The real-time vertical position control was enabled through the KUKA RobotSensorInterface software package and an adaptive force-torque motion control program created within the central LabVIEW environment. This was based on real-time measurements from a Schunk GmbH & Co. FTN-GAMMA-IP65 SI-130-10 force torque sensor mounted between the probe and the robot's end effector [226]. This sensor measures multi-axis force and torque, providing feedback for precise contact control during inspection. FT enabled 3-dimensional measurements of forces and torques, within a range of 400 N in the vertical direction and 130 N in the horizontal directions. FT also served as a fail-safe measure programmed to stop the movement of the industrial manipulator if the contact force exceeds a preset value, which was set to 150 N to protect the PAUT roller probe. The PAUT, FT, and industrial manipulator assembly are shown in Figure 30.

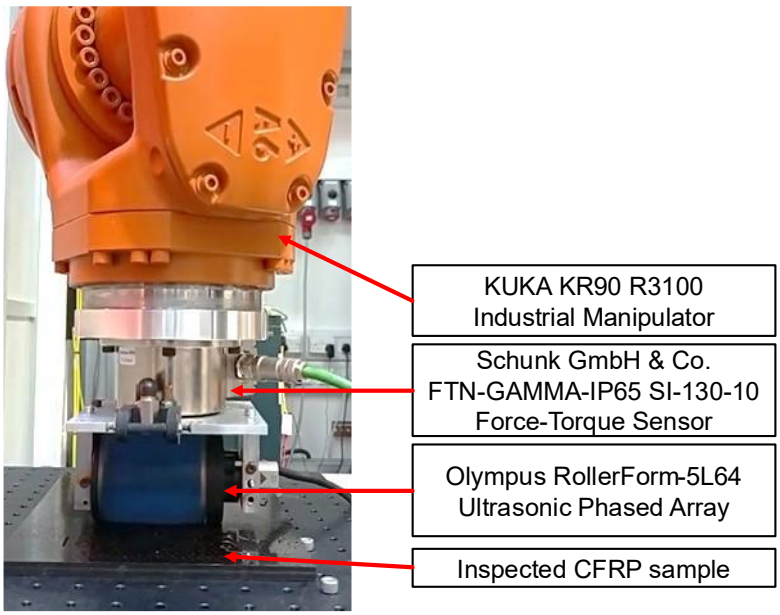


Figure 30 Experimental setup assembly used in this thesis.

All robotic movements, path-planning, and real-time tool pose corrections were programmed within the LabVIEW Virtual Instrument environment on a desktop PC connected to the robotic controller via ethernet cable and Transmission Control Protocol/Internet Protocol (TCP/IP) data communication protocol. The LabVIEW program also pushed PAUT settings to, and recorded data from the PEAK MP6, recorded data from the FT sensor, and communicated with the robotic controller, creating UT scanning data accompanied by the encoded robotic positions and FT sensor reading. Because of this, post-processing of the data allowed for precise rasterisation. As the active aperture equalled 48.8 mm, depending on the sample size, multiple robotic passes with an offset of 48 mm and a 0.8 mm overlap were performed to create a rasterised scan of samples. An example of a robotic path planning for a rasterised scan is presented in Figure 31, while a block diagram of the experimental setup is illustrated in Figure 32.

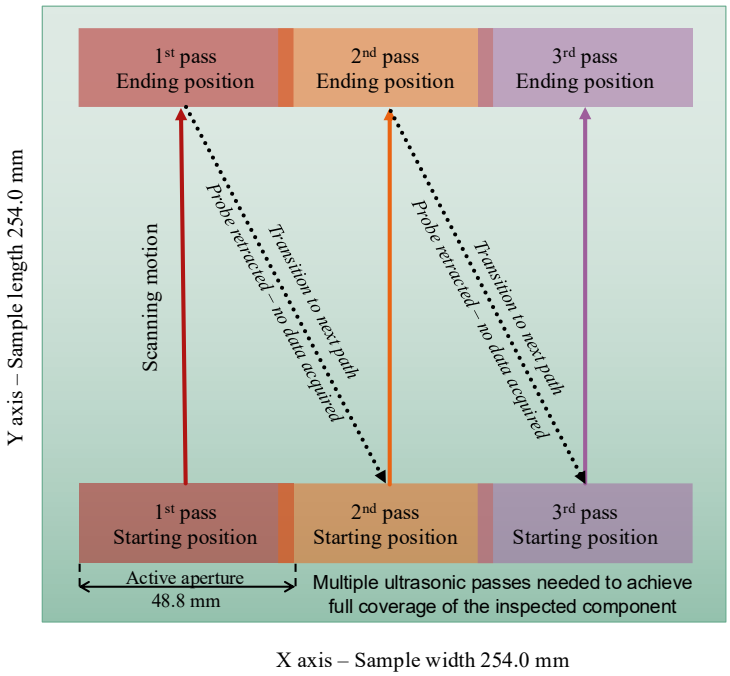


Figure 31 Robotic path planning for a rasterised scan, showing sequential passes starting at designated positions (coloured boxes) and moving along the Y-axis. Full lines indicate scanning motion, and dotted lines indicate transitions along the X-axis for complete sample coverage.

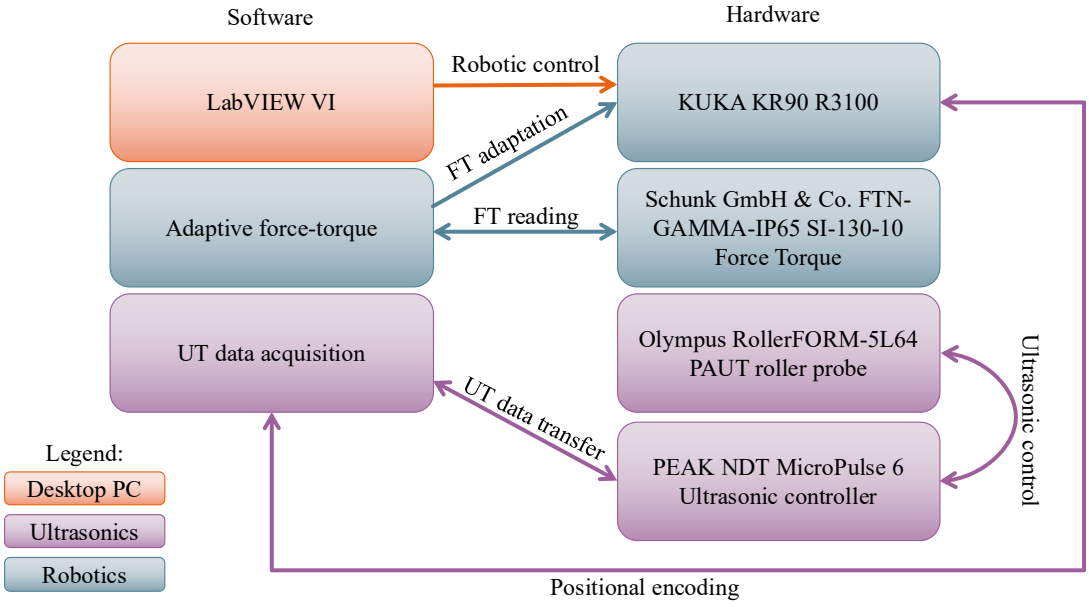


Figure 32 Block diagram of the experimental setup.

LabVIEW was also used to store the data in a custom binary format, storing the metadata, raw UT signals, positional data from the robotic controller, and FT readings. Next, a custom Python script was used to read and process this data, before storing it to the hard disk. All data were stored as a NumPy array [227], with a format of [Scanning step, Time samples, A-scans]. Data were normalised to the maximum value observed throughout the whole ultrasonic scan. In addition to this, the Hilbert transform processing was applied, resulting in separate files containing the raw, normalised, and Hilbert-processed data.

3.3 Carbon Fibre Reinforced Plastic Composite Samples

A total of thirteen CFRP samples, with thicknesses ranging from 2.20 mm to 21.2 mm, were manufactured to a Bombardier aerospace process specification standard and supplied by Spirit AeroSystems. These samples were produced using the resin infusion method, woven fabric sheets, and Cycom 890 polymer. Of these, eight pristine samples, with thicknesses ranging from 2.2 mm to 6 mm, are detailed in Table 3. The remaining samples contain intentionally introduced defects and are described in Sections 3.3.1 to 3.3.5.

Table 3 Technical details of pristine CFRP samples examined in this thesis.

Sample ID	Dimensions [mm]	Thickness [mm]	Number of B-scans [-]	Estimated number of layers [-]
1	254.0 × 254.0	2.20	1000	8
2	254.0 × 254.0	2.14	1000	8
3	254.0 × 254.0	2.75	750	10
4	254.0 × 254.0	2.75	1000	10
5	254.0 × 254.0	4.25	1000	10
6	254.0 × 254.0	4.25	1000	16
7	254.0 × 254.0	6.00	1000	22
8	254.0 × 254.0	6.00	1250	22

To capture acoustic responses similar to those produced by delaminations, Flat Bottom Holes (FBHs) were fabricated in two samples, and rectangular Teflon and other polymer inserts were embedded into three samples (detailed in subsequent sections).

FBHs and Teflon inserts are commonly used to mimic the acoustic responses of the delaminations that can occur during the manufacturing processes [228].

According to the current internal guidelines of Spirit AeroSystems for NDE inspection (internal document, not publicly accessible), critical defect sizes are specified according to their type and location on the aircraft. For delaminations, the largest allowable flaw area that would not be categorised as a defect, ranges from 60 to 500 mm^2. However, to challenge and understand the limits of the defect detection algorithms and PAUT inspection setup, FBHs with diameters ranging from 3.0 to 9.0 mm, and rectangular Teflon and other polymer inserts with dimensions of 4.0 × 4.0 to 20.0 × 10.0 mm were embedded into CFRP samples. This was done to scrutinise the performance of defect detection algorithms and test flaws with areas between 7.0 and 200.0 mm^2. 1.0 mm diameter FBHs were also trialled in this project, however, the current measurement setup was unable to capture them.

3.3.1 Sample A

The first defective CFRP sample (referred to as Sample A) measured 254.0 mm × 254.0 mm × 8.6 mm and comprised 32 layers. It featured drilled FBHs with diameters of 3.0 mm, 6.0 mm, and 9.0 mm, each with a tolerance of +/- 0.2 mm, positioned at depths of 1.5 mm, 3.0 mm, 4.5 mm, 6.0 mm, and 7.5 mm from the front face of the sample, with a depth tolerance of +/- 0.3 mm. The FBHs were spaced 35 mm apart on the X-axis and 30 mm apart on the Y-axis. Scanning of Sample A resulted in 750 B-scans. The amplitude C-scan with the exclusion surface and backwall echoes and the model of Sample A are presented in Figure 33 while its associated details are shown in Table 4.

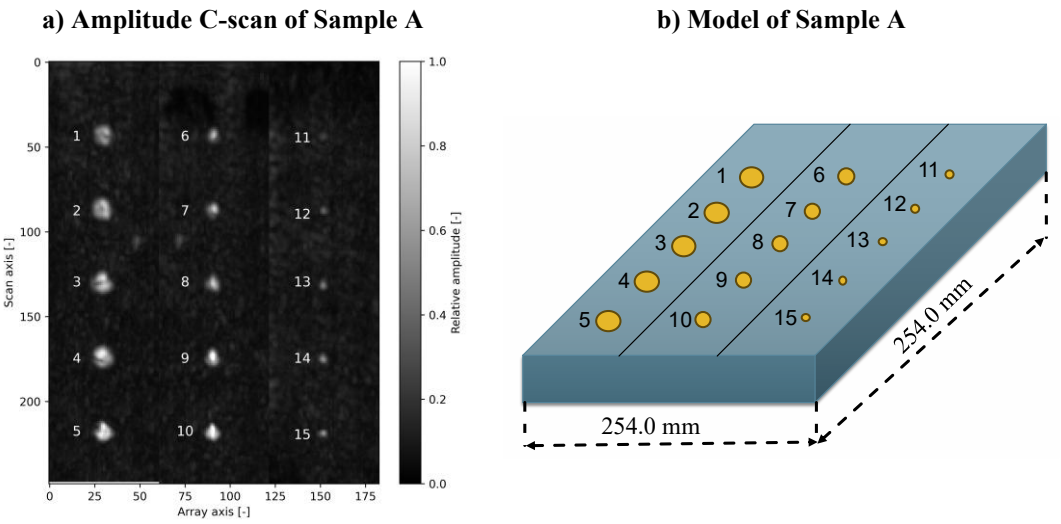


Figure 33 a) Amplitude C-scan of sample A and b) Model of a defective CFRP Sample A

Table 4 Technical details for defective CFRP Sample A

Defect ID	Diameter [mm]	~Depth [mm]	Defect ID	Diameter [mm]	~Depth [mm]	Defect ID	Diameter [mm]	~Depth [mm]
1	9.0	7.5	**6**	6.0	7.5	**11**	3.0	7.5
2	9.0	6.0	**7**	6.0	6.0	**12**	3.0	6.0
3	9.0	4.5	**8**	6.0	4.5	**13**	3.0	4.5
4	9.0	3.0	**9**	6.0	3.0	**14**	3.0	3.0
5	9.0	1.5	**10**	6.0	1.5	**15**	3.0	1.5

3.3.2 Sample B

The second defective CFRP sample (referred to as Sample B) measured 254.0 mm × 254.0 mm × 8.6 mm and was similarly constructed as Sample A, with the addition of 4.0 mm and 7.0 mm FBHs, resulting in a total of 25 defects. UT scans of Sample B resulted in 1150 B-scans. The amplitude C-scan and the model for Sample B is illustrated in Figure 34 while the details are presented in Table 5.

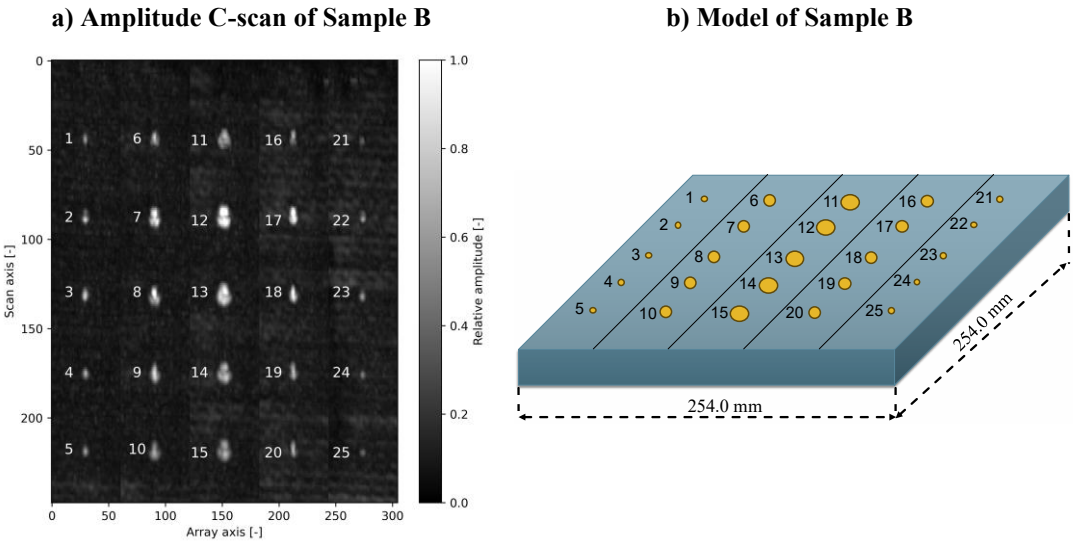


Figure 34 a) Amplitude C-scan of sample B and b) Model of a defective CFRP Sample B

Table 5 Technical details for defective CFRP Sample B

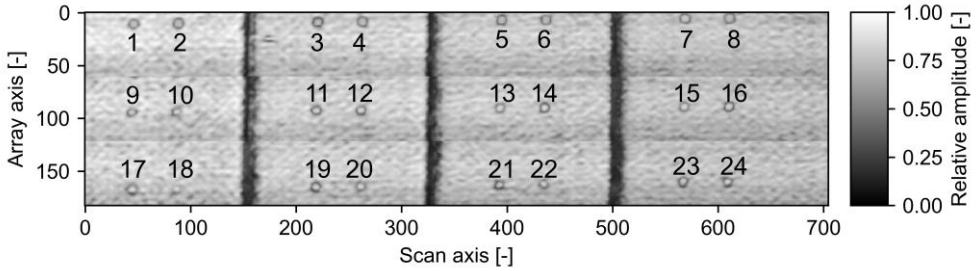
Defect ID	Diameter [mm]	~Depth [mm]	Defect ID	Diameter [mm]	~Depth [mm]	Defect ID	Diameter [mm]	~Depth]mm]
1	3.0	7.5	**10**	6.0	1.5	**19**	7.0	3.0
2	3.0	6.0	**11**	9.0	3.0	**20**	7.0	1.5
3	3.0	4.5	**12**	9.0	1.5	**21**	4.0	3.0
4	3.0	3.0	**13**	9.0	3.0	**22**	4.0	1.5
5	3.0	1.5	**14**	9.0	1.5	**23**	4.0	3.0
6	6.0	7.5	**15**	9.0	3.0	**24**	4.0	1.5
7	6.0	6.0	**16**	7.0	7.5	**25**	4.0	3.0
8	6.0	4.5	**17**	7.0	6.0			
9	6.0	3.0	**18**	7.0	4.5			

3.3.3 Sample C

The third defective CFRP sample (referred to as Sample C) was a large, stepped specimen measuring 780.0 mm × 200.0 mm, with thicknesses ranging from 7.5 mm to 16.0 mm in increments of 2.1 mm. At each thickness step, three square-shaped 6.0 mm × 6.0 mm Teflon inserts were embedded. These inserts were positioned immediately after the front wall, in the middle of the sample, and near the back wall. Inspection of sample C resulted in 2070 B-scans. The amplitude C-scan and model for sample C is

depicted in Figure 35, while the technical details are shown in Table 6. Due to the physical limitations of the used ultrasonic setup, discussed in section 5.10, the thickest section of the sample was excluded from the analysis.

a) Amplitude C-scan of Sample C



b) Model of Sample C

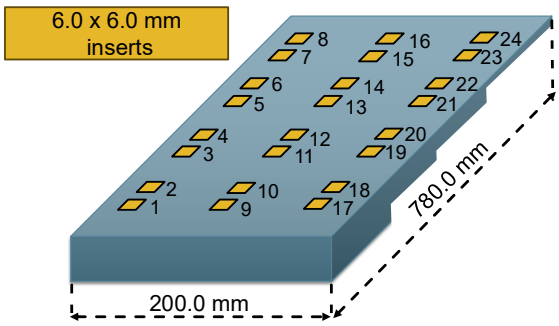


Figure 35 a) Amplitude C-scan of sample C and b) Model of a defective CFRP Sample C

Table 6 Technical details for defective CFRP Sample C

Defect ID	Start/end plies	~Depth [mm]	Sample thickness [mm]	Defect ID	Start/end plies	~Depth [mm]	Sample thickness [mm]
1	2 / 3	0.65	13.50	**13**	18 / 19	4.80	9.59
2	2 / 3	0.65	13.50	**14**	18 / 19	4.80	9.59
3	2 / 3	0.65	11.70	**15**	14 / 15	3.76	7.46
4	2 / 3	0.65	11.70	**16**	14 / 15	3.76	7.46
5	2 / 3	0.65	9.59	**17**	50 / 51	13.11	13.50
6	2 / 3	0.65	9.59	**18**	50 / 51	13.11	13.50
7	2 / 3	0.65	7.46	**19**	42 / 43	11.03	11.70
8	2 / 3	0.65	7.46	**20**	42 / 43	11.03	11.70
9	26 / 27	6.88	13.50	**21**	34 / 35	8.96	9.59
10	26 / 27	6.88	13.50	**22**	34 / 35	8.96	9.59
11	22 / 23	5.84	11.70	**23**	26 / 27	6.88	7.46
12	22 / 23	5.84	11.70	**24**	26 / 27	6.88	7.46

3.3.4 Sample D

The fourth defective CFRP sample (referred to as Sample D) was a smaller stepped sample measuring 300.0 mm × 90.0 mm, containing embedded rectangular Teflon tapes of sizes 12.0 mm, 6.0 mm, and 4.0 mm. The sample had varying thicknesses at different steps: 21.2 mm, 16.3 mm, 13.8 mm, 10.9 mm, and 7.9 mm (79, 61, 51, 41, and 29 layers respectively). At each thickness step, nine embedded inserts were positioned: one of each dimension near the front wall, in the middle of the sample, and near the back wall.

Scanning this sample was challenging due to its small size and the proximity of defects to the sample edges, posing risks to the PAUT roller probe when operating near the boundaries. As shown in the amplitude C-scan in Figure 36, this led to the omission of several defects. The amplitude C-scan and the model for sample D is illustrated in Figure 36.

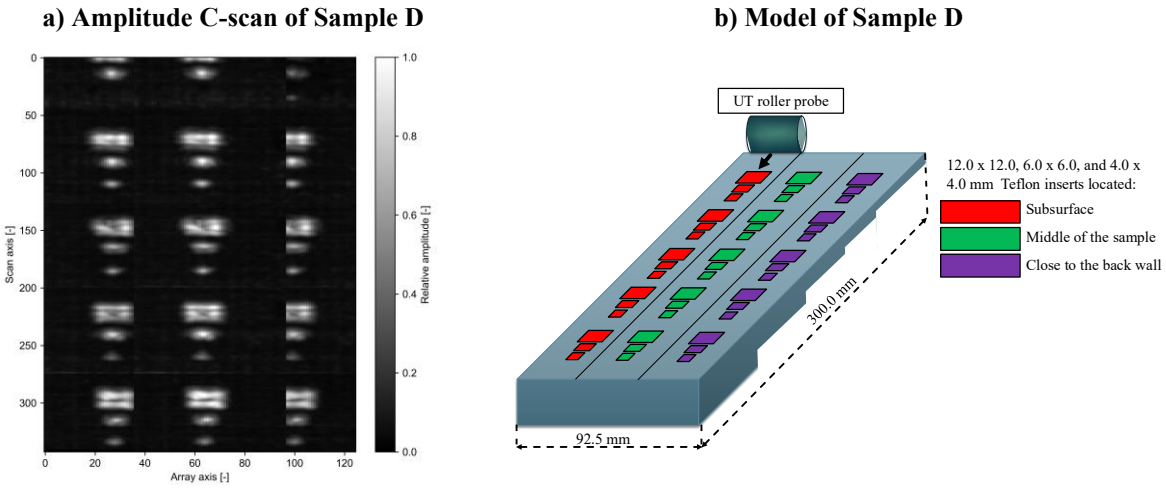


Figure 36 a) Amplitude C-scan of sample D and b) Model of a defective CFRP Sample D

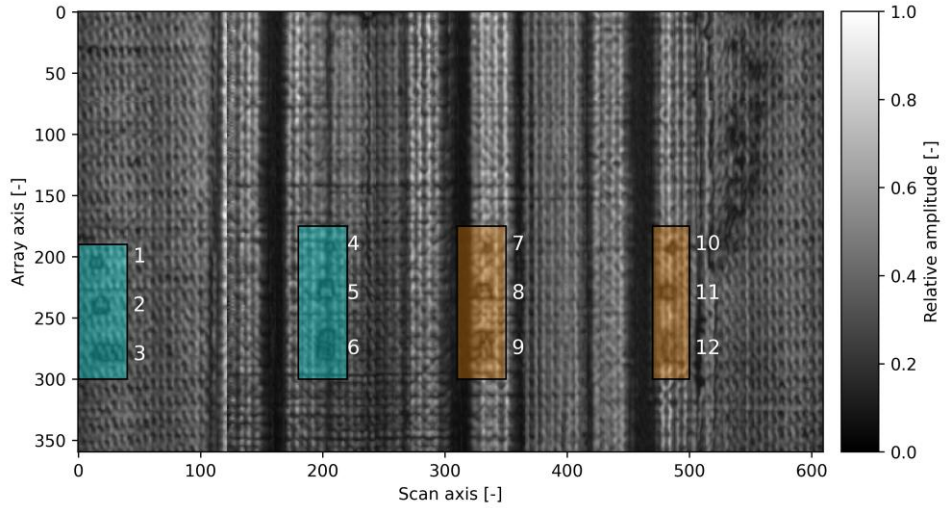
3.3.5 Sample E

The last defective CFRP sample (referred to as Sample E) was composed of a flat panel skin surface (thickness of 7.8 mm or 29 layers) co-cured with three stringer sections (thickness of 12.5 mm or 47 layers). The sample contained 12 Teflon inserts, with 6 located immediately beneath the surface and 6 beneath the stringer sections, as detailed in Table 7. The sizes of the inserts were 20.0 × 10.0 mm, 10.0 × 5.0 mm, and 5.0 × 5.0

mm. Overall, the sample consisted of 3600 individual B-scans. The amplitude C-scan and model of sample E is illustrated in Figure 37.

a) Amplitude C-scan of Sample E

Cyan: Subsurface inserts / Orange: Stringer inserts



b) Model of Sample E

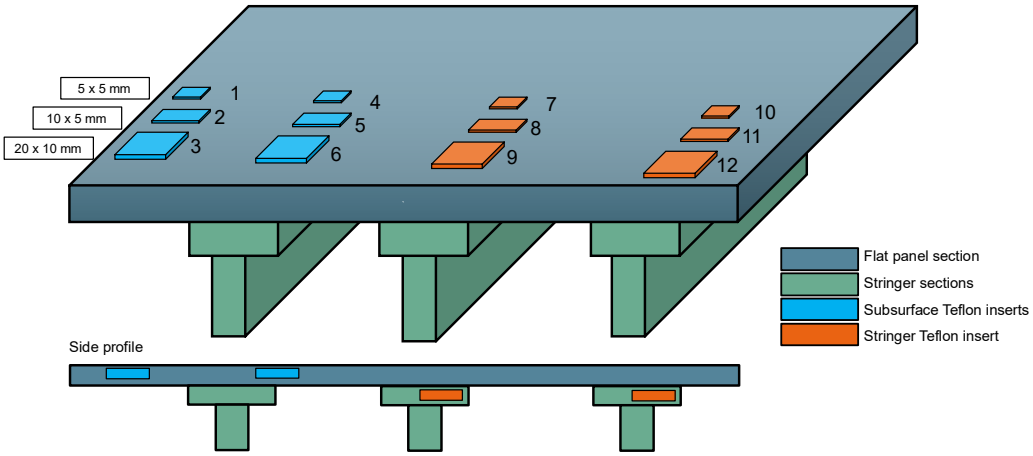


Figure 37 a) Amplitude C-scan of sample E and b) Model of a defective CFRP Sample E

Table 7 Technical details for defective CFRP Sample E

Defect ID	Insert size [mm × mm]	Start/end plies	~Depth [mm]	Sample thickness [mm]
1	5.0 × 5.0	2/3	0.9	7.8
2	10.0 × 10.0	2/3	0.9	7.8
3	20.0 × 10.0	2/3	0.9	7.8
4	5.0 × 5.0	2/3	0.9	7.8
5	10.0 × 10.0	2/3	0.9	7.8
6	20.0 × 10.0	2/3	0.9	7.8
7	5.0 × 5.0	18/19	6.8	12.5
8	10.0 × 10.0	18/19	6.8	12.5
9	20.0 × 10.0	18/19	6.8	12.5
10	5.0 × 5.0	18/19	6.8	12.5
11	10.0 × 10.0	18/19	6.8	12.5
12	20.0 × 10.0	18/19	6.8	12.5

3.4 Hardware and Software Setup

The AI models described in this thesis were initially prototyped on a Windows 11 Dell Precision 5570 laptop with an Intel i9-12900H 2.50 GHz Central Processing Unit (CPU), 64 GB of Random Access Memory (RAM), and an NVIDIA RTX A2000 8GB GPU. These models were developed using the Python programming language and the PyTorch [142] framework. Prototyping on a laptop allowed efficient use of available computing power before finalising the model architecture.

Following prototyping, the AI models were trained and tested on a high-performance desktop Windows 11 PC. This system was equipped with an Nvidia RTX 3090 Ti GPU, 128 GB of RAM, and two Intel® Xeon® Gold 6428 2.50 GHz CPUs. This setup provided enhanced computational capabilities necessary for extensive model training and testing.

Simulation work in EXTENDE CIVA [229] and POGO [230] was conducted on another high-performance PC setup. This system featured an Intel® Xeon(R) Gold 6248R CPU, Nvidia RTX 3090 Ti GPU, and 192 GB of RAM. The simulation setup and parameters are detailed in Section 4.3, where their relevance to the supervised defect detection approach is further discussed.

3.5 Conclusion

In summary, the experimental NDE setup comprised a PAUT roller probe and controller, a robotic manipulator for precise scanning, and a FT sensor to maintain consistent contact pressure. A set of CFRP test samples was scanned to acquire inspection data from specimens with varying thicknesses, layup configurations, and geometries. Additionally, simulation software was used to complement the experimental data and support the training of supervised AI models.

The collected and simulated ultrasonic data form the foundation for the analyses in the following chapters. In Chapter 4, simulated data are used for training supervised models, while experimental data serve for validation and performance evaluation. Chapter 5 adopts an unsupervised approach for the development of AI models, where both training and validation is performed collected data. Lastly, Chapter 6 integrates the methodologies from Chapters 4 and 5, applying them to the most complex dataset, captured from a real aircraft component (sample E), to evaluate how the developed techniques align with real-world NDE practices and inspection scenarios.

Chapter 4: Supervised Object Detection Machine Learning Approach Analysis of Amplitude C-scans

4.1 Chapter Overview

Given the limited past research investigations and the broad gap in the knowledge regarding automated defect detection, this chapter focuses on a comparison between the capability of various defect detection methodologies applied to amplitude C-scans of CFRP components. Firstly, an amplitude thresholding method, frequently used within the industry, was trialled as a baseline for comparison. Afterwards, an improvement was shown with the implementation of the statistical amplitude thresholding method, inspired by previous work in the fusion of ultrasonic data [231]. Lastly, the reliability of AI algorithms based on widely used object detection models such as YOLO, Faster Region-based Convolutional Neural Network (R-CNN), and RetinaNet was investigated, highlighting their key strengths and shortcomings. The training datasets were created using the semi-analytical simulation software CIVA and were further augmented with A-scan noise profiles based on the method proposed in [202]. This approach reduces reliance on large volumes of experimentally acquired defect data, which are difficult to obtain (especially for real defects in CFRP components). There are currently no publicly available datasets containing large, labelled collection of such defects. While mimicking the UT inspection process is a non-trivial task, it is considerably more feasible than replicating the manufacturing conditions under which representative defects may occur is extremely costly and challenging to control. As a result, gathering a sufficiently diverse and well-annotated experimental dataset for supervised learning is often infeasible. However, by generating synthetic but representative training data, the AI models can be effectively trained without requiring extensive experimental datasets (i.e., real ultrasonic data acquired from physical inspections). Their performance was subsequently evaluated on real ultrasonic amplitude C-scans of CFRP samples A, B, C, and D.

4.2 Contributions

This work introduces an object detection model training pipeline for ultrasonic C-scan images of CFRP components, characterised by its reliance on fully synthetic training data combined with explainable data augmentation (domain adaptation). The pipeline

begins with the generation of a dataset using CIVA software simulation, incorporating realistic variations in defect size, depth, and orientation. To improve model generalisation, real A-scan noise profiles were extracted from experimental CFRP scans and integrated into the synthetic images.

Three object-detection architectures (YOLOv5, Faster R-CNN, and RetinaNet) were adapted, trained, and validated on this synthetic data, then evaluated on independent experimental dataset. By performing multiple training runs with fixed random seeds, the study established that the training process is repeatable and stable across different data splits. The optimal model from each architecture was selected for final performance reporting, and a comparative assessment of computational efficiency and detection accuracy was conducted. Results demonstrate that augmented synthetic data can substitute for real-world training data, providing an effective and scalable solution for training ML models for ultrasonic NDE applications.

In addition to ML approaches, this work also benchmarked conventional industry-relevant defect detection technique (amplitude thresholding method) and an improved statistical thresholding method based on fitting mathematical distributions to pixel intensity histograms in C-scan data. This provided a clear baseline for assessing the relative benefits and limitations of AI-driven techniques in practical settings.

4.3 Introduction

In recent years, there has been an abundance of development of new object detection models with examples being R-CNN, Fast R-CNN, Faster R-CNN, Efficient-Det, and YOLO [232], [233], [234], [235], [236], [237]. These models use a complex architecture to extract regions of interest of an input image, outputting the bounding box and class of the object in the form of a vector. Despite the rise in the number of academic publications, object detection models have seen limited implementation with UT data. Performances of EfficientDet, RetinaNet, and YOLO models on B-scans of steel samples were compared in [178]. Authors have reported promising results with architectural changes to address the issue of extreme aspect ratios observed in UT B-scans. Similarly, object detection on ultrasonic B-scans was evaluated in [179], demonstrating the use of YOLO and SSD models and highlighting the differences in performance in inference speed between the tested models. Lastly, researchers in [177]

combined EfficientDet and several methods that enabled the processing of additional B-scans in the sequence, improving on the baseline results.

In industrial applications, defect localisation and sizing are usually performed manually through visual inspection of the C-scan, while applying different thresholds to the image. The most used method is a 6 dB drop where a threshold value is imposed on the signal to separate pristine and potentially defective regions. Researchers have used a 6 dB drop to separate damaged and undamaged areas in a C-scan image to assess the extent and size of impact damage [238]. The authors compared how sizing results vary with different methods and proposed a new algorithm that improves the sizing and shape of the damage. Limitations of the 6dB method were recognised in [239], especially when sizing defects that are smaller than the width of the ultrasonic beam. As an improvement the authors developed an AI approach that can automatically acquire different thresholding values, hence reducing the errors in quantification of defects. A semi-automated detection algorithm was proposed in [240]. This approach works on ToF C-scans, where the user defines areas of interest and threshold values which are in turn used for automated analysis. However, these traditional methods are inherently subjective and limited in flexibility. Although 6dB drop method remains commonly used, NDE operators often adjust thresholds depending on the scan location and the specific features under examination. Furthermore, thresholding performs well primarily when inspecting well-defined, stable signals, which are rarely the case in CFRP scans which exhibit complex signal characteristics [22]. Due to these challenges, ML approaches have been recognised as a promising solution due to their adaptability to non-linear signal patterns from data. However, it is noted that safety-critical industries such as aerospace remain cautious in adopting automation, primarily due to the safety concerns, the need for transparency, reliability, seamless integration into existing workflows, and the requirement for workforce upskilling [20].

Automatic defect localisation in CFRPs has been scarcely explored in the past; authors of [241] developed a time-dependent thresholding that improved the detection of micro flaws in UT C-scans of stainless-steel samples. Statistical analysis of backscattering noise to determine defect locations was used in [242], but the scope of this work was limited. Several works have used Otsu thresholding to segment ultrasonic images into

clusters of areas with similar acoustic properties [243], [244], [245]. Otsu's method is a global thresholding technique that determines the optimal threshold by minimising the variance between foreground and background pixels, based on the assumption of a bimodal histogram distribution of image intensities. However, this assumption limits its industrial application as ultrasonic data typically contains complex signals that do not follow a clear bimodal distribution [246].

The most recent work was presented in [247], where an AI object detection model successfully localised damage on ToF C-scans of aircraft wings. The authors demonstrated an accuracy of 94.5% for the best-performing model when training and testing on experimentally collected data.

While the experimental samples described in Chapter 3 provide valuable real-world data, their size remains relatively small, especially when compared to the large open-source datasets commonly used in ML research (e.g., COCO dataset [248]). As previously highlighted, obtaining large and well-annotated datasets from real ultrasonic inspections, particularly for CFRP composites, is challenging. Therefore, in this work, which focuses on supervised object detection models, the experimental data from Chapter 3 is reserved for model testing, while simulated datasets are used for training. The generation of simulated data is detailed in Section 4.3.

4.4 Generation of Simulated Data

For the generation of simulated data, a semi-analytical NDE UT software CIVA by EXTENDE S.A. was used [229]. This approach is more efficient and less computationally demanding than using the FEA software, as it uses ray tracing theory. Ray tracing theory models the propagation of ultrasonic waves as straight-line paths (rays) that reflect and refract at interfaces, based on geometric acoustics. While computationally efficient, it neglects complex wave phenomena like scattering, wave dispersion, and mode conversion, which are critical in capturing the true acoustic behaviour of composite materials. As a result, it fails to capture realistic noise characteristics caused by inter-laminar scattering and diffraction. FEA software, on the other hand, can simulate and calculate complex interactions between the wave and individual layers of the composite with greater accuracy, resulting in a more representative simulation. However, this often requires a painstaking definition of

individual layer's properties and dimensions. To gain an understanding of simulation time differences between CIVA and FEA, a similar scenario of a probe on a defective sample was modelled in CIVA and an FEA wave propagation software POGO [230]. Both simulations were executed on a high-performance PC detailed in section 3.4. The CPU-intensive CIVA simulation was completed in less than 2 minutes, whereas POGO FEA simulations even with GPU parallelisation took more than 2 hours of processing. In this study, 300 simulations of defect responses were created to include a variety of defect sizes at different depths in the CFRP sample. Given the large number of simulations and the observed time discrepancy between the CIVA and FEA, a semi-analytical modelling approach was used, with an attempt to reintroduce the compromised signal features in the post-processing stage.

Upon deciding on the simulation software, a square composite sample with dimensions of 100.0 mm × 150.0 mm × 8.0 mm was created and a range of FBHs were introduced in the model. A parametric sweep study was used for ease of data collection, where FBHs' diameters ranged from 3.0 to 15.0 mm, each placed at depths of 1.5 to 7.5 mm in steps of 0.5 mm measured from the inspection surface. Each simulation in the sweep contained one defect in the centre of the sample. The flow chart of the simulation process and an example output data for a defect of 6.0 mm at a depth of 4.5 mm are displayed in Figure 38 in a form of an amplitude C-scan.

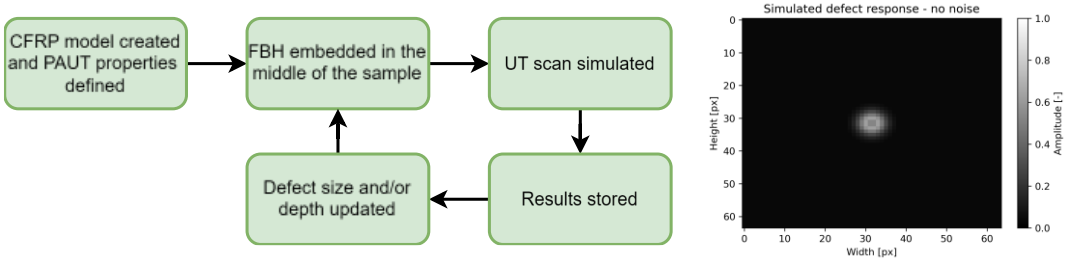


Figure 38 Simulation process flow chart for the parametric sweep of defect size and depth (left) and an example of simulated C-scan image of a 6.0 mm FBH at 4.5 mm depth (right)

The composite model was defined with a total of 8 carbon fibre layers in orientations of 0°, 45°, 90°, -45°, 0°, 45°, 90°, and -45 ° with the thickness of each layer being 1 mm. This was different from the experimental samples which were made with non-crimp fabric. Fibre layers were considered transversely isotropic with a density of 1670 kg/m^3 while polymer matrix was defined as isotropic material with a density of 1230 kg/m^3. Longitudinal and transversal wave velocities were set at 2488 m/s and 1134

m/s, respectively. These values were determined experimentally by conducting an ultrasonic scan on a pristine sample with known thickness. Next, on ultrasonic data, the distance between front and back wall reflections was calculated and correlated with the sampling rate of the ultrasonic controller. Lastly, the speed of sound was calculated with:

$$v = \frac{2 * d}{n_{samples}/f_s} \quad \text{Eq.21}$$

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

$$\alpha(f) = \sum_{p=1}^{n} \alpha_p * f^p \quad \text{Eq.22}$$

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

To create a scanning path simulation, an immersion linear phased array with 64 elements, 0.8 mm pitch and an element gap of 0.1 mm was modelled with a stand-off of 20.0 mm from the sample filled with water with no assumed attenuation and a velocity of 1483 m/s. The operating frequency of the array was set to 5 MHz with Hanning windowing. Scanning was performed in linear mode with a sub-aperture of 4 elements to match the experimental setup. The step of array movement was set to match the array element pitch of 0.8 mm and was moved across the defect in a total of 64 steps. Overall, this resulted in a total of 300 simulations of FBHs, that were used for the training of the AI models.

4.5 Signal Processing and Imaging

Simulated data and captured experimental data were stored as 3D arrays comprised of all A-scans collected along the electronic and mechanical scanning direction of the array and the robotic arm, respectively.

Data were normalised with respect to the maximum amplitude occurring across all captured A-scans. Next, a Hilbert transform was applied to each A-scan to extract the

envelope of the signal, followed by time gating to remove the front and back wall responses. Time gating was done manually for each sample due to the varying material thicknesses. Lastly, maximum amplitudes of gated signals were used to construct amplitude C-scan images. To illustrate the difference between a simulated and real amplitude C-scan, a comparison is shown in Figure 39.

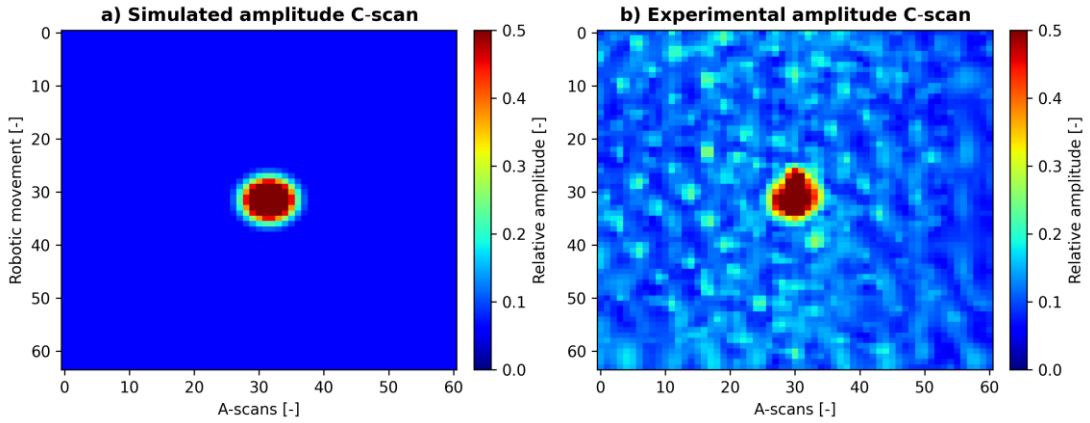


Figure 39 Illustration of a C-scan containing a defect a) Simulated amplitude C-scan; and b) Experimental amplitude C-scan

4.6 Augmentation of Synthetic Data

When comparing the C-scans presented in Figure 38 and Figure 39, there is a clear difference in structural noise that the CIVA model failed to capture. This also adversely affects the defect response as the defect indication from the CIVA model looks undisturbed and uniform. AI models benefit from training on data that represents reality as accurately as possible; therefore, a post-processing approach should be used to overcome the lack of modelling noise which is present in the experimental data. To this end, the method of A-scan noise addition proposed by [202] was implemented. The foundation of this noise augmentation approach stems from the fact that each A-scan is composed of structural noise, resulting from interactions between individual material layers, and random noise from sources such as electrical interference. The authors have demonstrated that this method improves the performance of AI models compared to the use of raw simulated data. For noise profile analysis, a pristine CFRP sample was used. In this context, the noise augmentation method can also be seen as a form of regularisation, helping to prevent the model from overfitting to idealised, noise-free simulations. Moreover, because the added noise profiles are grounded in the physical characteristics of ultrasonic wave interactions with CFRP structures, this

approach introduces a limited but meaningful degree of physics-informed modelling into the training process.

To separate structural noise from random noise in the A-scan signals, a two-step process was used.

Step 1: Estimation of structural noise

The goal here was to isolate consistent noise patterns caused by the internal structure of the composite (i.e., inter-laminar reflections).

1. First, all A-scans across the full scan were averaged to obtain a global mean A-scan. This averaging reduces the influence of random noise and helps capture the stable structural features that are constant across scan.
2. Then, for each B-scan in the dataset, all its A-scans were averaged to produce a B-scan-specific mean A-scan.
3. By subtracting the global mean A-scan (from step one) from each B-scan's mean A-scan, variations pertaining to structural noise were isolated.
4. This process was repeated across all B-scans. The resulting structural noise data was compiled into a histogram and modelled using a normal distribution, with a standard deviation of 0.003.

Step 2: Estimation of random noise

This step aimed to capture the random signal fluctuations due to electronic or environmental sources.

1. For each B-scan, the B-scan average A-scan (from the structural noise step) was subtracted from each individual A-scan within that B-scan.
2. This subtraction removes consistent features (i.e., structural noise), leaving behind the purely random noise component.
3. Again, this was repeated for all B-scans. The aggregated random noise data was approximated with a normal distribution, with a measured standard deviation of 0.013.

The process for the calculation of structural and random noise is presented in Figure 40.

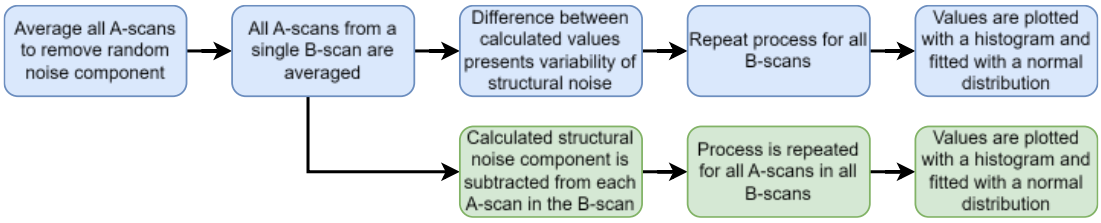


Figure 40 Process for determination of structural (blue) and random (green) noise components.

The generation of a new noise profile was performed by applying mean structural noise and adding a variance that corresponds to the normal distribution. Following this, the random noise component is added with the mean and variances calculated in the previous steps. Figure 41 illustrates the simulated response, generated noise, and the final combined synthetic image.

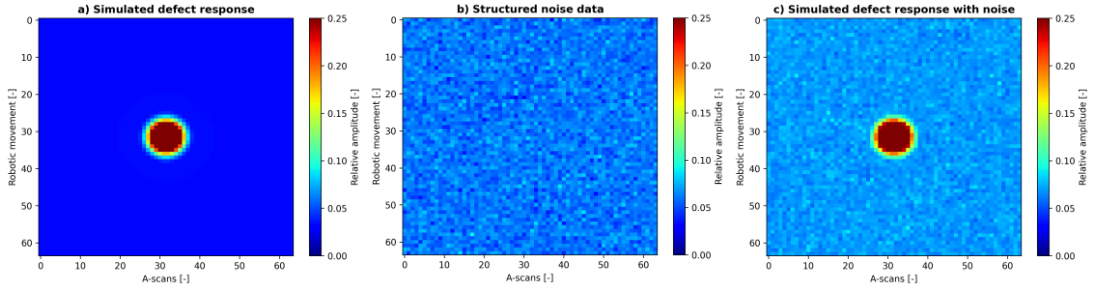


Figure 41 Representation of augmentation results. A) Simulated defect response; b) Generated noise; and c) combined image.

4.7 Amplitude Image Thresholding

The first method of defect detection that was explored in this work was amplitude image thresholding. In industry, a 6 dB drop on A-scans is often used for defect sizing, but in this work, the approach is adapted for defect detection and localisation. Physically, the 6dB drop on an A-scan signal represents the positions/time samples at which the maximum signal response loses half of its amplitude. In the ideal case, assuming A-scans were gated properly to exclude the front and back wall echoes, the maximum signal response would be created from the strong scatterers such as delaminations. Similarly, this loss of amplitude can be examined at an amplitude level other than half of the original value (e.g., 9, 12 or 18 dB). While used frequently, the 6 dB drop often performs poorly when it comes to larger defects of irregular shape and defects that are smaller than the beam width of the acoustic wave [239]. A smaller

body of research tackled this issue and analysed alternatives to the 6 dB method [238], [239], [249]. The amplitude thresholding approach is hereby used as a baseline for comparison.

To apply the amplitude thresholding method to the experimental data, the maximum pixel value of the resulting C-scan image was found, and the image was thresholded for 6-, 9-, and 12-dB drops (corresponding to 50%, 65% and 75% losses of amplitude). All pixels that had values lower than the calculated threshold were set to 0, while those with values larger than the threshold were set to 1, creating a binary map of the original image. Next, the spaghetti algorithm [250] was used to find connected components. The algorithm selects an unmarked pixel and assigns it to a new connected component, and afterwards, it moves to neighbour pixels and assigns them to the same connected component. This process is repeated until all pixels are assigned. Furthermore, the algorithm produces coordinates and areas of connected components. Lastly, resulting coordinates are used to create rectangles that encapsulate the corresponding defective area. For display purposes, these rectangles were overlaid over the original image.

4.8 Statistical Image Thresholding

In addition to the previous method, a statistics-based approach was also evaluated. This process is based on work presented in [231] where no prior knowledge about defects is needed, only that they have sufficiently different acoustical responses than defect-free areas. Firstly, a representative defect-free section of the amplitude C-scan from a pristine sample was extracted and used for statistical analysis. The goal of this method is to convert pixel values to probability values, where a higher value indicates a higher probability that an individual pixel belongs to a defect class. The pixel amplitudes in the extracted section were represented by a histogram, with a number of bins calculated with the Freedman-Diaconis rule [251]. Next, the SciPy Python [252] package was used to test theoretical distributions and determine the best Probability Density Function (PDF). PDF is the mathematical representation that describes the likelihood or probability of observing different values for some continuous variable. By extension, a Cumulative Density Function (CDF) is also computed. CDF is a related concept to PDF, as it indicates the probability of encountering a value that is less or equal to a point described by the PDF. Lastly, each pixel value from the original

image was remapped to a corresponding probability according to the CDF. For the current set of data, an f-distribution was determined to provide the best fit to the histogram. When using a pristine scan as the baseline, pixels with values similar to this reference are unlikely to be flagged as defects, since their probabilities remain low, reflecting the defect-free nature of the data. A range of probabilities was used (99, 99.5, 99.9%) to determine defective areas in the remapped image. An example of generated PDF and CDF for the pristine sample is presented in Figure 42.

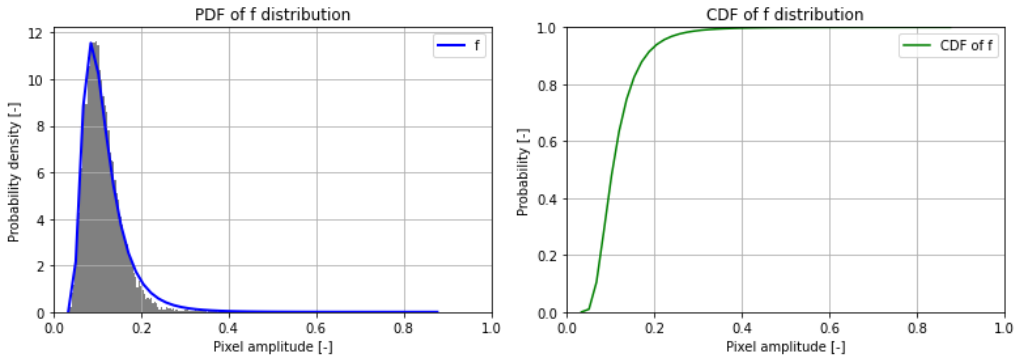


Figure 42 Probability density function (left) and resulting cumulative density function (right) of a pristine sample.

4.9 Object Detection Neural Networks

In this chapter, the defect detection performances of YOLO, Faster R-CNN and RetinaNet family of models were compared. The choice of networks stems from their track record as state-of-the-art models on various object detection datasets, and from variations in their architecture that influence their inference speed and performance. To leverage prior knowledge and improve training efficiency given the limited availability of annotated ultrasonic data, transfer learning was employed. All networks were pre-trained on the Microsoft Common Objects in Context (COCO) dataset [248], which contains 80 object classes. Using these pre-trained weights as a starting point improves training stability and convergence compared to random parameter initialisation. During fine-tuning on the defect detection task, all layers of the network were updated (i.e., none were frozen). Furthermore, the final classification layer was replaced to output only one class corresponding to defects. To ensure consistency, all tested AI and thresholding methods were evaluated on the same images.

4.9.1 Faster R-CNN

Faster R-CNN is an ML architecture released in 2015 as an improvement over the earlier R-CNN models for object detection in images [232], introduced in [233]. The authors recognised that the region proposal step in R-CNN was the main bottleneck in terms of computational time and to address this the Region Proposal Networks (RPNs) to generate region proposals more efficiently was introduced. Faster R-CNN comprises two components, RPNs and Fast R-CNN that perform object classification on the areas proposed by RPN. These two structures share convolutional layers which enable end-to-end training.

The RPN operates on the feature maps produced by earlier convolutional layers. It uses a sliding window approach, where a 3x3 kernel is moved over the feature map. At each position, the RPN predicts a set of region proposals, which are potential bounding boxes that might contain objects. These proposals are generated based on anchor boxes that have different scales and aspect ratios. Anchor boxes are predefined bounding box shapes that serve as a reference for generating region proposals. Following the RPN, the Fast R-CNN takes proposed regions and performs feature extraction using pooling layers that convert variable-sized region proposals into fixed-size outputs. These outputs are then propagated through fully connected layers that perform classification.

For training, the authors used a multi-task loss function which combines classification and bounding box regression losses. ResNet50-FPN, a variation on ResNet architecture introduced in [253], was used for feature extraction and creation of feature maps. Recommended hyperparameters used by the original authors were used, with changes to the batch size and training epochs. For robustness and faster training convergence, initial pre-trained weights from a Faster R-CNN model that was trained on the Microsoft COCO dataset were adopted.

4.9.2 You Only Look Once

You Only Look Once object detection models were initially introduced in [254] , with multiple iterations being released in recent years from various research teams [235], [236], [237], [255]. Compared to the region proposal and sliding windows methods used in R-CNN and Fast R-CNN, YOLO introduced techniques that improved both accuracy and inference speed. These include single-stage detection where both

bounding box coordinates and class are determined with a single pass through the network and mosaic augmentation which enhances the training datasets. In this study, the implementation by the company Ultralytics [255] was utilised. This implementation includes architectures of varying sizes and complexities that were pre-trained on the COCO dataset. All architecture variants share the common underlying structure consisting of a Cross Stage Partial (CSP) network in the backbone, a Path Aggregated Network (PAN) in the neck, and a YOLOv3 detection head.

The CSP network [256] was implemented in the backbone due to its efficiency and the ability to deploy trained models to setups with weaker CPUs and GPUs. CSP is based on DenseNet and introduces the splitting of the gradient flow, which increases speed and performance. The focal point of CSP gradient splitting is the convolutional layer with 1x1 kernel size, which is computationally efficient and used to increase the complexity of the architecture. The spatial pyramid pooling [257] layer block is located at the end of the backbone and allows YOLO networks to accept input images of any resolution by max pooling of the same input multiple times with different kernel sizes and strides before concatenating them. With this method, the output is always of the same dimension, making it compatible for use in the subsequent layers. The neck is the central part of the YOLO structure, which comprises a series of network layers that collect and integrate various characteristics obtained from the backbone before passing them to the final detection layers. The neck of the YOLO model is the PAN, developed in 2018 [258] and first introduced in YOLOv4 [237]. In short, it is an improvement over the previous Feature Pyramid Network (FPN), which is based on feature maps of varying sizes. The improvement came from additional lateral connections between low- and high-level feature maps in the feature pyramid. Lastly, the head portion of the network produces predictions in the form of a vector with the class of the object and coordinates for the proposed bounding box.

4.9.3 RetinaNet

In 2017, the authors of [259] developed a single-stage object detection model called RetinaNet that achieved better performance than its two-stage counterparts. The novelty in this work is a new loss function that addresses the issues of class imbalances that can happen if cross-entropy is used as a loss function during training. The new

loss function is called “focal loss,” and it diminishes the losses by an order of one magnitude for high-probability examples like pristine CFRPs, while still retaining high losses for low-probability examples such as the occurrence of defects.

Like YOLO, RetinaNet is a one-stage object detector that uses an FPN network for multi-scale feature representation. The classification and bounding box regression are handled by two smaller task-specific neural networks. The new loss function was combined with ResNet-101 and FPN to create RetinaNet, a model that achieved state of the art on the COCO dataset. However, with an inference time of 200ms, the final performance was less suitable for real-time tasks. In this implementation, ResNet50-FPN was used as the backbone. Hyperparameters used by the original authors were followed, with changes to training epochs and batch size. Similar to previous two models, pre-trained weights and biases were used. Table 8 summarises main characteristics of each network

Table 8 Comparison of object detection models used in this study

Model	Stages	Backbone	Proposal mechanism	Anchor boxes
YOLOv5	One-stage	CSPDarknet53	Dense prediction	Yes
Faster R-CNN	Two-stage	ResNet50-FPN	Region Proposal Network	Yes
RetinaNet	One-stage	ResNet50-FPN	Dense prediction	Yes

4.10 Model Training

The training dataset consisted of 300 synthetically generated C-scan images of size 64 × 64 pixels. The range of generated defects ranged from 3.0 mm to 15.0 mm, at 12 different depths starting at 1.5 mm measured from the front surface and extending to 7.0 mm. To support model validation, 10% of this synthetic dataset was randomly set aside as a validation subset. The data splitting was performed using fixed random seeds to ensure repeatability of the training and validation splits across all network trainings. All defects were circular, with no deviations in shape or position within the generated image. A separate experimentally acquired testing dataset used for model testing consisted of 8 amplitude C-scans, containing a range of defect types and sizes. FBHs were present in samples A and B, while samples C and D contained Teflon, and other polymer inserts which were rectangular.

All networks were trained using a desktop PC detailed in the section 3.4. An overview of the training hyperparameters for all the models used for training are presented in Table 9. Hyperparameters were chosen according to values proposed by the original authors of the used models. While some degree of hyperparameter optimisation has already been conducted in previous studies by the original authors, further task-specific tuning within the present context could yield additional performance gains. However, such optimisation was beyond the scope of the current study. During training, data augmentation in the form of random image translation, scaling, and vertical, and horizontal flipping was introduced, except for the YOLO family of models, which additionally employed mosaic augmentation. During model deployment onto the test dataset, no augmentations were used. All models were trained for 50 epochs, and model weights at the lowest validation score were saved.

Table 9 Overview of used training hyperparameters and other technical information

Hyperparameter	**YOLO Medium**	**YOLO Large**	**Faster R-CNN**	**RetinaNet**
Epochs	50	50	50	50
Learning rate	0.01	0.01	0.005	0.0005
Momentum	0.937	0.937	0.9	0.9
Optimizer	SGD	SGD	SGD	SGD
Batch size	32	32	32	32
Weight decay	0.0005	0.0005	0.0005	0.0005
Model size in MB	42 MB	92 MB	159.7 MB	130.27 MB
Parameters	21.2M	47M	41.8M	34.0M

To demonstrate the convergence of models, training and validation losses are shown in Figure 43. Due to the small size of the training dataset, convergence in validation loss is achieved relatively early.

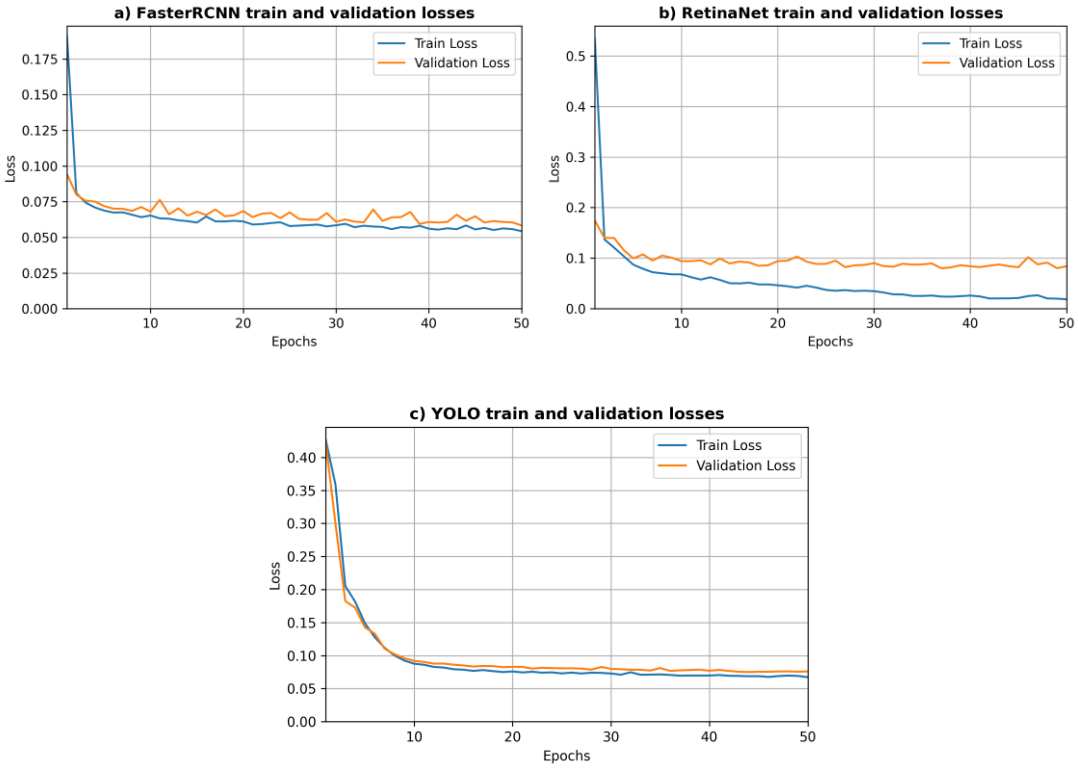


Figure 43 Training and validation losses for: a) Faster R-CNN; b) RetinaNet; c) YOLO

4.10.1 Performance Metrics

Each experiment involved training 10 networks with different fixed random seeds to verify training stability and convergence. The best-performing model from each set was then used to compute the reported performance metrics. In this study precision, recall, and F1 score were used as evaluation metrics. Precision is defined in Eq.23, and it illustrates the percentage of positive predictions that are correct according to the ground truth:

$$P = \frac{TP}{TP + FP} \qquad \text{Eq.23}$$

Where *TP* annotates true positives and *FP* annotates false positives. Recall is defined as the likelihood of detecting objects determined by ground truth. Mathematically this is represented in Eq.24:

$$R = \frac{TP}{TP + FN} \qquad \text{Eq.24}$$

Where *FN* denotes false negatives. F1 score is defined as a harmonic mean of precision and recall and is shown in Eq.25:

$$F1 = 2 * \frac{P * R}{P + R} \quad \text{Eq.25}$$

For NDE applications, recall is more important as it is crucial to not miss any defects while having some false positives is tolerable at the expense of adding to the analysis time. To evaluate which predictions are considered positive, Intersection Over Union (IoU) is used. IoU is represented in Eq.26:

$$IoU = \frac{Pred \cap GT}{Pred \cup GT} \quad \text{Eq.26}$$

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

4.11 Results and Discussion

In sample A, the application of amplitude thresholding with a 6 dB drop failed to identify four 3.0 mm FBHs and two smaller delaminations. This failure was attributed to the presence of stronger reflectors in the scan, specifically shallower 9.0 mm FBHs, which contained the maximum amplitude of the image. Similar observations were made in sample B containing FBHs, where a single 4.0 mm FBH and several 3.0 mm FBHs went undetected. Furthermore, in both samples C and D, the 6 dB method proved inadequate in identifying shallower and smaller indications, resulting in a poor overall defect detection performance with only 38.8% of the defects being correctly identified. Such low performance is also attributed to the defined IoU level of 0.25, as some predictions were made in the correct area but were much smaller than the provided ground truth.

The use of a more aggressive 9 dB drop method led to the identification of more defects. However, in the samples with FBHs, the shallowest 3.0 mm defect and two small delaminations were once again missed. The 9 dB drop method performed well in detecting all Teflon inserts, however several false negative indications started to appear. This issue was particularly prominent in sample C, which exhibited brighter areas in the scan due to imperfections during the scanning process and the application

of gating parameters for image creation. In this case, the gating process for C-scan generation incorporated some reverberations from the front wall, which were misinterpreted as defective areas. Compared to the 6 dB drop method, the 9 dB drop method achieved a much higher defect detection rate of 72.5%.

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

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

By increasing the probability to 99.5%, the number of false positives decreased. This adjustment had a positive impact on both the overall precision and F1 score, resulting in increases of 8.8% and 8.1% (from 34.7% to 43.5% and from 50.8% to 58.9%). With a recall rate of 91.2%, the statistical image thresholding method outperformed the amplitude threshold technique, but with lower precision. Furthermore, similar to the amplitude thresholding method, the statistical approach generated false positive

indications when features other than defects with higher amplitudes were present in the image. The statistical method exhibited high sensitivity to gating parameters, which greatly influenced the number of false positive indications. Notably, when testing this method on pristine samples, no false detections occurred as the pixel values were close to the mean of the statistical distribution, without obvious outliers. A similar trend continued for the 99.9% threshold, where precision increased to 64.9%, but the recall dropped to 76.2%. The precision of the statistical thresholding method could be improved by imposing an additional area threshold in the predictions, but this creates a risk of filtering smaller defects. Overall, the presented method provides an improvement over the amplitude thresholding method, especially in the recall values, with room for improvement when it comes to its precision.

The Faster R-CNN implementation trained on raw data detected all larger defects, but it consistently struggled to detect the smallest and deepest FBH defects. On average, the Faster R-CNN model performed well in generalising to rectangular-shaped defects and FBHs that are 4.0 mm or larger in size. The major benefit of this implementation is an improved precision score of 98.6% when compared to the statistical method and amplitude thresholding methods. This improvement is attributed to the ability of AI models to learn complex features that describe defective areas, whereas previous methods relied solely on amplitude values. As a result, the robustness of the model matches that of previous methods while providing increased resilience to imperfections in the scanning process and signal gating. When data augmentation was performed, it resulted in minor increases in precision, recall, F1, and AUC scores (1.1%, 1.2%, 1.1%, and 1.1% increase, respectively). Nevertheless, both the Faster R-CNN trained on raw data and the augmented data show improvements over previous techniques by providing a more robust detection mechanism, with significant enhancements in precision and F1 metrics.

Similar to the Faster R-CNN, RetinaNet provided an increase in precision when compared to statistical thresholding, but still oftentimes missed smaller 3.0 and 4.0 mm FBHs. RetinaNet generalised well on the rectangular-shaped defects in samples C and D but had several false indications that were very close to the positive indications. Such false indications were refined with a better choice of Non-Maximum Suppression (NMS) thresholds. Furthermore, there were some instances where indications captured

multiple defects under the same bounding boxes. These were considered false positives as the clear separation between the defects is important, even when they are close in distance. Upon augmentation, precision dropped by 3.6% (98.4% to 94.8%), but the recall rate increased by 4.0% (90.9% to 94.9%). The main difference was that 3.0 mm and 4.0 mm defects were identified with greater recall rates than with the Faster R-CNN.

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

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

produce more bounding boxes of lower confidence, and the results are heavily influenced by NMS and confidence thresholds.

Overall, all models provided an improvement over the statistical thresholding and amplitude thresholding methods, even when trained on raw simulation data. The augmentation of training data positively impacted recall and F1 scores of all models, with the minor exception of the large YOLO model. The augmentation led to the most prominent increases in FasterRCNN, which overall produced the best results for this dataset. Furthermore, with the optimisation of confidence, IoU, and NMS thresholds, this score could be refined further. Another improvement would be the implementation of an ensemble of different AI models, where several models would process the same input and provide a combined output. The good results achieved in terms of recall with amplitude and statistical thresholding were always followed by a low score in precision and vice versa, making it hard to strike a balance between all performance metrics. On the other hand, AI models provided more balanced and robust results throughout different tests. Visual representation of test results is illustrated in Figure 45, with ground truth being presented with red bounding boxes and test results with green bounding boxes. Precision and recall curves and evaluation metrics of all tested methods are presented in Figure 44 and Table 10. Precision and recall scores in Table 10 are reported based on the maximum achieved F1 score.

Unlike ML models that produce continuous confidence scores (enabling creation of full precision-recall curves), amplitude thresholding methods rely on fixed, physically meaningful criteria. Commonly used thresholds such as 6, 9, and 12 dB have clear interpretability in ultrasonic inspection and are standard in practice. Sweeping these thresholds continuously would be physically arbitrary and, more importantly, not reflective of how such methods are applied in real NDE scenarios. Moreover, since precision-recall curves for ML models are based on varying confidence scores, including many discrete thresholds for rule-based methods would require a different approach to plotting.

For consistency and direct comparison, the same discrete approach was adopted for the statistical thresholding method. Although a continuous sweep of percentiles is technically feasible, lower thresholds (e.g., below 99.0%) would be expected to yield

an unusually high number of false positives, distorting the practical value of the analysis. Therefore, for both amplitude and statistical thresholding, a small number of discrete, interpretable thresholds were selected to reflect realistic usage in practice, and to enable a fair comparison on the same precision-recall plot as the ML-based methods. This links back to the issues on the use of performance metrics in AI and NDE research discussed in Section 2.10.

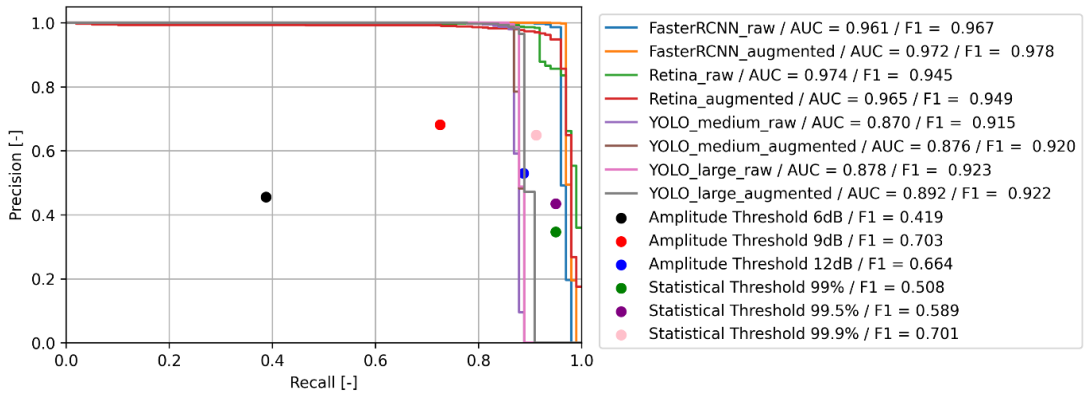
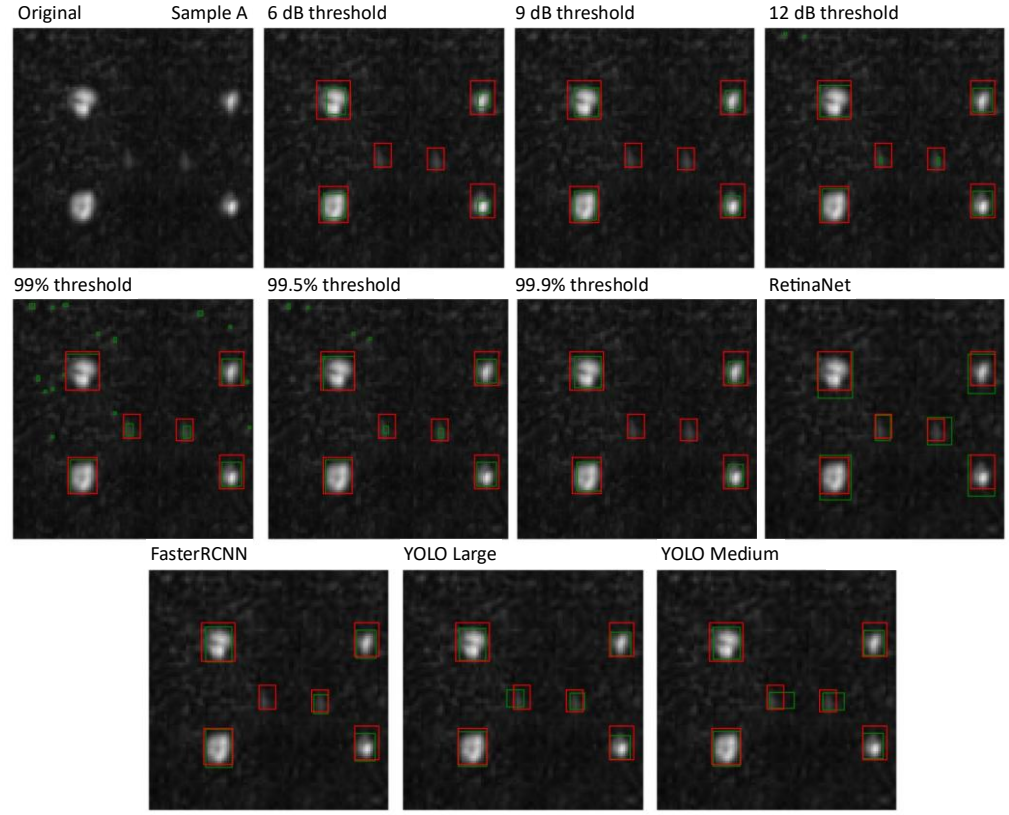
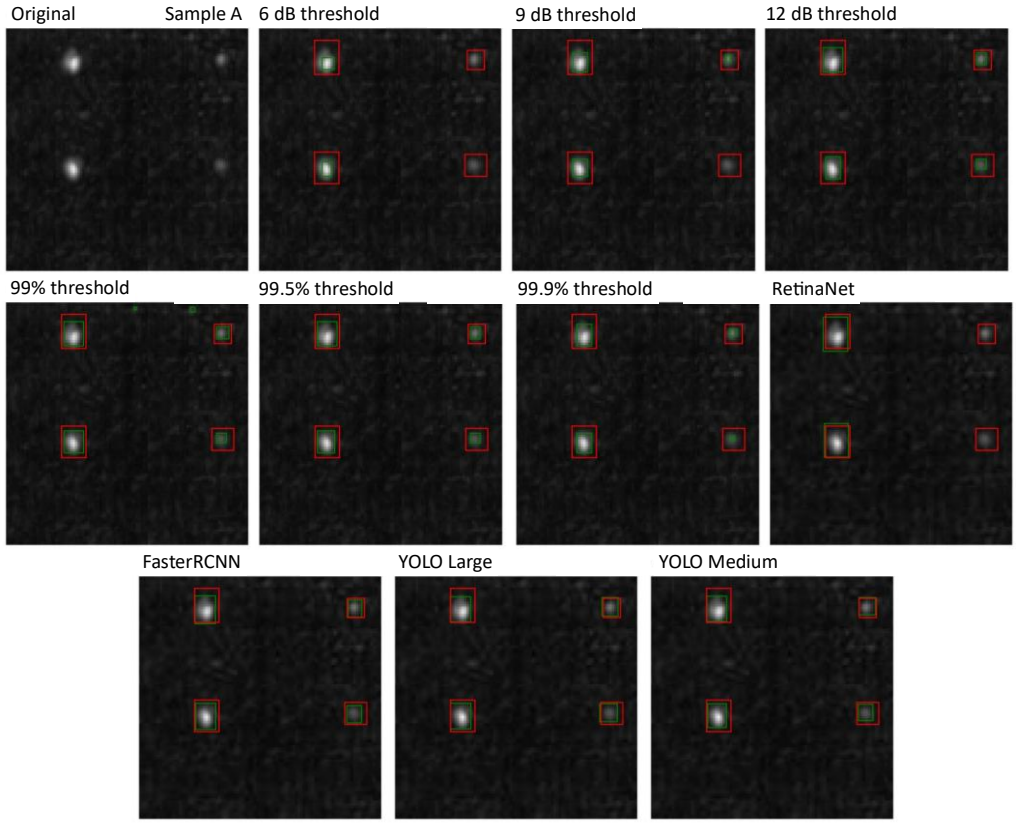


Figure 44 Precision and recall curves for all tested methods of defect detection, with IoU was set at 0.25.

Table 10 Evaluation metrics for the experimental dataset for IoU set at 0.25.

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


Original
Sample A
6 dB threshold
9 dB threshold
12 dB threshold
99% threshold
99.5% threshold
99.9% threshold
RetinaNet
FasterRCNN
YOLO Large
YOLO Medium


Original
Sample A
6 dB threshold
9 dB threshold
12 dB threshold
99% threshold
99.5% threshold
99.9% threshold
RetinaNet
FasterRCNN
YOLO Large
YOLO Medium

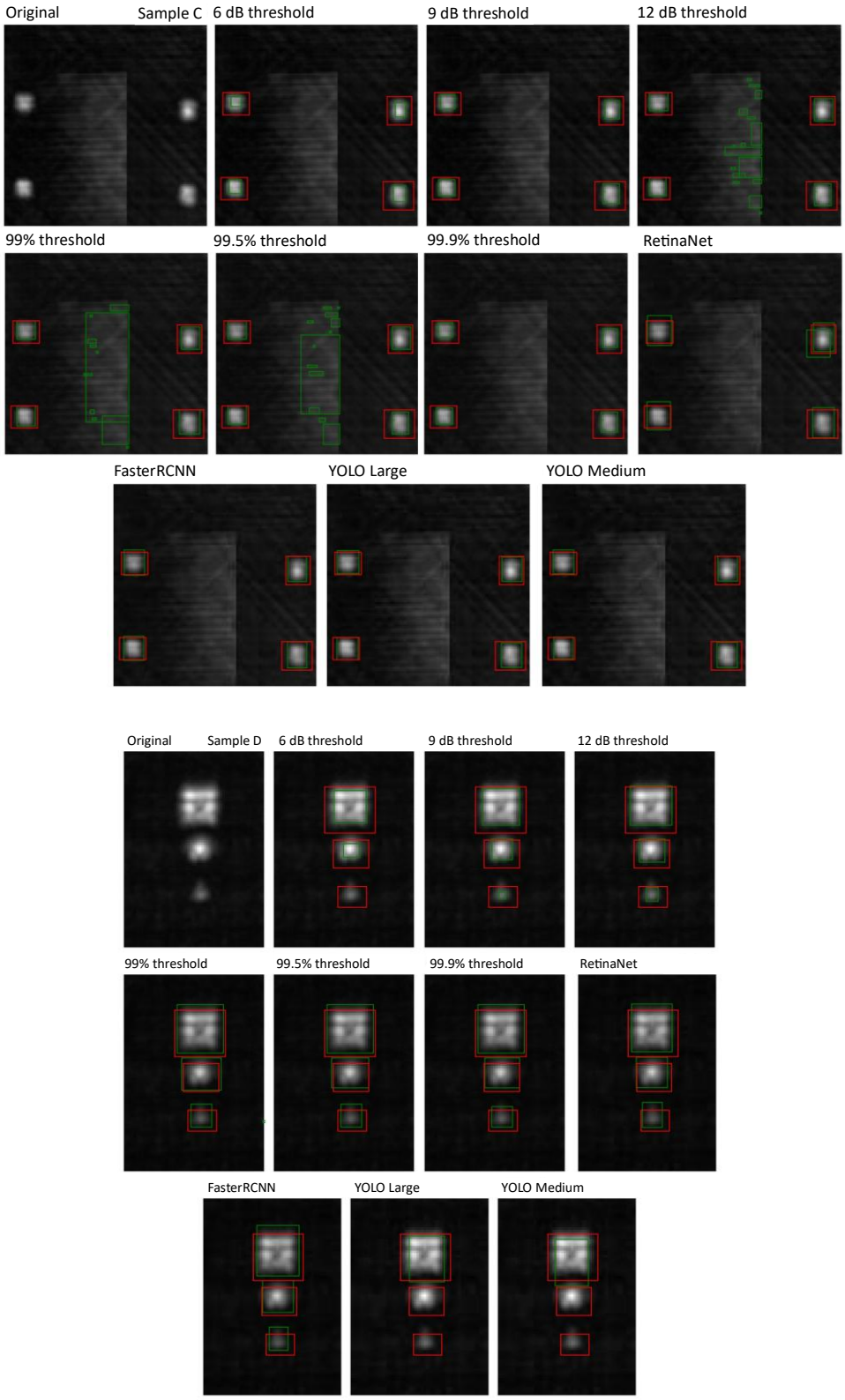


Figure 45 Extracted sections of several testing images. Names of samples and used detection method are listed above each example, with the ground truth bounding box marked in red and test results in green.

Training times were modest due to the small dataset size, and the full computational times are presented in Table 11. The medium YOLO model was the fastest to train and the fastest model during inference. Amplitude thresholding was the overall fastest method, due to its simplicity. Statistical thresholding took 1250.8 ms per image, but this time is heavily influenced by the number of tested statistical distributions. The fitting process is repeated for each new image, but for scans with similar backscattering noise, it is possible to perform this process only once and significantly speed up the inference times. In the reported results, five candidate distributions were tested.

Table 11. Computational times for examined methods, including training and testing times.

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

4.12 Conclusion, Limitations, and Future Work

The main novelty of this chapter lies in demonstrating that ML-based defect detection in ultrasonic amplitude C-scan images can be trained and validated exclusively on synthetic data, eliminating the need for experimental data in the training pipeline. This was made possible by a data augmentation (domain adaptation) approach in which realistic A-scan noise profiles, extracted from experimental CFRP scans, were injected into simulated data. This strategy represents a practical and scalable solution to the limited availability of annotated experimental ultrasonic datasets.

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

method, removing the need for the use of experimental data in the training loop. Through the investigations, it was concluded that:

- The amplitude thresholding method is suitable for the detection and localisation of large reflective defects. However, this method was unable to detect smaller defects and was heavily reliant on the absence of any other reflective features that trigger false indications. Furthermore, this method performed poorly on scans where no defects were present.
- The improvement to this method is the statistical image thresholding method, which outperforms amplitude thresholding by reducing false indications in pristine samples. However, the performance of this method was also reliant on the absence of high amplitude artefacts in the images.
- Lastly, the four different AI models tested for defect detection provided improvement over the statistical method, by accurately identifying defective areas. These models also demonstrated robustness in processing pristine samples without producing false indications, overcoming a key limitation of the amplitude thresholding method. The performance of trained models was further improved by the application of a noise profile augmentation method to simulated data for domain adaptation. Validation of the models during the training process on a subset of the synthetic dataset was valuable as it diminished the need for the acquisition of a separate validation dataset.

The study presented in this chapter has several limitations. While the trained models demonstrate good results for detection and localisation, they do not distinguish between different types of defects. Although the overall aim is to expand training and testing on datasets that include various types of defects, such as porosities and inclusions, the current lack of available data prevents addressing classification tasks. Additionally, while multiple training runs were conducted to assess training stability, the performance metrics reported here were based on the best-performing model from each set. A more rigorous approach would involve reporting average performance across runs and including statistical measures of variability.

The AI models explored in this work represent a targeted subset of widely adopted high-performing object detection architectures from the broader computer vision

domain. While alternative architectures such as EfficientDet, different YOLO variants, and more recent transformer-based models (e.g., DETR [260], ViT [145]) could offer further performance gains, many recent developments in this space prioritise speed, scalability, or deployment efficiency rather than significant leaps in the detection accuracy. For example, Vision Transformers (ViT) are DL models that include the transformer architecture (originally developed for natural language processing) to apply attention mechanisms to object detection tasks. In theory, ViTs could offer improved generalisation for defect detection. However, in practice, they typically require very large datasets to outperform convolutional models [261]. Given the limited size of the datasets used in this work, such architectures were considered impractical. These challenges and limitations are further discussed in Section 7.3.

Furthermore, the primary aim of this study was not to exhaustively benchmark architectures, but to investigate whether object detection models, when trained solely on synthetically generated and noise-augmented data, could reliably generalise to real ultrasonic C-scans. It is acknowledged that a broader architectural comparison and more exhaustive hyperparameter optimisation could potentially improve detection performance, and these are seen as valuable directions for future work.

Lastly, it is important to highlight a broader limitation of the research landscape. In contrast to mainstream AI research, which benefits from standardised open datasets (e.g., COCO, ImageNet) and well-defined benchmarking practices, UT NDE research such resources and norms. Existing studies rely on proprietary or in-house data, making direct comparisons, reproducibility, and cross-validation of published results challenging. As a result, applying and comparing “state-of-the-art” ML approaches across different studies is inherently limited, as models are rarely evaluated on common data or under consistent conditions. Addressing this challenge represents a high-impact opportunity for the NDE research community.

The next chapter will expand on the application of AI, particularly in automated gating methods, as manually gating individual samples in this work slowed down the process and detracted from full automation potential. Furthermore, the unsupervised training method will be tested to compare its performance to the supervised models presented in this chapter, focusing on reducing the time-consuming generation of datasets and

associated ground truth labels by framing the defect detection problem as one of anomaly detection. The generation of ground truth labels in this study was labour-intensive and heavily influenced by the individual performing the task, making it challenging to label defects, particularly smaller ones, with precise and uniform accuracy. This subjectivity may have impacted the reported detection performance metrics, but it was not explored in this chapter. Several avenues can be pursued to address this issue. One approach is to involve multiple annotators and generate a consensus ground truth, which helps to remove personal bias (such approach was adopted in Chapter 5). Another direction would be to adapt active learning or weak supervised approaches, where a smaller subset of relevant data is manually labelled and used to train an initial model, which then assists in labelling the remaining data [262]. Lastly, in the case of training and validation datasets, some simulation environments do allow for precise generation of ground truth as defect location and size are predefined by design [263].

Chapter 5: Unsupervised Anomaly Detection for B-scan Analysis

5.1 Chapter Overview

Following the work presented in the previous chapter, several key challenges and potential for further research have been identified:

- The generation of C-scans required manual setup of time gates tailored to individual samples, particularly for samples of non-uniform thicknesses. Therefore, the development of an automated gating method to tackle this task is deemed promising to increase the level of automation in data analysis.
- The reliance on datasets with defects for supervised model training presents a challenge, as such datasets are not readily available, and manufacturing large numbers of defects is economically unfeasible. While simulation software was utilised in the previous chapter to address this problem, an alternative approach would be to frame the problem as one of anomaly detection. Ultrasonic data is often analysed by examining different views of ultrasonic data. Compared to the C-scan view, examining the data from a B-scan perspective allows for the creation of larger training and testing datasets.
- Ground truth labelling depends heavily on the precision and consistency of the individual performing the task. Marking defects, especially smaller ones, with pixel-level accuracy is challenging and can introduce variability in reported performance metrics. The impact of this variability was not explored in the previous chapter.

To address these challenges, this chapter presents a two-step defect detection process for ultrasonic B-scan data (explained in Section 2.1.12). The first step is an automatic gating process based on the DBSCAN algorithm, providing a robust and flexible method applicable to samples with non-uniform geometries. In the second step, a CNN-based AE model is employed to identify defective B-scans. The combination of outlined methods results in successful detection of 38 out of 40 embedded defects in samples A and B, and 22 out of 24 defects present in sample C, with 2070 B-scans processed in 1.26 ± 0.09 seconds. Lastly, a study of uncertainties was conducted to assess the impact of human labelling on the reported AI performance metrics. While the AE model was trained in an unsupervised manner using only pristine data, its

performance was evaluated against a labelled test set. Variability in human annotations, provided by three different NDE operators, influenced the reported detection metrics.

5.2 Contributions

This work presents a two-step defect detection framework for ultrasonic B-scan data of CFRP samples with complex geometries and embedded defects. The first step introduces an automated gating method leveraging the peak finding and DBSCAN clustering algorithms, addressing challenges associated with manual gate setup for samples exhibiting non-uniform thickness. This approach improves the scalability and robustness of preprocessing for diverse sample geometries.

The second step employs a CNN-based AE trained in an unsupervised manner on pristine B-scans, framing defect detection as an anomaly detection problem. This approach mitigates the dependence on large, defect-labelled datasets, which are costly and difficult to obtain. The combined method achieves successful detection of the majority of embedded defects across multiple samples while processing thousands of B-scans efficiently.

Additionally, the work examines uncertainties related to scan quality and repeatability, as well as the impact of inter-operator variability in ground truth labelling on performance metrics, highlighting challenges in defect annotation consistency and its influence on reported detection accuracy. Overall, this chapter advances UT data analysis by integrating automated data gating with unsupervised learning, offering a practical and efficient solution for real-world inspection scenarios involving limited defect data and samples with complex geometries.

5.3 Introduction

In practical industrial applications, the manual analysis and interpretation of UT scans typically start with a focused examination of sectioned C-scans. If these analyses reveal defects surpassing predefined area limits established by industrial guidelines, subsequent analysis of individual B-scans is performed to further examine such areas. In an industrially representative example, considering a pristine sample comprising approximately 4500 individual B-scans, an NDE inspector expends around 1.5 hours

to complete the analysis. In contrast, an additional hour or more is added to the process when defective areas are found. This additional time is allocated to the inspection of individual B-scans and the generation of quality reports. It is crucial to note that the individual inspection of every B-scan is unattainable within a reasonable timeframe. Therefore, the underlying idea behind the method proposed in this chapter is to serve as a supplementary tool for NDE inspectors, enabling the processing of all B-scans without incurring significant computational costs.

Autoencoders (AE) represent a category of ML networks that have found applications in the detection of potential security threats [184], denoising of data [155], and undertaking various NDE tasks [181], [264], [265], [266], among many others. AEs can be divided into three distinct components: the encoder, the latent space, and the decoder. The encoder's task is to process the original input data with a series of layers into a representation of features known as the latent space. The decoder part of the network aims to approximate the inverse of the encoder, taking the latent space as an input, and attempting to reconstruct the original inputs. AE as a concept has seen multiple iterations and improvements over the years, including: a) GANomaly [184] where additional encoder and discriminator structures are introduced; b) VAE [267] where the latent space is a statistical distribution; and c) U-NET models that adopt the encoder-decoder structure and are used for the segmentation of medical images [268]. In the scope of this study, attention is directed toward the utilisation of AEs as unsupervised defect detectors. The fundamental structure of a basic autoencoder is illustrated in Figure 46.

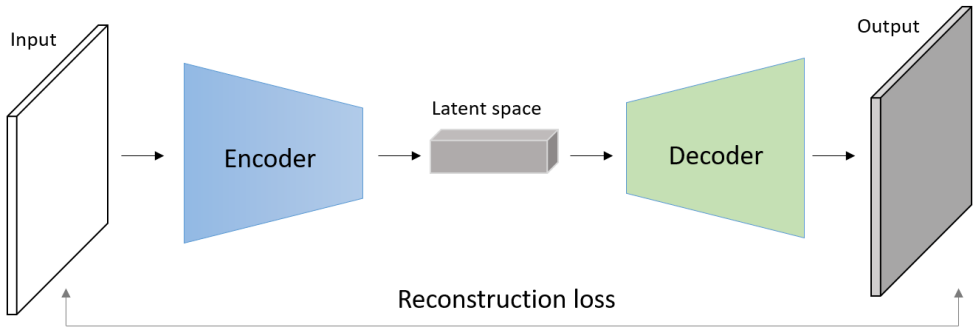


Figure 46 Basic autoencoder structure.

Anomaly detection is a method used in data science, statistics, and ML, primarily focused on identifying abnormal patterns or outliers within a dataset. Oftentimes data

follows regular patterns or can be approximated by specific statistical distributions. Anomaly detectors aim to capture these regularities while extracting outliers or data points that fall outside the probable distribution. The research of AI in the domain of NDE faces significant challenges due to data scarcity. Specifically, acquiring genuine defective data is often accompanied by costly manufacturing and testing procedures (as highlighted in section 4.1). A key risk of using artificial or manufactured defects is that they may not accurately represent the complex morphology, signal response, and variability of real defects. This can lead to over-optimistic model performance during training and evaluation, with reduced generalisability when applied to real-world inspection scenarios. AE-based anomaly detectors offer the advantage of being trainable and deployable in an unsupervised manner, diminishing the need for extensive acquisition and labelling of data containing defective samples. Instead, training can be based on undefective/pristine data, allowing the AEs to learn distributions and representative features of undefective samples. Their performance can then be evaluated with the reconstruction error observed between the input and output data, as undefective data should exhibit successful reconstruction with lower errors in comparison to defective data. In practice, PAUT systems are typically operated in automated setups where data acquisition occurs in real time as the probe is robotically moved across the material surface. At each scan position, a B-scan is generated based on predefined focal laws and sent to a controller unit, which buffers the data before transferring it to a PC. During deployment, each B-scan can be passed through the trained model on-the-fly, as inference times are negligible compared to acquisition rates. The reconstruction error is computed for each B-scan and stored in a corresponding error array. As the scan progresses, this error map is incrementally built up and can be visualised in real time or post-processed to flag regions of interest (those with elevated reconstruction errors) indicating potential anomalies.

In the field of NDE research, AEs have been previously used as anomaly (defect) detectors and denoising mechanisms. In [181], authors have developed a VAE model to detect defects within ultrasonic B-scans of bulk metallic materials. To enhance training and performance, taking inspiration from GANomaly, another encoder was added to the AE structure after the decoder. The study revealed that the model successfully identified larger defects but struggled with smaller defects that cause

minor reconstruction errors. Ultrasonic scans of rails were explored in [269] and [264]. In [269], AEs were deployed on a dataset comprising ultrasonic guided waves A-scans, achieving promising results. In [264], the authors have demonstrated that AE structures work well in the identification of different flaws that are visible in ultrasonic B-scans. The consistent geometry of inspected material made the application of signal gating easier, simplifying the dataset and positively impacting the final performance. In [266], authors focused on the ultrasonic dead zone, an area of an ultrasonic pulse close to the transmission source that can heavily mask reflections from near-field features. Positive results were achieved in the identification of near-surface defects, with the recommendation that the method be further validated on different material specimens and alternative defect types. In [265], through-transmission UT was used for the task of identification of defective adhesive bonds. Explainable anomaly scores were demonstrated as the sigmoid activation function was added to the calculated MSE between inputs and outputs of the model, presenting the anomaly score as a percentage. Overall, the authors have reported valuable quantitative results, but the scope of the study was limited. Lastly, in [155], AEs were utilised as a tool to denoise A-scans before classification. This method yielded great success as the classification performance was improved.

During the NDE of CFRP components, various signal processing techniques are used to improve the interpretability and quality of visualisation, with signal gating being crucial when preparing C-scan representations. Appropriate gating allows for the exclusion of surface and backwall echoes, which often have significantly higher amplitudes compared to potential volumetric defects, from the selected time window. This results in images that provide better visibility for lower amplitude features. The gating process is usually performed by a human operator while automatic gating is rarely discussed in the academic literature. In [270], authors developed a back wall echo filter method based on the computation of gradients of ToF variations and thickness tolerances. However, this method is incompatible with complex and stepped materials. In regions with abrupt changes (such as steps or sharp slopes) these gradients exceed the algorithm's tolerance and are mistakenly flagged as internal features, leading to false detections. Authors in [271] have introduced an automated UT analysis software that performs automated gating in two steps. The proposed approach is

complex and was tested only on samples with smooth and slight curvatures. The authors expanded the previous work in [272] by incorporating the Amplituden Laufzeit Ortskurven algorithm, characterising the front wall responses as echoes with the overall smallest ToF while other echoes correspond to defects, back wall, or repetitions. While yielding positive results, this method is not an all-round solution for gating complex geometry samples. In the study [273], authors achieved automatic gating on a per A-scan basis, by identifying amplitude peaks and assigning the leftmost peak label of the initial pulse and the rightmost peak the label of the back wall echo. However, this approach can misclassify echoes in the middle of the scan as defects in cases involving thinner samples with multiple backwall reflections. Lastly, in [274], the authors illustrated two adaptive gating processes for detecting defects: the first relies on the Computer-Aided Design (CAD) model, and the second is grounded in back wall echo tracking. Both approaches draw upon external knowledge, either through possessing a comprehensive CAD model of the inspected specimen or, in the latter method, by manually establishing a back-wall gate width that is larger than the maximum thickness of the component.

The scope of this study encompassed the examination of 11 CFRP samples of varying characteristics described in section 3.3. All but three samples were undefective, enabling the acquisition of a relatively large number of healthy ultrasonic scans that were used for training the ML model. Overall, 64 manufactured defects were examined in this study, and the used CFRP samples are summarised in Table 12. For a visual representation of the CFRP samples used, readers are referred to Chapter 3, Section 3.3.

Table 12 The range of defective/undefective CFRP samples used in this chapter.

ID	Dimensions [mm]	Thickness [mm]	B-scans total	Defective B-scans	Use
1	254.0 × 254.0	2.20	1000	N/A	Training
2	254.0 × 254.0	2.14	1000	N/A	Training
3	254.0 × 254.0	2.75	750	N/A	Training
4	254.0 × 254.0	2.75	1000	N/A	Training
5	254.0 × 254.0	4.25	1000	N/A	Validation
6	254.0 × 254.0	4.25	1000	N/A	Training
7	254.0 × 254.0	6.00	1000	N/A	Training
8	254.0 × 254.0	6.00	1250	N/A	Training
A	254.0 × 254.0	8.60	750	153	Testing
B	254.0 × 254.0	8.60	1150	239	Testing
C	780.0 × 197.0	7.50 – 16.0	2070	215	Testing

5.4 Machine Learning

Before conducting experiments, a preliminary small grid-based search was performed to assess the range of hyperparameters and potential architectures. To ensure compatibility between the encoder and decoder, a reflective padding technique was applied before passing the B-scans to the model. This adjustment expanded the input dimensions [batch size, 61, 1000] to the nearest multiple of 32, allowing for flexibility in the input sizes the model could accommodate.

The full network schematic is presented in Figure 47. The encoder part of the network consists of four convolutional layers, each coupled with a batch normalisation layer [142] and a hyperbolic tangent activation function. Convolution was performed with square kernels of size 7 and stride of 2, resulting in dimensionality reduction as each convolutional layer reduced the input size by a factor of 2. The integration of batch normalisation served to mitigate the risk of overfitting, whereas the hyperbolic tangent activation function was selected due to its alignment with the amplitude extremities of raw B-scan data after normalisation to front wall response, which ranged between -1 and 1.

For the decoder architecture, bilinear interpolation upsampling layers with a factor of 2 were integrated. This upsampling was followed by the inclusion of a convolutional layer, configured with a kernel size of 7, a stride of 1, and padding to preserve the spatial dimensions of the feature maps. This approach deviates from the utilisation of

transpose convolution blocks that are often used. The shift was induced due to the periodic artefacts that were observed in reconstructed images. These artefacts are a consequence of the overlap inherent in striding transpose convolutions, resulting in certain pixels being subjected to multiple passes by the kernel, while others receive only one pass. This phenomena has been observed in various works such as [205] and [206]. The decision to adopt the upsampling approach, as suggested in [277], has proven effective in mitigating artefacts, thus resulting in a clearer reconstruction.

The training process employed MSE as the chosen loss function, measuring the difference between the input and output images. During training, only undefective data was used for both training and validation. Measures to avoid overfitting also included the use of L2 regularisation in the form of weight decay which was set at 0.0001. For all trained models, ADAM optimiser was used [140], with $\beta 1$ and $\beta 2$ values of 0.9 and 0.999 respectively. To determine the length of training, early stopping with patience of 10 epochs was used on calculated MSE losses on the validation dataset. Lastly, a batch size of 64 individual B-scans was used, with a learning rate of 0.0005. During inference, each B-scan is passed through the trained autoencoder, and the pixel-wise MSE is calculated between the input and its reconstruction. Elevated reconstruction error indicates a deviation from normal (undefective) patterns, and this error is used to detect potential defects. In this chapter, no specific thresholding is applied to the reconstruction error, and performance is evaluated using ROC curves. Thresholding strategies are introduced and discussed later in Chapter 6 Section 6.6.1.

MSE was selected as the loss and evaluation metric primarily due to its simplicity and stable performance during training. While MSE can be sensitive to outliers and may not optimally emphasise small, localised anomalies, it provided a consistent global reconstruction measure across entire B-scans. Alternative loss functions such as SmoothL1 (a combination of MSE and MAE) were considered; however, it requires setting a transition threshold prior to training, which introduces additional tuning complexity. Mean Absolute Error (MAE) could also be used during inference, as both MSE and MAE would rank B-scans similarly based on reconstruction quality.

For this study, MSE was deemed sufficient given the objective was to flag anomalous B-scans rather than detect precise defect boundaries, and inference was performed over

full scan datasets. Importantly, since the task involves comparing input to reconstruction, multiple metrics could be adapted during inference without retraining the model. More localised or patch-wise comparison strategies could enhance sensitivity to subtle features and are highlighted as a potential avenue for future research.

All models were trained and tested using the desktop PC described in section 3.4. Full network architecture is presented in Figure 47.

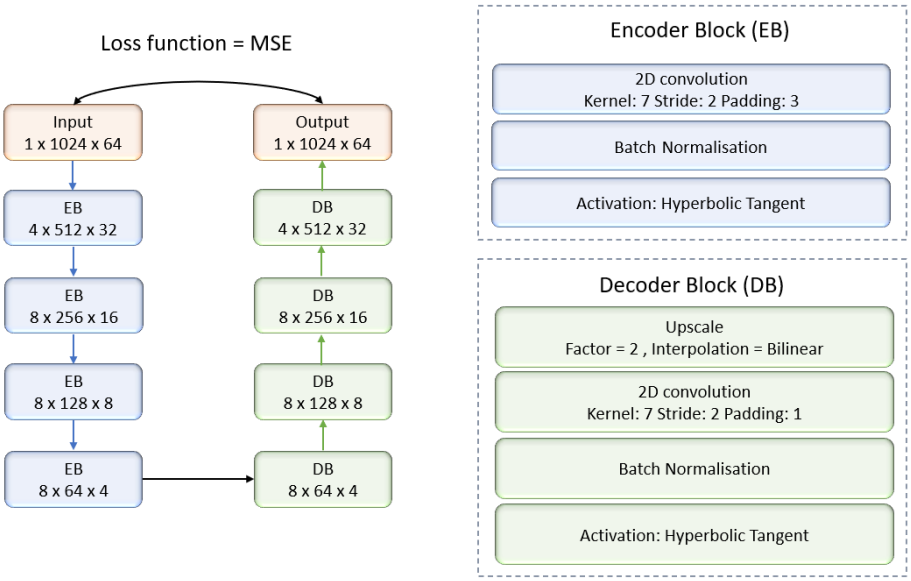


Figure 47 Autoencoder architecture used in this study, with details for encoder and decoder blocks.

There are several reasons why a custom autoencoder was developed. Currently, there are no publicly available AE models with pretrained weights specifically tailored for UT NDE data. While it is theoretically possible to adapt general-purpose AEs or variants designed for other domains, these models would require substantial architectural modifications to accommodate the extreme aspect ratio of B-scans and the normalised amplitude range used in this work (−1 to 1). For instance, GANomaly uses five convolutional layers in its encoder, which would reduce the input dimensions by an aggressive factor of 256, which may disregard smaller structural details in the data. Additionally, the bounded nature of input data does not align well with the use of unbounded activation functions in GANomaly (LeakyReLU and ReLU).

An additional avenue for exploration involves analysing the latent space representations learned by the AE. While the current approach focuses solely on

reconstruction error for anomaly detection, the latent embeddings may contain information that could be leveraged to enhance defect classification and/or clustering. For example, applying dimensionality reduction techniques such as PCA or Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [278] could help visualise and cluster B-scans, potentially revealing separable patterns between normal and anomalous data (or even between different defect types). Furthermore, with some labelled data, it may be possible to fine-tune or constrain the latent space to support supervised classification, thereby creating a hybrid approach that combines anomaly detection with defect type discrimination. Several studies have explored such strategies. For example, in [279], an AE is combined with a representation network designed to encourage a more meaningful and structured latent space to support downstream classification/clustering tasks. Study detailed in [280] uses AE latent space representation for classification of cancer types, while [281] uses both reconstruction loss and classification losses on the latent space for improved performance in metabolomic analysis. These directions were not pursued in the present study but offer promising opportunities for future work.

A key benefit of the outlined unsupervised method outlined is its lack of reliance on positive examples of defects during training. Evaluating B-scans against an expected defect-free reconstructing loss turns defect detection from a positive classification problem to an outlier identification/anomaly detection problem. While this method is unsuitable for defect classification, its strength lies in its robustness for defect detection. This is attributed to its training process, which did not involve the use of examples of defects, resulting in increased generalisability. To demonstrate the method's effectiveness, testing was conducted on various manufactured defects. While naturally occurring defect examples are limited, the generalisability of the method is expected to be robust for most defect types if the defect response yields an anomalous B-scan.

5.5 Automatic Gating Method

Following the data acquisition, prominent geometric patterns of higher amplitude were identified in front and back wall responses. The internal structure consists of much lower amplitude levels. When such data was used in conjunction with the AEs, it was

difficult to distinguish defective and undefective B-scans in the dataset based on reconstruction errors, especially for B-scans containing smaller defects (these observations are further discussed in sections 0 and 5.8). This complexity arose from the fact that defects which occupy small areas of a B-scan, whilst locally producing large MSE around the defect, would often be lost in the global reconstruction error which considers differences from all individual pixels. Similar observations were reported in [181] where authors have developed a VAE that provides good defect detection when encountering large defects, while sometimes failing to reconstruct undefective images or identify smaller defects. Furthermore, the authors state that geometrical signals usually have large reconstruction errors which in turn cause false positive indications. An approach that provided the solution to this was presented in [264], where constant geometry of the scanned specimens allowed researchers to effectively apply gating of the captured ultrasonic signals to remove the front and back wall indications, thereby reducing the complexity of the data. An example of the undefective/pristine B-scan and the defective B-scan are presented in Figure 13.

Although manual gating was possible, an automated gating setup would be beneficial as a range of samples of various thicknesses was used to generate training and validation datasets. In the aerospace industry, CFRPs are used for critical components such as fuselage, wing covers, spoilers, and stabilisers and they are manufactured in various geometries and material thicknesses to serve the functional purpose and meet the required performance criteria for different components. For instance, CFRPs are used for wing components with thickness variation, having thinner measurements towards the wingtip and thicker measurements near the root of the wing. For this specific application, thicknesses of around 24 mm are used [282] with wing spars reaching thicknesses of 40 mm or more [283]. An example of a complex geometry wing cover with stringers is illustrated in Figure 48.

Figure 48 CFRP wing cover component with complex geometry and varying thickness.

To this end, an automated gating approach that leverages the DBSCAN algorithm [284] was introduced. DBSCAN is a robust unsupervised ML clustering algorithm, characterised by two adjustable parameters: ε, which defines the maximum distance between a pair of data points for them to be considered neighbours, and *minimum_points,* a parameter specifying the minimum count of neighbouring data points necessary to form a distinct cluster. The proposed workflow is initiated by the definition of minimum amplitude threshold and minimum distance between the peaks used in the peak-finding algorithm. The utilisation of the peak finding algorithm was crucial to induce dimensionality reduction as clustering of the raw data is impractical due to the large number of data points, which results in poor outcomes. For this method, a full 3D data set was Hilbert transformed and processed at once. The minimum threshold amplitude used in peak finding was defined in Eq.27:

$$Minimum\ threshold\ amplitude = RMS(noise) * \alpha \qquad \text{Eq.27}$$

Where α signifies the scaling factor that adjusts the threshold in relation to the Root Mean Square (RMS) of the noise level. Next, the minimum peak distance is mathematically presented with Eq.28:

$$Minimum\ peak\ distance = \frac{f_{sampling}}{f_{operating}} * \beta \qquad \text{Eq.28}$$

Where $f_{sampling}$ and $f_{operating}$ correspond to the sampling and operating frequency of the ultrasonic setup, respectively. The quotient of these values represents the wave packet, which is scaled by a tuneable parameter β. The rationale underlying this formulation is to decrease the overall number of identified peaks. This strategy is useful to produce a separation between the identified peaks in the front wall and subsequent peaks produced by stronger subsurface reverberations. An example of such separation

between the clusters is presented in Figure 49, where front wall peaks are coloured in blue while subsurface reverberations are indicated in red.

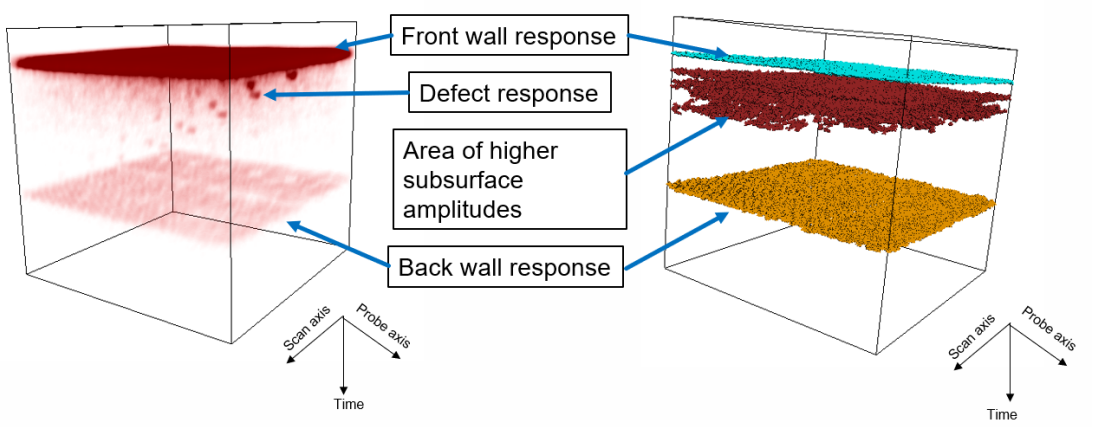


Figure 49 Example of the automatic gating progress; Hilbert processed 3D data (left), DBSCAN formed clusters (right)

The specified parameters were implemented in the peak finding algorithm sourced from the Python Library SciPy [252], which was used as a form of dimensionality reduction. Specifically, the algorithm identifies local maxima in a 1D signal by evaluating each data point against its neighbours within a defined window and applying threshold criteria for height, prominence, and distance. This allowed the most relevant signal peaks to be retained, effectively compressing the input representation. The resulting peak coordinates were then used as input to the DBSCAN clustering model. For flat samples, the two largest identified clusters corresponding to the front and back walls were chosen for the automatic gating process. However, in cases where samples exhibited complex stepped geometries, an additional user input variable *n* was introduced, representing the anticipated number of steps within the sample. This variable was used to extract produced indices of *n* number of clusters, starting with the one with the most points.

In both scenarios, a loss of back wall response might occur due to the presence of reflective defects, which in turn could result in no detections from the peak finding algorithm in the back wall areas. Therefore, clusters that belong to the back wall were checked for such occurrences and interpolated, similar to the work presented in [271]. To finish the process, the identified clusters generated indices indicative of areas for

exclusion during the gating process. This outcome yielded a scan focused exclusively on the internal portion of the material.

In experiments, a range of parameter values were tested to determine the optimal ones for the dataset used in the study. These parameters are listed in Table 13. Parameter α showed the largest influence on the overall results discovering that for thinner CFRP samples, all tested α values yielded satisfactory results. However, this was not true for the thicker samples, as higher α values generated a much higher threshold for peak detection and the back wall was not identified correctly. This was attributed to the high attenuation of CFRP materials (around 1.5 dB/mm), causing significant amplitude loss over longer acoustic paths. For β, higher values led to better separation between the front wall and subsurface reflections. For DBSCAN, smaller ε values produced better results, while higher values sometimes resulted in the unwanted merging of distinctive clusters. *The minimum_points* parameter had very little to no impact on the overall performance.

Table 13 Tuneable parameters in the automatic gating process

Parameters	Used in	Tested values	Selected value	Influence
$\boldsymbol{\alpha}$	Peak finding algorithm	3 – 10	4	High
$\boldsymbol{\beta}$	Peak finding algorithm	1.5 – 6	6	Medium
$\boldsymbol{\varepsilon}$	DBSCAN	7 – 12	7	Medium
minimum_points	DBSCAN	3 – 10	3	Low

This process offers a twofold advantage. Firstly, it serves as a valuable component within the NDE inspection workflow, facilitating the automated generation of C-scans by excluding the front and backwall echoes. Secondly, the process contributes to a reduction in data complexity in the interest of the anomaly detection process, focusing exclusively on the material's internal structure. This enhances the effectiveness of subsequent approaches in detecting defects with greater ease and accuracy. A step-by-step explanation of the automated gating workflow is provided below.

Automated Gating Workflow:

1. The full 3D ultrasonic dataset is Hilbert-transformed to extract the envelope of the signal.
2. Peak Detection (Dimensionality Reduction):
    a. A peak-finding algorithm is used to identify local maxima in each A-scan.
    b. Thresholds for amplitude and distance between peaks are used to reduce the number of candidate points.
3. Clustering of Peaks:
    a. The resulting peak coordinates are clustered using the DBSCAN algorithm.
    b. For flat samples, the two largest clusters (front wall and back wall) are automatically selected.
    c. For stepped/complex samples, a user-defined parameter n is used to extract n clusters corresponding to wall reflections at different thicknesses.
    d. Alternatively, for curved or irregularly shaped samples, n can be replaced with an area threshold, allowing more flexible identification of wall-related peaks.
4. Handling Missing Back Walls:
    a. If the back wall is not detected (e.g. due to high attenuation or strong reflection from a defect), the corresponding cluster is interpolated.
5. Apply Gating:
    a. Identified front and back wall indices are used to mask out the geometry-related signals, leaving only the internal structure for analysis.

A comparison of a defective ungated and gated B-scan is shown in Figure 50.

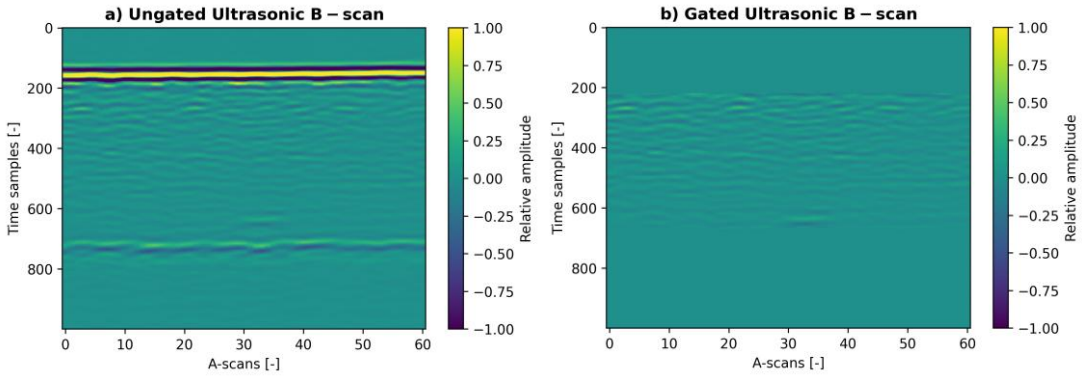


Figure 50 A comparison of a) Ungated ultrasonic B-scan; and b) Gated ultrasonic B-scan

5.6 Performance Metrics

In this study, a set of performance metrics to evaluate trained models were considered. Among these metrics, the False Positive Rate (FPR) and True Positive Rate (TPR) were used for assessing the accuracy of binary classification tasks and are given in Eq.29 and Eq.30 :

$$FPR = \frac{False\ Positives\ (FP)}{True\ Negative\ (TN) + False\ Positives\ (FP)} \quad \text{Eq.29}$$

$$TPR = \frac{True\ Positives\ (TP)}{True\ Negatives\ (TN) + False\ Positives\ (FP)} \quad \text{Eq.30}$$

FPR, denotes the ratio of undefective scans incorrectly identified as defective. On the other hand, TPR quantifies the rate at which models correctly identify defective B-scans as such. Together, FPR and TPR form the foundation for constructing the Receiver Operating Characteristic (ROC) curve, a graphical representation for binary classification tasks. Mathematically, the coordinates of ROC curve can be presented with Eq.31:

$$ROC\ curve = \{(FPR_1, TPR_1), (FPR_2, TPR_2), \dots, (FPR_n, TPR_n),\} \quad \text{Eq.31}$$

The ROC curve enables the visualisation of the trade-off between TPR and FPR at various thresholds. Furthermore, it enables the comparison between the presented methods with the AUC metric. AUC condenses the model's overall performance into a single scalar value and is given through Eq.32:

$$ROC\ AUC = \int ROC\ Curve(FPR, TPR) dFPR \quad \text{Eq.32}$$

The performance metrics used in this chapter differ from those used in Chapter 4 due to the difference in task formulation. While Chapter 4 evaluated object detection using spatial metrics such as IoU and precision-recall curves, this chapter addresses scan-level anomaly detection, for which ROC curves and AUC are more relevant.

5.7 Training and Deployment on Simple Geometry Samples

Training on the gated and ungated datasets yielded consistent inference results across repeated runs (meaning the models produced stable evaluation metrics regardless of random parameter initialisation). However, this consistency does not imply that the gated and ungated datasets performed equally as the models trained on ungated dataset consistently demonstrated better absolute performance, as detailed below. What varied between runs was the number of epochs required for convergence due to the stochastic nature of training and random initialisation of weights and biases. To quantify this, ten training repetitions on both datasets were conducted. In the case of training on ungated datasets, convergence was typically achieved after an average of 118 epochs, with a standard deviation of 62 epochs. Similarly, for gated datasets, the training process converged on average at 122 epochs, with a standard deviation of 77 epochs. The relatively high standard deviation highlights the neural network training's sensitivity to the initial weight and bias values. Nonetheless, it is important to note that while the random initialisation may impact the time taken for convergence, the performance of the models after convergence was not influenced, showing good convergence to a global minimum. An example of training convergence for both datasets is presented in Figure 51.

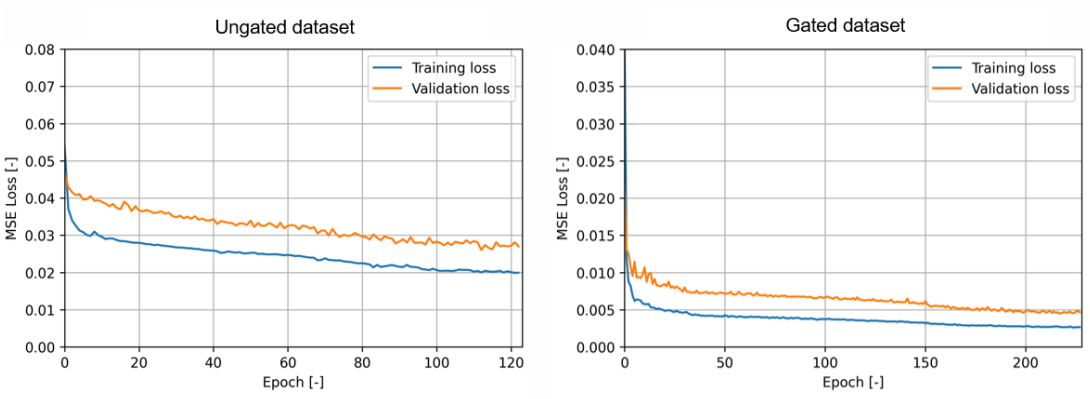
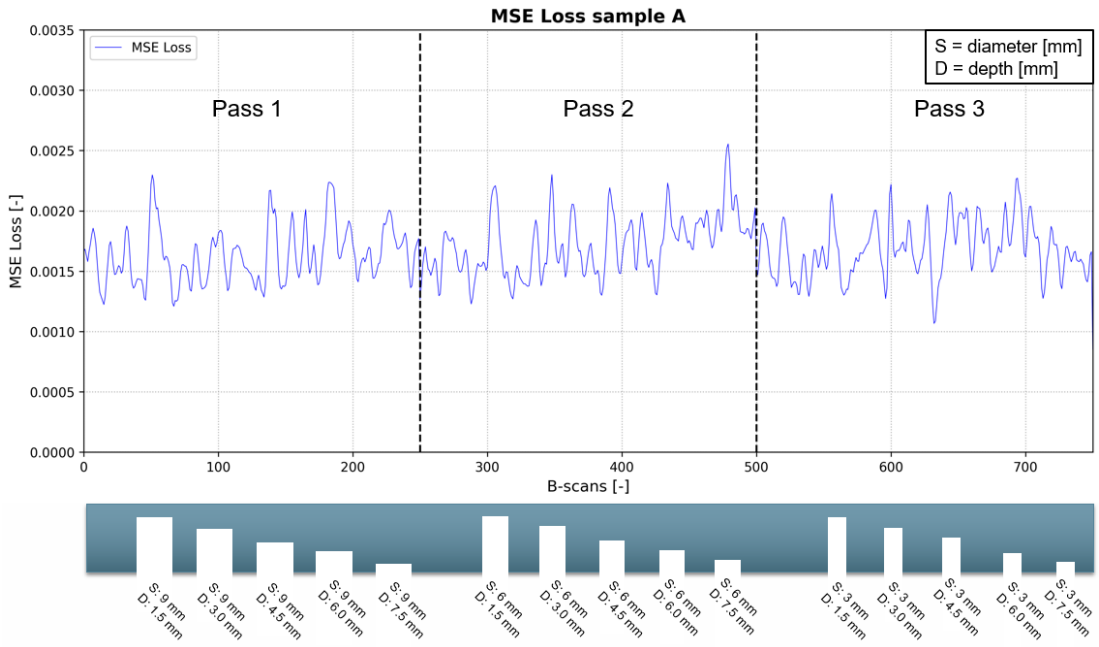


Figure 51 Training and validation losses for ungated and gated datasets.

When assessing the ungated data reconstruction losses for samples A and B during the inference, several observations were noted. For sample A, both defective and undefective B-scans were reconstructed with comparable reconstruction loss levels, giving rise to interpretation difficulty as the separation between data was not discernible. For sample B, slightly improved outcomes were observed as spikes in reconstruction losses for shallow defects, such as the 3.0, 4.0, 6.0, 7.0-, and 9.0-mm diameter defects were identifiable. However, the deeper defects remained undetected and were masked with higher reconstruction errors of undefective B-scans. ROC AUC scores of 0.763 and 0.863 were achieved for samples A and B, respectively. From an NDE application viewpoint, this performance would be deemed unsatisfactory as important defects would remain undetected. The reason for the poor performance was that the reconstruction error associated with the front wall and back wall echoes outweighed the error corresponding to the acoustic responses of defects in the internal part of the material. Reconstruction losses for ungated samples A and B are presented in Figure 52.



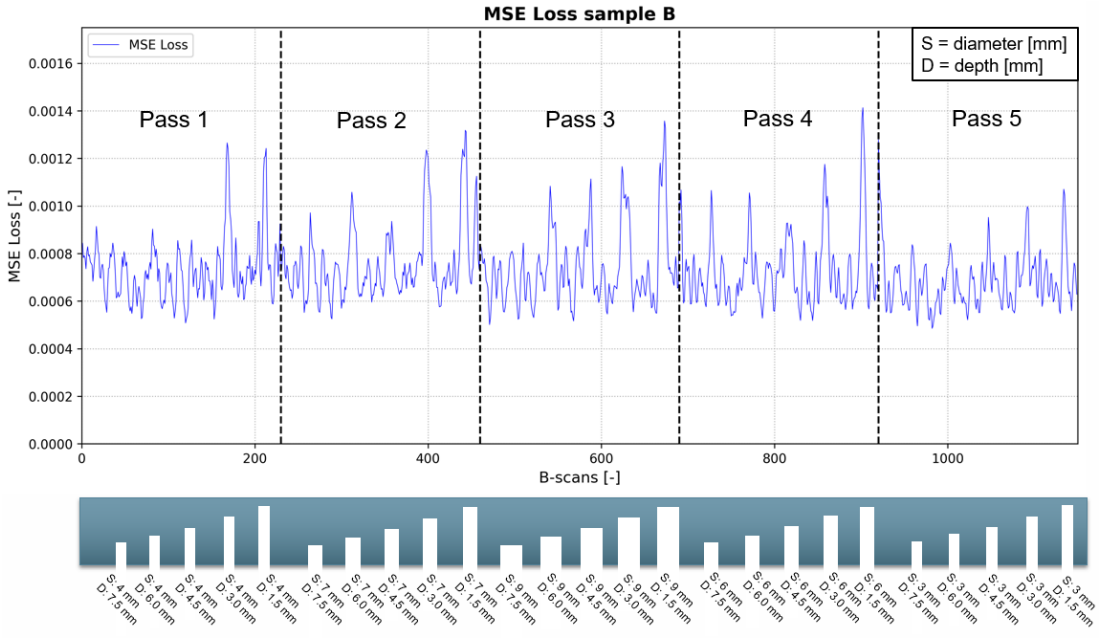


Figure 52 Reconstruction losses and side view schematics for sample A (top) and for sample B (bottom) for the ungated dataset.

When using gated datasets during training, as suggested by the results presented in Figure 53, major detection improvements were observed. In both samples, all 6.0, 7.0, and 9-mm diameter defects produced evident spikes in the reconstruction loss making them easily identifiable. Two 4.0 mm diameter defects located close to the front wall also produced elevated reconstruction error values, but three deeper defects produced very small deviations from the undefective B-scans. Lastly, 3.0 mm diameter defects produced the smallest reconstruction losses, with three defects closest to the surface producing small but visible spikes. The last two 3.0 mm diameter defects in both samples were not identified. ROC AUC scores were improved, with 0.920 and 0.922 for samples A and B respectively as compared to the ungated data, indicating a clear improvement. ROC curves are presented in Figure 54.

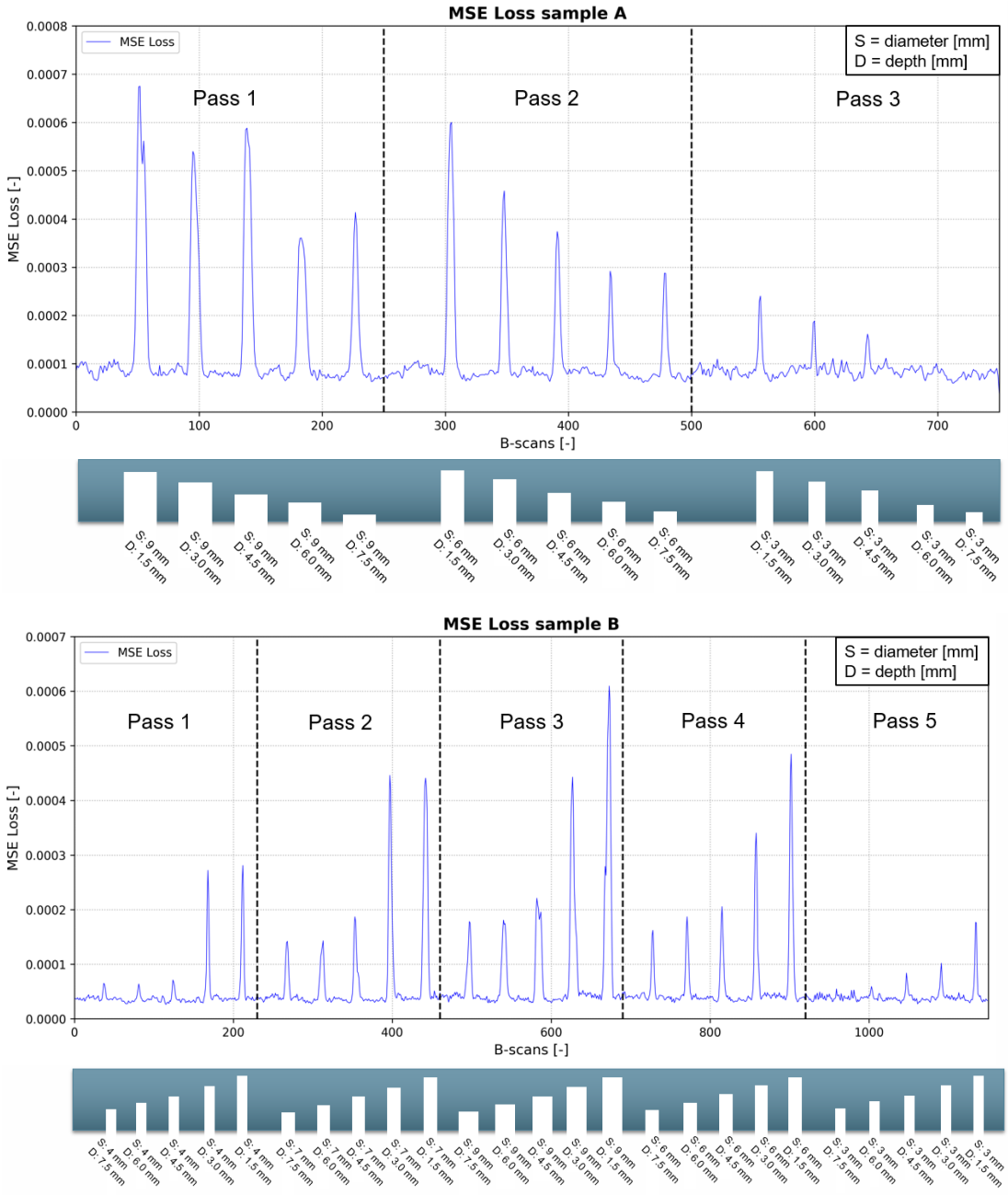


Figure 53 Reconstruction losses and side view schematics for sample A (top) and for sample B (bottom) for gated dataset.

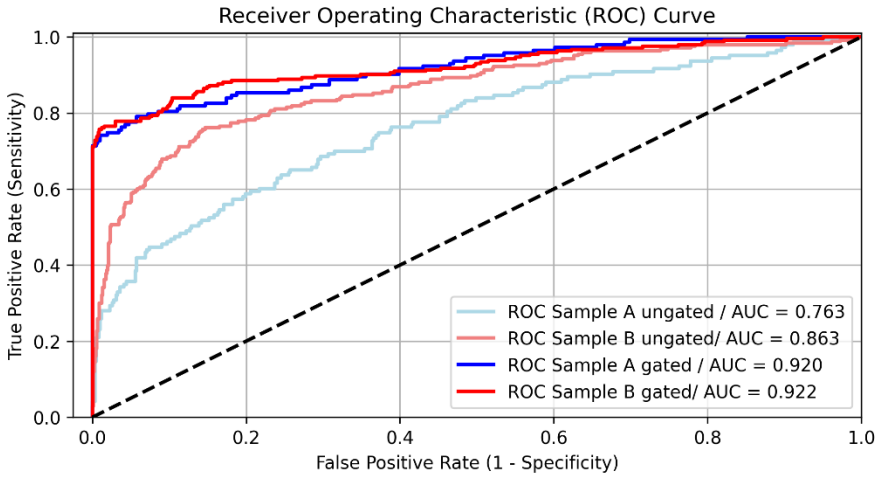


Figure 54 Receiver Operating Characteristic curve for samples A and B

For industrial application, the typical practice would involve establishing a threshold to distinguish between defective and undefective samples. This approach is flexible and can be adjusted based on the specific requirements of the industry. For example, in NDE of critical components where defect detection is imperative, the threshold may be set more conservatively, even if it allows for some false positive indications. A threshold can be defined using several strategies depending on operational requirements and the available data. One method involves tuning the threshold using a separate validation dataset, aiming to minimise false negatives while maintaining acceptable false positive rates, and then applying the selected threshold during testing. Alternatively, statistical techniques can be applied, for example by selecting a threshold based on a specific quantile (e.g. 99th percentile) of the reconstruction error, similar to the statistical thresholding approach discussed in Chapter 4. Another practical method involves computing the median reconstruction error across a scan and defining the threshold as the median plus a fixed percentage margin (e.g. 20 - 30%) to account for natural variation in defect-free B-scans. An example of such a threshold-based approach is demonstrated in Chapter 6. However, more advanced strategies involving localised analysis of reconstruction errors remain outside the scope of this work and are left for future research. While global metrics like MSE offer stable performance and consistent scoring across B-scans, they may overlook small, localised anomalies. Implementing a patch-wise evaluation could improve sensitivity to subtle defects, but would also introduce new challenges such as handling sparse data, as discussed in section 5.4.

5.8 Uncertainties Associated with the Repeatability of Ultrasonic Scans

When employing gated datasets, two distinctive characteristics were observed within the reconstruction losses. First, sudden, and large spikes in the reconstruction loss, have already been discussed as they pertain to defects. The second feature is the underlying reconstruction value of undefective B-scans, which tends to average around specific values with minimal variances. This was expected since undefective B-scans do not contain features that were not observed during training; therefore, these are reconstructed well with consistency. However, upon further analysis, it was found that the scan-to-scan error level for undefective scans is varied. There are two main reasons for this, the inference of the data that falls outside the distribution of observed training data, and mathematical implications due to the loss function. The former is represented by several factors:

- **Sample finish quality:** The training dataset contains high-quality scans performed on the samples with a similar surface finish. While some variance in surface finish is covered in the training data, significant deviations during testing can have a substantial impact on reconstruction errors.
- **Material anisotropy:** CFRPs have complex, anisotropic structures that can interact unpredictably with UT. Therefore, test samples may feature different macro and microscale properties not observed in the training data.
- **Variability within the equipment**: An inherent variance between the performance of individual array elements is present, as arrays are manufactured to operate within certain tolerances to pass the quality assessment by the manufacturer.
- **Variance in coupling:** Performed scans vary from each other due to changes in coupling conditions during scans. While the process is automated, the coupling dynamics during the scanning are unpredictable due to the manual application and contribute to fluctuations in reconstruction errors across different scans. This has a significant impact on the energy transfer which is discussed in the later example.

Mathematical variances are exhibited due to:

- **Type of reconstruction loss**: The chosen MSE loss calculates a summation of differences between all pixels. As a result, areas with higher amplitudes exhibit larger absolute errors compared to areas with lower amplitudes. This relationship is directly linked to coupling variance, as it is a significant factor in energy transfer during the scanning process.
- **Sample thickness**: Scans of thicker samples result in larger B-scans in terms of time samples, leading to greater reconstruction error. With a larger number of data points to be reconstructed, the likelihood of errors occurring in the reconstructed image increases. Consequently, the global sum of errors in thicker samples tends to be larger.

All the aforementioned occurrences lead to challenges in the repeatability of PAUT scans which in turn influences the performance of ML models. When the input data varies in quality, the model may misinterpret acquisition-related artefacts as structural anomalies or overlook genuine defects in poor-quality scans. This is especially problematic in unsupervised settings, where no ground truth is available to guide the learning process. As such, maintaining stable and repeatable scanning conditions becomes critical for human data interpretation and for the developed ML model.

In the experiments, coupling inconsistency exhibited the highest influence on the reconstruction errors. As the roller-probe tyre moves over the sample surface under a fixed contact force, it pushes out the water film which can cause inconsistencies in the coupling. The tyre's surface is slightly roughened by the manufacturer to retain water; however, this can sometimes lead to uneven wetting and improper coupling which would impede the path of the ultrasound beam. Such variations become apparent if an amplitude C-scan of the front wall is created, as displayed in Figure 55, where brighter areas indicate portions of the scan where higher amplitudes were recorded, compared to dimmer/darker areas of lower amplitudes corresponding to good and poor coupling, respectively. In the same figure, an example where the coupling between the PAUT and the sample was inconsistent is shown, where inconsistency creates a large variance in the reconstruction errors when it comes to undefective B-scans. The underlying reason for this issue was the excessive application of couplant onto the sample. In this specific scenario, reconstruction losses of certain undefective B-scans exceed the

reconstruction losses of defective responses, which negatively influences the final performance of the deployed model. While this behaviour is detrimental to defect detection accuracy in the current application, it also presents a valuable opportunity. Excessive and inconsistent reconstruction errors may indicate poor scan quality rather than structural anomalies. Such elevated errors could be symptomatic of acquisition issues such as poor coupling (e.g., dry spots), misalignment of the probe with the material surface, or inconsistent contact pressure. If the AE was deployed in real-time during data acquisition, it could serve as a quality assurance mechanism. Unusually high reconstruction errors could trigger alerts, prompting the operator to repeat the affected segment of the scan. This proactive feedback loop could be used as a tool to maintain consistent scan quality and is left for further research.

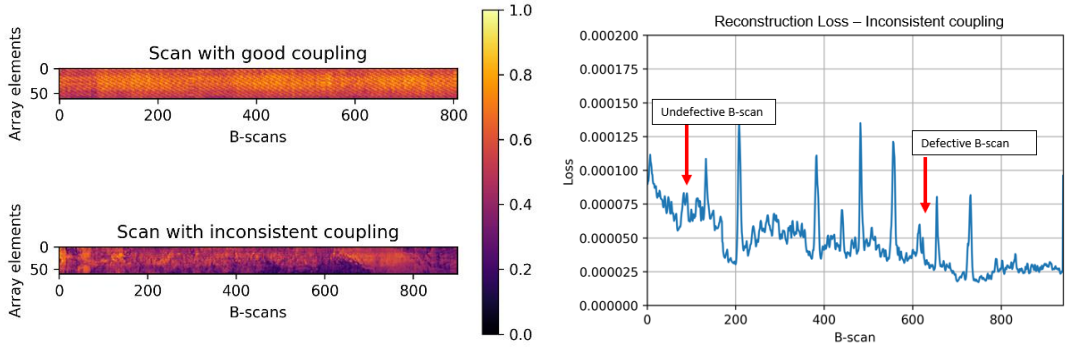


Figure 55 Amplitude C-scan comparison of scans with good and poor coupling (left) and resulting reconstruction loss from the scan with inconsistent coupling (right)

5.9 Uncertainties Associated with Human Annotations.

The process of labelling the test datasets involved three operators (PhD students and postdoctoral researchers with expertise in UT NDE) who were presented with visualised B-scans accompanied by corresponding scan indexes (positions where the B-scans were acquired). As each defect has breadth and depth parallel to the inspection surface, it can provide indications captured across consecutive B-scan frames. Therefore, the operators were tasked with identifying and marking the beginning and concluding indexes at which defects were believed to appear within the dataset. However, this task proved to be quite challenging, as the achievement of consensus among the operators was infrequent. Notably, disparities in the identified starting and ending indexes exhibited variances of up to 3 indexes. This discrepancy carries significant implications, as frames were captured at a robotic displacement of 0.8 mm.

The underlying cause for this challenge stemmed from the operators' approach, as they relied on observing sequences of B-scans to pinpoint defects. By navigating back and forth within these sequences, they searched to identify the precise starting and ending points of defects. This deviation from the way the AE model processes B-scans introduces a discrepancy, as the model doesn't operate sequentially or retrospectively review datasets to arrive at conclusions. To address this, future work could incorporate architectural changes that explicitly model sequence and context. One possible direction is the use of sliding windows of consecutive B-scans (such as the sequence approach proposed in [177]) to provide the model with immediate contextual information. Alternatively, sequential models such as LSTM architectures could be explored to capture longer-range dependencies across scan data. These approaches are well-suited to mimicking the way human operators evaluate defects over multiple frames and are therefore recommended for future research. An example of B-scan sequence where the individual defect is observable over several instances and a corresponding C-scan image with ground truths generated by operators is shown in Figure 56.

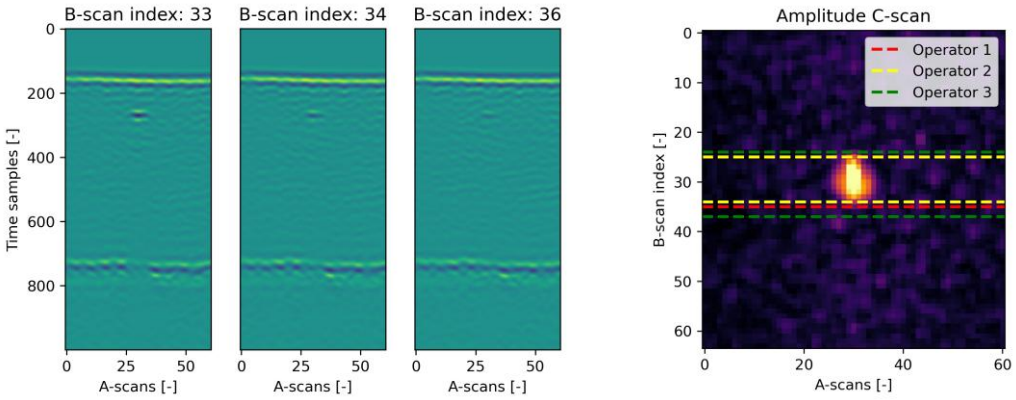


Figure 56 A sequence of B-scans with an observable defect (left) and a C-scan of the same defect with marked labels from each operator (right)

Given the lack of consensus for most defects, a pragmatic approach was adopted: labels from all observers were averaged, resulting in a composite ground truth. Outlined uncertainties associated with the labelling process, result in several observations. Firstly, it validates the significance of developing a robust automated approach for defect detection in the context of NDE, as ground truth was challenging to produce since the process heavily relied on the manual human interpretation of data. Secondly, it directly influences the reported performance of the deployed model (while

data and model remain unchanged) as ROC AUC varies up to 10.2% between different operators. To shed light on how the network's reported performance would be impacted by the labelling discrepancies of different operators, the ROC curves constructed by using different ground truths are presented in Figure 57.

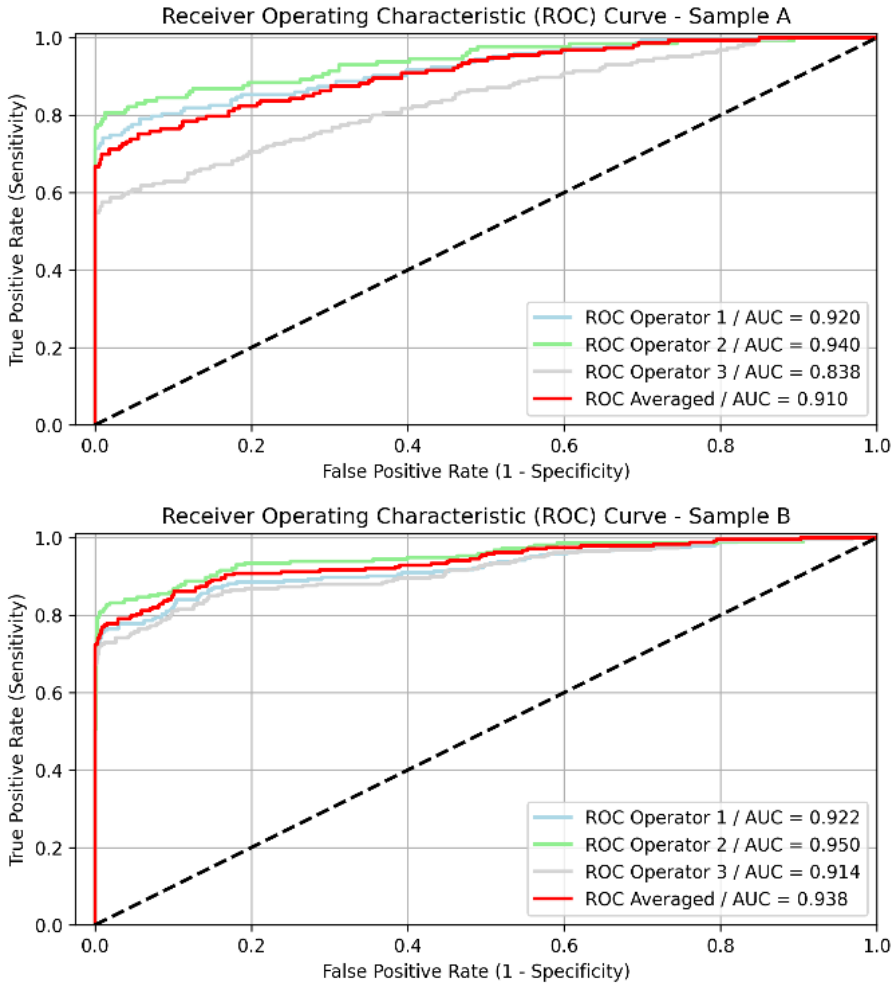


Figure 57 ROC Comparison with respect to ground truth produced by different operators.

The challenges in achieving consensus among operators reflect real-world practice. In many industrial settings, multiple NDE operators independently examine the same data without direct communication, and their findings are cross validated with any discrepancies investigated further. Such disagreements are therefore expected. Increasing the number of operators is impractical due to resource constraints and the specialised training required. As a result, developing automated or semi-automated solutions that support operators is a promising approach, with some industry initiatives

reportedly combining AI models with human expertise, though specific details are not publicly available.

5.10 Deployment on Complex Geometry Sample

To evaluate the automated gating method and the developed AE model, an additional testing dataset was created from sample C. The range of defects positioned close to the front wall, in the middle of the sample, and near the back wall serves as a challenging example to scrutinize the performance of the model. Furthermore, as the sample is stepped, it demonstrates the ability of the automated gating process to cope with sudden changes in the geometry. The sample's geometry and the output from the gating process are presented in Figure 58.

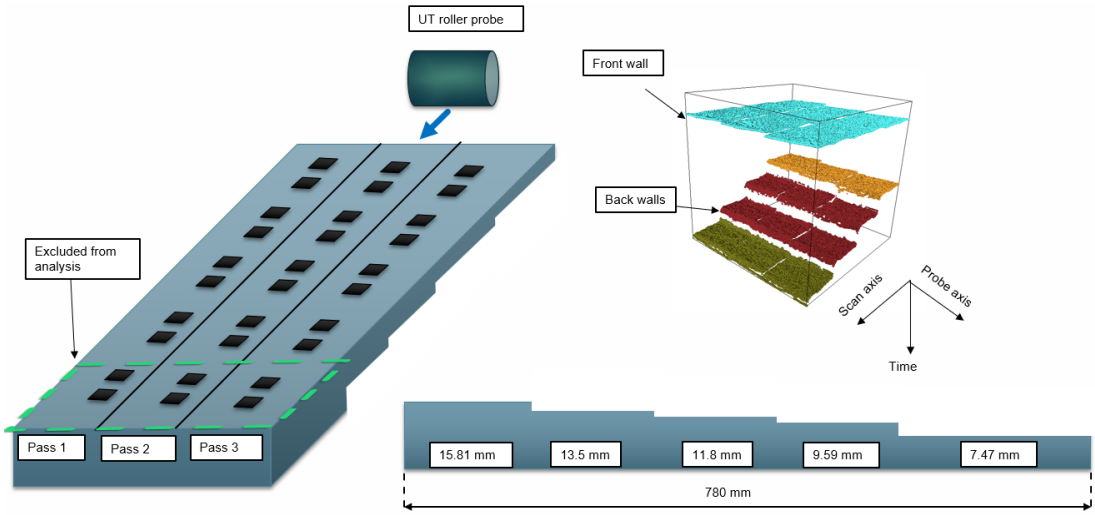


Figure 58 Schematic of sample C (left), DBSCAN output for automated gating (top right), and side view of the sample (bottom right).

After the data acquisition process, it was observed that the scan performed on the thickest part of the sample exhibited a visible repetition of the tyre reflection. This occurred due to the disparities in velocities of CFRPs (~3000 m/s) and glycol/tyre (~1638 m/s), causing the reception of the second tyre reflection before the first back wall echo from the thickest composite material section. This ultrasonic indication was a limiting factor for the successful extraction of unique clusters during the automated gating process. To this end, that part of the scan was excluded in this case study and altering of the ultrasonic setup to eliminate such reflections is left for future work. Potential solutions to this problem include using a larger roller probe diameter to

extend the acoustic path inside the tyre or using an alternative liquid filler material in the roller probe with a lower speed of sound. These options were not explored in this work, as they would have required the development of a custom probe. Instead, a standard off-the-shelf Olympus roller probe (widely used in industry and available only in one size) was employed to maintain relevance to current field practice. The scan was performed with three robotic passes over the sample with the results presented in Figure 59.

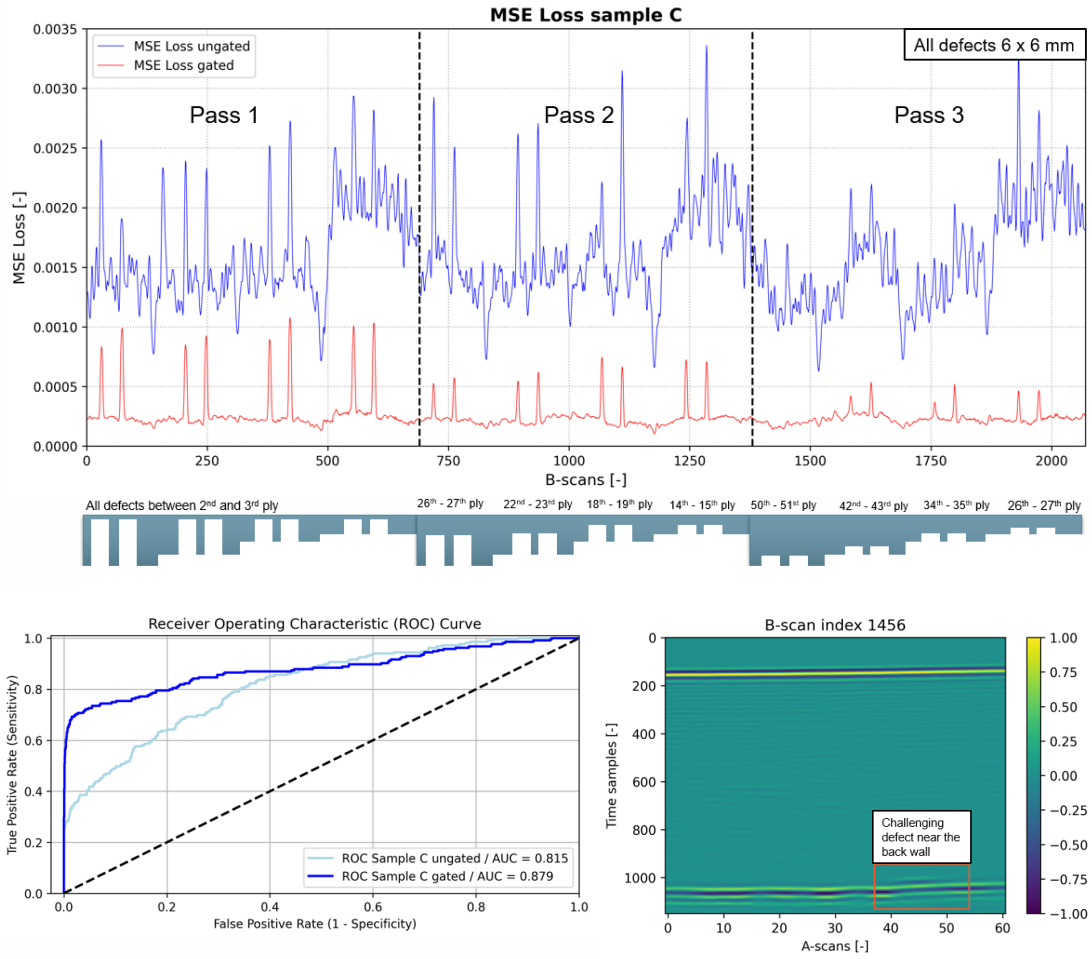


Figure 59 Reconstruction losses for gated and ungated sample C (top), ROC comparison (bottom left), and an example of a challenging defect (bottom right)

All defects in the first two passes were identified successfully. In the third pass, containing defects close to the back wall, two defects in the thickest section were missed. These missed defects present a current limitation of the proposed NDE workflow and the performance on these types of defects could be improved with a better ultrasonic setup and a more in-depth analysis of CFRP attenuation properties which in turn would result in a more effective TVG procedure. Furthermore, these

defects were challenging to observe both due to their position and low acoustic response. An example of such a missed defect is also included in Figure 59. The achieved ROC AUC on the gated dataset was 0.879, an improvement from 0.815 when the ungated dataset was employed. Overall, this case study presents a practical application with a realistic scan conducted on a complex geometry sample. The model's deployment results in fast inference, processing 2070 samples in 1.26 ± 0.09 seconds on a GPU-accelerated machine.

The rapid inference achieved can be attributed to the model's lightweight architecture. To further enhance inference times, exploring serialisation and saving models in environments better suited for production deployments, such as Open Neural Network Exchange or similar formats, rather than running them directly from Python scripts, could contribute to additional improvements in efficiency; however, this is outside the scope of this work.

5.11 Conclusions, Limitations, and Future Work

In this chapter, a two-stage defect detection method based on automated gating and unsupervised ML model was developed for analysing ultrasonic B-scans images of CFRP components. Unlike the past efforts at automated gating, the proposed method is agnostic to the geometry of the scanned sample, resolving limitations and heavy reliance on currently available methods on operators' fine-tuning. For this stage, unsupervised clustering through the DBSCAN algorithm was employed to isolate the front wall and backwall echoes of the scanned component to prepare the data in a way to maximise the model's performance in the next stage.

The subsequent stage featured an AE-based model, tasked with processing ultrasonic datasets captured using an automated NDE setup that emulates an industrial environment. The proposed approach is fully unsupervised removing the need for ground truth labelling of the B-scans. This saves expert NDE operators time in preparing training datasets while still achieving good detection performance. The study yielded several key findings:

- The AE performance on the ungated data resulted in unsatisfactory results as the separation between undefective and defective B-scans in terms of reconstruction error was not achieved.

- The DBSCAN-based automated gating has proved to be a practical technique, effectively extracting front and back wall indications from 3D datasets without geometric constraints, making it a strong candidate towards true automation of the interpretation process.
- Implementing the automated gating process significantly increased the performance of the AE-based defect detector. The ROC AUC increased from 76.3% to 92.0%, from 86.3% to 92.2%, and from 81.5% to 87.9% across the three different testing datasets.
- Overall, 36 out of 40 defects produced visible reconstruction error spikes in the simple geometry samples, and 22 out of 24 defects in the complex geometry sample. Inference on a GPU-accelerated machine was rapid, processing 2070 B-scans in 1.26 ± 0.09 seconds.
- The overall performance of AE models was significantly influenced by the consistency of conducted scans. This was controlled by the stability of energy transmission into the sample, a factor greatly influenced by coupling quality.
- Uncertainties stemming from variations in producing ground truth had a direct impact on the reported results, highlighting the potential advantages associated with robust automated systems within the NDE workflow.

Limitations of this work include missed detections of the smallest defects closest to the back wall of the samples. Due to the nature of analysing B-scans, in industrial situations, this could extend to missed detection of thin defects oriented parallel to ultrasound beam propagation. Furthermore, controlled scans were performed where the entirety of the defects were captured within the active aperture of the ultrasonic setup. If defects are not fully captured within the active aperture, there is a risk that the AE model may fail to flag the B-scan containing the defect as defective. As demonstrated in the results section, the model performs better with larger defects. When a defect is fragmented across two ultrasonic passes, it effectively gets split into smaller pieces. These smaller fragments may not have strong enough features on their own to be recognised as defects by the AE model, causing them to be missed. Therefore, fragmentation reduces the effective size of the defect seen in each ultrasonic pass, increasing the likelihood of false negatives. A potential solution to mitigate the risk of defect fragmentation is to apply the AE model multiple times to the same data,

simulating a sliding window approach by generating B-scans from the captured data. This would reduce the likelihood of defects being fragmented, thereby improving detection accuracy. The additional overhead would be minimal in terms of computational cost, though it would introduce slightly more complexity in the code implementation. Furthermore, while the automated gating method improved the overall results of the AE model, by removing the back wall of the scan valuable information that pertains to the loss of the back wall is lost, which is often used in the analysis performed by an expert NDE operator.

To improve the AE model, several architectural changes could be implemented and tested. For instance, adding another encoder, drawing inspiration from approaches like GANomaly, could be beneficial. Alternatively, incorporating and computing feature reconstruction errors, as demonstrated in [285], could be another avenue for enhancement.

The next chapter will focus on combining supervised and unsupervised approaches detailed in Chapters 4 and 5 to analyse B-scans and amplitude C-scans concurrently, aiming to increase the automation level of data analysis in NDE and to mimic the human operator workflows. Furthermore, as the academic works focused on the use of AI in NDE often lack proposed integration strategies, different scenarios with varying levels of data analysis automation will be examined in line with the NDE 4.0 paradigm. Lastly, this combined approach will be applied to an industrial CFRP sample (Sample E), providing a challenging scenario to evaluate the developed AI-based workflow.

Chapter 6: Multi-Model Aggregation Strategies for Data Analysis

6.1 Chapter Overview

NDE 4.0 represents the integration of recent advancements in robotics, sensor technology, and AI, transforming and automating traditional NDE in line with Industry 4.0 principles. Despite these advancements, data analysis in NDE is still largely performed manually or with traditional rule-based tools such as signal thresholding. These tools often struggle to effectively manage complex data patterns or high noise levels, leading to unreliable defect detection as examined in Chapter 4. Additionally, they require frequent manual adjustments to set appropriate parameters for varying inspection conditions, which can be inefficient and error-prone in dynamic or fast-paced environments.

In contrast, AI-based analysis tools have demonstrated improvements over traditional methods, offering greater accuracy in defect detection and adaptability to higher variability within captured signals. However, their adoption in industrial settings remains limited due to challenges associated with model trust and their “black box” nature. Additionally, practical guidelines for implementing AI tools into NDE workflow are rarely discussed, motivating this work to explore various integration strategies across different automation levels.

Three levels of automation are explored, ranging from basic AI-assisted workflows, where tools developed in this thesis provide suggestions, to advanced applications where multiple AI models simultaneously process data in a comprehensive analysis, shifting human operators to a supervisory role. Proposed strategies of AI integration into the NDE automation workflow were evaluated on inspection of the two most challenging defective CFRP samples C and E. These samples were considered the most challenging due to their complex geometries. Sample C featured five distinct thickness steps, while Sample E (an aircraft wing cover) included both varying thicknesses and integrated stringer sections. These features introduce additional reflections and signal variations, making interpretation of the captured UT data more complex compared to the uniform-thickness samples examined in this thesis.

Unlike manual inspections, which take hours for larger components, the proposed approach completes the analysis in 94.03 and 57.01 seconds for the two inspected samples, respectively. This performance is directly compared to manual analysis, as described in section 6.3. While a comparison to other AI-based methods would be more ideal, this remains a challenge in the NDE research due to the lack of publicly available datasets and models, unlike the broader ML research field where benchmarking and direct comparison is more established (see section 4.12).

6.2 Contributions

This chapter introduces a flexible, multi-model framework for UT data analysis based on the aggregation of complementary AI models presented in Chapters 4 and 5. Rather than relying on a single end-to-end solution, the proposed approach distributes decision-making across specialised models, enabling defect detection across multiple ultrasonic views (B-scans, C-scans, and full 3D volumes). The methodology was validated on two CFRP samples containing 36 embedded defects, acquired using a robotic inspection setup. By designing workflows corresponding to varying levels of model aggregation, the chapter demonstrates how combining model outputs can improve detection reliability, reduce inference times, and support explainability through cross-model validation. Furthermore, changes to the experimental setup were made to enable deployment of models concurrently during data acquisition.

Additionally, this work explores the trade-offs between automation, operator involvement, and system complexity across different inspection scenarios. It introduces an arbitration mechanism using a 3D self-supervised model to resolve disagreements between 2D models (AE and Faster R-CNN), effectively enabling scalable deployment of more expensive 3D model on lower-powered hardware. The chapter also examines the limitations of achieving higher levels of automation and discusses how model aggregation may serve as a practical intermediate strategy toward building trust and robustness in AI-assisted NDE systems.

6.3 Introduction

As mentioned in Chapter 5, current industry NDE practices in the aerospace sector begin with automated robotic sensor delivery and data acquisition. This initial stage is followed by data preparation, which includes signal processing techniques such as

frequency filtering, signal enveloping with the Hilbert transform [72], and signal gating. Next, NDE inspectors review segments of the C-scans, and if indications exceeding industry guidelines for allowable defect size or amplitude threshold are identified, the corresponding B-scans are further examined. Lastly, areas of interest are extracted for quality certification report creation. Automated robotic data acquisition for components like wing covers of midsize civil aircraft models typically takes around 40 minutes, with data analysis requiring a similar amount of time for pristine components. However, this step may be extended by an additional hour or more if artefacts and defects are detected. This additional time is allocated for further inspection of different views of the data, primarily individual B-scans around areas of interest, and the report generation process. The overall workflow is illustrated in Figure 60.

Figure 60 Standard NDE workflow in the aerospace sector.

Apart from defect detection, defect sizing is another critical step in the data analysis workflow. Current industrial guidelines for NDE inspection describe allowable defect sizes based on their type and location on the aircraft. For instance, in the case of delaminations, the largest allowable flaw area that would not be categorised as a defect range from 60 to 500 mm², depending on the specific location on the aircraft. Traditionally, defect sizing is achieved using the 6 dB drop method, where an operator manually moves the probe to find the maximum amplitude and then determines the defect boundaries by identifying points where the amplitude drops by 6 dB (i.e., to half of the maximum amplitude). This method allows for fine-tuned probe positioning, making it highly dependent on operator skill. A similar approach can be applied to automated PAUT testing. However, instead of manual movement, the PAUT array is manipulated using industrial robotics, significantly improving repeatability, precision, and scanning speed. Despite these advantages, the resolution for defect sizing is constrained by the fixed pitch between individual transducers and the predefined scan step.

Following detection and sizing, operators are tasked with categorising defects based on their physical properties, which are inferred from ultrasonic signal features. This classification step is critical in distinguishing between common defect types in CFRPs such as delaminations, porosities, and foreign object inclusions, each of which exhibits distinct patterns in ultrasonic data. This manual process is not only time-consuming and labour-intensive but also prone to inconsistencies as different operators may interpret the same dataset differently. The variability in human judgment introduces additional challenges in reproducibility and makes a fair assessment of performance difficult. The reliance on contextual judgment, global understanding of data, and external knowledge about the inspected components further highlights the complexity of the operator's role.

The above-mentioned tasks and workflow highlight the potential of automation in NDE data analysis, particularly in the aerospace industry, where large volumes of data are routinely handled. While data acquisition is predominantly automated, the subsequent stages of data analysis, defect identification, sizing, and classification, remain heavily reliant on NDE operators. In certain scenarios, basic automation tools can be used to analyse stable and well-defined signals. In [240], the authors introduce tools to assist with thickness measurements, detection of delaminations in areas with varying thickness, and evaluation of porosity content. These tools require human interaction to narrow down areas of interest and provide some input parameters, resulting in a reduction of analysis time by 70%. Another approach is presented in [271], where data analysis is based on a multi-step algorithm. However, for complex signals heavily influenced by geometrical features of components, overlapping ultrasonic echoes, or external factors such as poor scan quality, the use of advanced solutions is needed [22].

It has been demonstrated that AI models are capable of outperforming humans in certain tasks. The study detailed in [286] explored the capability of NDE inspectors to distinguish between real UT data and data created by generative AI models. The study concluded that artificial data is indistinguishable from real data, making it an ideal candidate for training future inspectors and for supplementation of training datasets for alternative AI models. In [21], the authors compared the defect detection performance of an AI model with that of three NDE operators. The results showed that

human operators made a larger number of false calls, while the AI correctly identified all defects present in the data. This trend extends to other fields as well. In [32], the researchers demonstrated that an AI model designed for analysis and diagnosis of three-dimensional optical coherence tomography data matches or exceeds the accuracy of medical professionals with years of experience. Similarly, the researchers in [287] leveraged an ensemble of AI models to outperform human experts in medical diagnosis based on medical sonography. Despite the highlighted advancements in AI models, data analysis in industry remains predominantly manual, with limited adoption of new AI-based automation tools. Two key reasons for this are a lack of trust in the models, which includes concerns from both industry users and regulators, particularly in safety-critical processes [288], and the "black box" nature of AI, where the reasoning behind decisions is obscured. This lack of transparency leads to greater risks in evaluating safety-critical components, as inaccurate predictions from an automated system could result in unpredicted catastrophic in-service failures. Therefore, while these studies confirm the potential of incorporating new AI tools into NDE workflows, advancing to higher automation levels will depend on building trust in these systems.

Definitions of automation levels vary across fields and applications [289]. In the context of NDE, the authors of [22] propose a taxonomy for the entire NDE process, categorising it into Classical NDE (Level 0), Operator assistance (Level 1), Partial automation (Level 2), Operational automation (Level 3), and Full automation (Level 4). In recent years, there has been a notable shift towards adopting automated solutions in NDE workflows, leveraging advancements in robotics, AI, and other technologies, recognised as NDE 4.0 [20], [22]. This transition aims to redefine the roles of human NDE operators, transitioning them to more supervisory positions where they oversee and address specific parts of the process, while automated systems manage the bulk of repetitive tasks. The overarching objective is to enhance efficiency while improving the precision and repeatability of the overall NDE workflow.

However, this evolution introduces several challenges. First, the increased complexity of automated systems can make troubleshooting and maintenance more difficult, as operators may need to develop new skills to manage these systems effectively. At the same time, the mental workload on staff is likely to increase [20]. Additionally, there is a risk of inappropriate reliance on automation, where tasks requiring human

judgment are delegated to machines, potentially leading to errors or oversights. A study detailed in [288] assigned NDE operators detection and sizing tasks using automated tools and found significant levels of both disuse (operators disagreeing with the automation when it is correct) and misuse (operators agreeing with the automation when it is incorrect). To address this, the authors recommend incorporating discussions on the limitations of automation tools into the training of new personnel. Furthermore, by providing reasons behind potential automation failures, operators can develop a more informed and appropriate approach to using these tools, while also building trust through direct experience with the technology. In the context of automation, the term "human-in-the-loop" refers to systems where human operators remain actively involved in decision-making processes, while "out-of-the-loop" refers to systems where automation takes over tasks with no direct human involvement. Over-reliance on fully automated systems can result in out-of-the-loop performance degradation, where operators lose the ability to identify system errors and perform tasks manually. Studies, such as [290], have highlighted that operators relying on automation tools have diminished manual task performance compared to those who perform tasks without automation. To address these issues, it is suggested that humans maintain a high level of control through periodic interventions, which can help minimise system failure rates [291].

Trust can be defined as subjective anticipation of future behaviour [292], often based on reported performance metrics on a subset of data used in the study. This approach shapes the human perception of trust, which is more effectively demonstrated through direct interaction with the model and observation of its decisions [293]. Some implementations leverage the human-in-the-loop method to enhance trust, where the human operator oversees and supervises decisions made by the AI, facilitating continuous improvement of the existing models. Such an approach was explored in [294] where humans collaborated with AI models to build trust and enhance accuracy. This was achieved by identifying anomalous instances of data, labelling them, and incorporating them into subsequent iterations of model development. Additionally, allowing the human operator to question and have control over AI predictions is another way to build trust in probabilistic models [295]. An alternative approach is to adopt explainable and interpretable AI [296]. Examples of explainable AI include

Shapley Additive Explanations (SHAP), which quantifies feature contributions to predictions by assigning each input feature an importance value based on its contribution to the model output [297]. Another example is Gradient-weighted Class Activation Mapping (Grad-CAM) [298], which visualises the feature heatmaps from CNN layers to highlight regions that influenced the model's decision the most. However, these strategies are rarely explored in the field of NDE, with the most notable works being [201], where the authors used a novel dimensionality reduction method to strengthen the explainability of the AI model used for the sizing of defects from UT data, and [202] which used Grad-CAM to demonstrate that the CNN learned similar important features when trained on real, GAN-generated, and other simulated data for defect detection.

While there is a clear need to increase automation in data analysis, and some progress has been made with traditional methods that offer significant time savings [240] [271], guidelines for the practical implementation of AI tools in NDE are often lacking. Moreover, existing research on the adoption of AI methods for analysing UT data tends to focus on a single ultrasonic view (refer to Sections 2.5, 2.6, and 2.7). This approach does not accurately reflect how human inspectors conduct NDE, as they utilise multiple views to form conclusions about the inspected material. Relying on only one view can also overlook the strengths of other ultrasonic views, which may be better suited for inspecting varied locations, and detecting different types of defects or features. To address these gaps, this chapter focuses on:

- **Proposal and discussion of automation levels in data analysis**, ranging from operator assistance level (Level 1) to full automation (Level 4), with a focus on integration strategies to minimise the risk of critical system failures.
- **Development of a comprehensive PAUT data analysis workflow** utilising three distinct AI models that analyse B-scan views, C-scan views, and full 3D volumetric data in a coordinated manner.
- **Presenting a case study** involving an automated robotic inspection system for PAUT of CFRP materials used in the aerospace industry. This case study examines two reference industrial samples with complex geometry using an

experimental setup that closely mimics industrial practices and employs industrial manipulators for accurate and precise measurements.

6.4 Data Analysis: Levels of Automation

6.4.1 Level 0: Classical NDE

Taking inspiration from the automation levels defined for the entire NDE process defined in [22], Figure 61 illustrates the proposed automation levels for data analysis. Data analysis at level 0 of automation corresponds to classical NDE, where the operator manually examines all data, performs preprocessing tasks, and makes decisions independently. This manual approach, although still widely used due to the historical industrial approach in training operators and reliance on individual decision-making, relies heavily on the operator's NDE expertise and judgment. While offering high traceability and explainability, it also results in longer data analysis times, higher operator workload, and increased likelihood of human-induced errors, particularly during prolonged repetitive tasks [20].

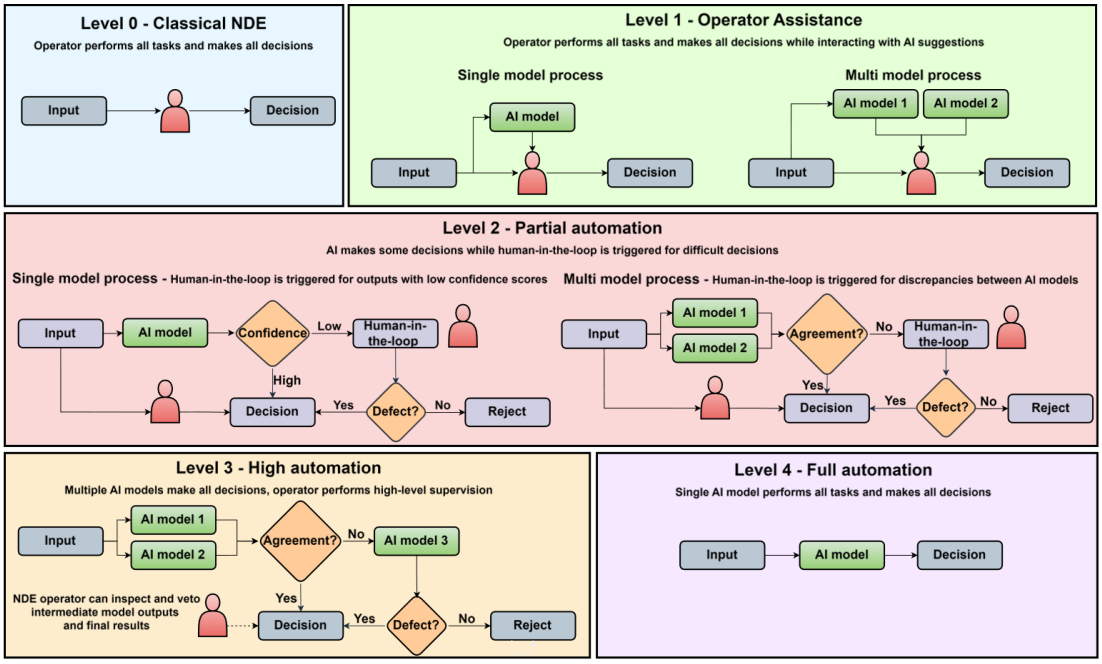


Figure 61 Proposed data analysis workflows for different levels of automation.

6.4.2 Level 1: Operator Assistance

At Level 1, operator-assisted data analysis, the human operator retains responsibility for all decisions and tasks. AI models assist by providing suggestions and highlighting areas of interest, but the final decision regarding the presence/absence of defects (sentencing) remains with the operator. However, the risk of failure at this level of automation is higher than at Level 0, primarily due to the potential for inappropriate reliance on automation, which could lead to misuse or disuse of the AI tools. Therefore, this level requires moderate trust in the AI models, which are expected to generate suggestions while focusing on minimising false negative (failing to detect an actual defect) calls to accelerate the analysis process. When used correctly, operator-assisted automation is ideal for gaining insights into scenarios where AI models may underperform without compromising the quality of the final NDE inspection. Additionally, this approach allows for continuous improvement by using those findings in future model re-training. Lastly, allowing operators to interact with the models, fine-tune inference parameters, and observe outputs during deployment could help build trust over time, as suggested in [293]. However, it is important to note that, as AI technologies are not yet widely implemented in the industry, NDE operators have not been trained in refining or adjusting AI tools. Therefore, additional training would be necessary for operators to confidently undertake this task.

6.4.3 Level 2: Partial Automation

Partial automation of NDE data analysis at Level 2 relies on a combined system of single- or multi-model processing with human-in-the-loop decision-making. For single-model setup, predictions with confidence scores above a set threshold are accepted automatically, while lower-confidence predictions are passed to a human-in-the-loop mechanism for further review. On the other hand, multi-model configuration involves two AI models collaborating to identify areas with potential defect indications, automatically accepting them if their decisions coincide, and activating human-in-the-loop decision-making to resolve any disagreements. This approach accelerates data analysis by focusing human intervention solely on resolving model discrepancies rather than manually processing all data.

The prerequisite for this level of automation is a high trust in the models to identify all defective areas while tolerating some false positives. False positives are managed through two mechanisms: first, by cross-verifying outputs between two detection models, which are unlikely to produce identical false positives, and second, by engaging human-in-the-loop decisions when models disagree. Overall, Level 2 of automation is characterised by faster data analysis and reduced human workload, albeit at the expense of higher system complexity and an elevated risk of failure. To prevent human-out-of-the-loop performance degradation, operators retain the ability to intervene and take control at any time. They can audit AI decisions and examine intermediate outputs from each stage, thereby improving both explainability and traceability.

6.4.4 Level 3: High Automation

Level 3 automation operates as a fully automated multi-model system, where a third, higher-precision AI model resolves disagreements between the initial two AI models. In this work, the two initial models are selected based on a logic of mirroring the manual approach taken by human operators, who typically examine C-scan data first to identify defects and then use B-scan data for further investigation. Therefore, the two models work independently on C-scan and B-scan data, and with their rapid inference offer a balanced combination of efficiency and accuracy for defect detection. The third model, which operates on full 3D volumetric data, offers the highest precision and is reserved for the final verification of areas where the first two models disagree. However, it is the slowest of the three and scales the least efficiently with increases in data size. Instead, it is selectively applied to specific sections, replacing the human-in-the-loop mechanism from Level 2.

It is important to note that different inspection scenarios may benefit from different combinations of AI models depending on requirements such as inference speed, precision, and explainability. This multi-model arrangement enables accelerated data analysis and reduced human workload, but introduces a higher risk of failure, necessitating high model trust and accuracy. At this level, used AI models must be scrutinised and fine-tuned for the specific application, aiming to achieve optimal accuracy with no tolerance for false negatives. Human operators, while removed from

direct involvement, transition to a supervisory role, retaining the ability to intervene, monitor, and override AI decisions, as necessary. This configuration delivers many advantages of an ideal automated system, albeit with slightly slower analysis and increased computational power required to run multiple AI models in parallel.

The hierarchical multi-model approach described here is conceptually generalisable beyond the specific UT inspection case presented. The core principle of using multiple models that independently analyse complementary data representations, with a higher precision but slower model reserved for the final verification stage can be adapted to many inspections where multiple data modalities or views exist. For instance, in other NDE contexts, models could be tailored to work on different sensor modalities and combined in a similar manner.

An example is in-process inspection of additive manufacturing components, which increasingly employs multi-modal inspection techniques using both eddy current and ultrasonic data [225]. In such scenarios, separate models can be trained on eddy current data and B-scan data, while a third model trained on full volumetric or C-scan data could serve as a verification tool.

However, the specific models, data partitioning logic, and integration strategy would naturally need to be adapted for each new application. Therefore, while the overall architecture and principles are broadly applicable, the implementation details and model choices remain application specific.

6.4.5 Level 4: Full Automation

Level 4 automation represents an idealised long-term goal where an AI model surpasses human capabilities in both speed and accuracy. In this setup, a single end-to-end model is responsible for all decision-making, eliminating the need for human NDE operators to inspect the data. While this approach would offer the fastest analysis, it comes with the highest risks and requires very high trust in the AI model, which can only be achieved through rigorous testing and parameter tuning. This level also represents an extreme case of automation, where human out-of-the-loop performance issues might arise. While this may be achievable and desirable in certain non–safety-critical industries, in safety-critical domains such as aerospace, Level 4 automation is

more realistically envisioned as an assistive tool used in combination with human operators. Achieving Level 4 automation would require overcoming a series of technical, regulatory, and trust-related challenges. Firstly, on the technical side, the model must demonstrate long-term performance that surpasses that of trained human operators across a wide variety of conditions, test samples, and edge cases. This includes generalisation to out-of-distribution scenarios not present in the training data, such as new defects, unexpected material variations, or acquisition artefacts. Additionally, some form of uncertainty quantification must be built in, to enable the model to recognise when it is uncertain or likely to fail. From a regulatory perspective, AI-enabled systems (like any other tool or a process) would need to meet defined standards, but current regulatory frameworks for such systems are still underdeveloped, as highlighted in Chapter 1. In this context, the path forward likely involves explainable AI approaches, where the reasoning behind decisions is interpretable and auditable. Lastly, in terms of trust, widespread adoption would depend on building confidence among operators, engineers, and regulators that the AI system is both reliable and predictable. This trust would have to be earned through extensive validation, transparency in decision making, and demonstrable alignment with expert human judgement over long periods of time. This is further discussed in the conclusion of this chapter.

It is important to note that different inspection scenarios may benefit from different combinations of AI models depending on requirements such as inference speed, precision, and explainability. As automation levels increase, several key system characteristics change. Higher automation levels lead to faster analysis speeds, with significant reductions in human workload. However, this comes at the cost of increased risks and system complexity. Lastly, trust in the AI system becomes crucial at higher levels of automation.

The lower risk associated with human operator performance stems from their ability to demonstrate inspection competency through rigorous training and testing. This acquired expertise is expected to generalise to out-of-distribution cases, as it is based on fundamental principles rather than solely on pattern recognition. In contrast, AI-based approaches often struggle with out-of-distribution scenarios, leading to higher

inspection risks. However, the risk level for human operators is not fixed and varies significantly depending on individual skill and experience.

6.5 Data Stream Handling

To this point, the ML models presented in Chapters 4 and 5 were evaluated in an offline setting (i.e., after the ultrasonic data had already been acquired and stored). However, deploying these models concurrently with data capture can improve overall analysis speed by reducing the time between acquisition and data interpretation. To explore this capability, the following section presents a practical implementation of real-time data acquisition, processing, and ML inference, incorporating several modifications to the experimental setup introduced in Chapter 3. A simplified schematic of the system architecture is provided in Figure 62 to illustrate the data flow and interaction between components.

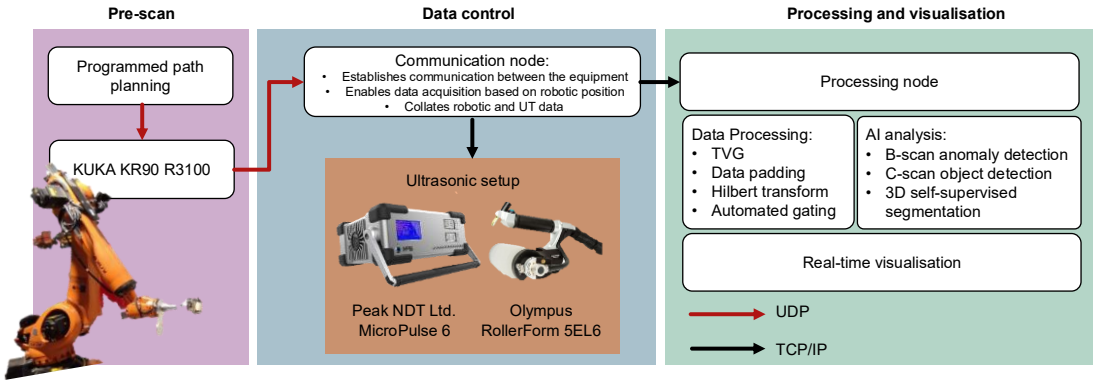


Figure 62 Flowchart of the experimental setup, integration of PAUT and robot, and data flow.

The data processing, data capture, and synchronisation between individual hardware elements were performed on a laptop setup described in Section 3.4 . Robotic control was executed with JAVA code wrapped in Python syntax, while UT and AI processes were performed in Python 3.8. Data acquisition and processing were split into two Python nodes.

The acquisition/communication node first sets up a TCP/IP connection to the UT equipment and listens for the User Datagram Protocol (UDP) connection established by the KUKA robotic controller. Once all connections are active, another TCP/IP connection to the processing node is initiated to maximise the utilisation of available computing resources. In Python, the global interpreter lock restricts true

multiprocessing and parallelism within a single interpreter process, therefore running two separate scripts concurrently allows the scripts to utilise different CPU cores effectively. Alternative programming languages such as C++ offer more straightforward solutions for parallelism but would result in more complex code and make implementation of the developed AI models more difficult. Another potential solution includes using Robotic Operating System (ROS) framework which is optimised for real-time applications.

Upon establishing the connection to the processing node, the robotic controller begins monitoring and broadcasting its' positions. The communication node continuously checks the Euclidean distance between subsequent position updates, triggering the UT data capture command if it surpasses the predetermined distance threshold of 0.8 mm (*i.e.* scanning step used in this study). The threshold aligns with the pitch of the used PAUT assembly, ensuring a square aspect ratio in the final data representation. Upon receiving the data, the robotic positions and UT readings are correlated and transmitted immediately to the processing node, repeating the process until the scan is completed. The processing node, upon receiving the data, performs basic data manipulation, including reshaping, normalisation, data padding, TVG, and Hilbert transform, before feeding the data into AI models.

Additionally, it is important to address the limitations and set reasonable expectations for the positional triggering setup. The current configuration, with a UDP connection between the KUKA controller and laptop, provides a positional update rate of 250 Hz. This update rate may present challenges at higher scanning speeds, potentially resulting in positional overshooting for data capture triggers. In the conducted experiments, scanning speeds of up to 30 mm/s were tested and deemed satisfactory.

Another critical aspect to consider is the resolution of ultrasonic scans. In the aerospace sector, the primary objective of NDE is to detect defects classified as critical based on their size and location on the structure. Quality control documents from Spirit AeroSystems indicate that delaminations ranging from 60 to 500 mm^2 may be allowed, depending on their position within the structure. When converted to equivalent circular defects, these areas correspond to defect diameters ranging from 8.8 to 25 mm. Given

this context, acquiring data at intervals of 0.8 mm ensures at least five frames per defect are captured.

6.6 Artificial Intelligence Models

6.6.1 Anomaly Autoencoder Model

The first AI detection model and automated gating workflow used in this work are detailed in Chapter 5. Modifications to the automated gating method include removing the previously used number of steps (n) parameter, as testing showed that imposing a minimum cluster size was more effective for identifying clusters corresponding to back walls. Previously, n defined the expected number of echoes in the scan, thereby introducing prior assumptions about the part geometry. In contrast, the use of minimum cluster size threshold (required by the DBSCAN algorithm) removes this dependency and allows the algorithm to adaptively detect meaningful clusters associated with geometrical features, regardless of how many are present. This represents a more scalable and robust approach, especially when scanning components with varying geometry. Moreover, a suitable minimum cluster size can be reasonably estimated based on the known scale of scanned CFRP components or inferred from the robotic path planning.

Surface echo is removed automatically due to a constant known offset of the inspected sample from the ultrasonic array, determined by the roller probe's outer diameter. This deviation is justified in well calibrated and consistent scanning setups, where the front wall echo is stable and predictable. If probe misalignment occurs, the more flexible clustering-based method from Chapter 5 could be reinstated for front wall removal. This adaptability makes the overall approach suitable for a range of inspection scenarios.

The peak-finding algorithm employs a normalised amplitude threshold of 0.25, chosen within the range of 0 to 1, and requires a minimum distance of 5 time samples between peaks to filter out minor peaks, thereby reducing data dimensionality and processing times. This approach deviates from the one used in Chapter 5, where an RMS-based thresholding method was applied. During testing, it was observed that the computed thresholds consistently fell within a narrow range (0.2 – 0.25), so a fixed value of 0.25 was adopted for simplicity. The risk of missing meaningful defect signals is minimal,

as observed delaminations tend to produce amplitudes well above this threshold (especially after TVG correction and Hilbert transformation). For more sensitive or variable signal conditions, the threshold can be adjusted, or the original RMS method reintroduced.

For DBSCAN clustering, the eps value is set to 7, consistent with values from the original publication. The *min_number_of_peaks* parameter is set manually at 250 to ensure defects up to 20.0 × 10.0 mm are included while excluding irrelevant clusters. Alternatively, this threshold can be adjusted automatically by analysing the size of the captured data and estimating the expected changes in material thickness.

The AE structure remained the same, and the discrepancy between the input and output is quantified using MSE. To differentiate pristine from defective B-scans, an anomaly threshold is applied to the observed MSE errors. A single threshold is applied across all automation levels, set as the median value of all observed MSE errors, increased by 50% of the median value. This approach is based on the expectation that most B-scans in the scanned sample are pristine. As a result, the median value of MSE will effectively represent the typical value for pristine B-scans, while the additional offset helps capture only significant deviations, accounting for smaller variations. The fixed threshold used in this chapter represents a simplified approach, chosen to reflect realistic deployment constraints in which ground truth labels are not available during inference. While the selected threshold is adjustable, its main purpose here is to support demonstration of full system functionality under automated conditions. More rigorous approaches to threshold selection are acknowledged and recommended for future work. In particular, using a separate validation sample to define and optimise the threshold (discussed previously in Chapter 5, Section 5.7) is seen as a promising avenue for more robust deployment. This method is consistent across all levels of automation, prioritising the safety considerations and detection of defects while accepting a higher rate of false positives. It is worth noting that this threshold can be adjusted based on the specific application scenario, and could also be defined using other statistical methods, such as standard deviations.

6.6.2 Object Detection Model

The second model utilised in this work is the Faster R-CNN model detailed in Chapter 4. During deployment, Faster R-CNN requires a confidence threshold to filter generated predictions (value between 0 and 1). Following the same logic for setting anomaly thresholds for the AE model, confidence thresholds are set at 0.001 for all automation levels. The inference process begins with the generation of amplitude C-scans using the automated gating method described in Section 6.6.1. These C-scans are then fed into the Faster R-CNN model, which outputs bounding boxes that highlight the defects within the inspected material. Compared to the AE model, the Faster R-CNN offers superior detection performance and provides the ability to precisely locate defects in the inspection plane.

The primary drawback of both AE and Faster R-CNN models is their "black box" nature, where the reasoning behind inference results is obscured. In industry sectors requiring clear, interpretable outputs, this is a disadvantage, as it limits transparency in the decision-making process. Furthermore, since the model was trained on 64 × 64 resolution images, it can struggle when processing input images with significantly different aspect ratios or sizes. To overcome this challenge, a workaround involves applying the model on smaller sections of the scans and then collating the results. Although this slightly complicates the deployment of the code, it allows for efficient inference and reliable defect detection.

6.6.3 Self-supervised model

The third model was a 3D Ultrasonic Self-Supervised Segmentation (3-DUSSS) model designed to process full 3D volumetric data, as presented in [299]. This lightweight model operates by pre-training on pristine 1D scan series through the component, where the model attempts to predict the likely distribution for the next value in the sequence. The model was trained to minimise the Negative Log-Likelihood (NLL) loss for the Weibull distribution, as shown in Eq.33:

$$Weibull\ NLL\ Loss = -\log \prod_{i=1} f(a, b|x_i) = -\sum_{i=0}^{n} \log f(a, b|x_i) \qquad \text{Eq.33}$$

Here, a and b represent the scale and shape parameters, respectively, of the two-parameter Weibull distribution predicted by the model for each input x_i.

During inference, the model predicts the parameters of a Weibull distribution (scale *a* and shape *b*) for the next point in a 1D scan sequence. These predicted distributions are then compared against the experimentally measured values to identify anomalous voxels. A datapoint is flagged as defective if its observed value falls in the tail of the predicted distribution, based on a specified confidence threshold (e.g., a false-call rate of 0.001%). The model utilises a sliding window approach (a fixed-length window is moved sequentially across the 1D scan series), whereby if a point is considered defect free it is added to the series to ground the model in relation to experimental data. If the point is considered defective, the model uses the mean of the predicted distribution as a best proxy for the expected defect-free datapoint and flags the voxel as defective.

Similar to the AE model, training was performed on pristine data, allowing the 3-DUSSS model to learn the amplitude responses specific to carbon fibre structures. To ensure no data leakage, the training, validation, and testing datasets originated from different physical samples. The training dataset included both front and back walls, which minimises the impact of poor gating which could lead to removal of defect signatures. During inference, the model requires two parameters: the allowable false call rate and an area threshold. The allowable false call rate defines the maximum deviation a voxel can have from the predicted distribution before being considered defective (in this work, this was set to 0.999999). The area threshold filters out smaller voxel groups to minimise false positive calls, with this threshold set to 10 in the current work.

The developed model excels in localisation, depth estimation, and sizing of defects, effectively detecting flaws as small as 3.0 mm in diameter. It offers voxel-level precision, allowing for accurate defect sizing and depth localisation, unlike the AE model (which does not offer spatial localisation) or Faster R-CNN (which tends to overestimate due to bounding box geometry). Furthermore, it enables volume-based representation, which is particularly valuable in NDE applications, where understanding the full defect geometry can influence repair decisions. The 3DUSSS method was found to perform well across the tested datasets by successfully detecting all known defects in samples C and E, as confirmed in earlier publications [299], [300].

However, when processing large datasets, this method encounters challenges due to the computational demand of handling the entire scan volume, requiring a powerful GPU with significant memory capacity (the original study employed a setup with three NVIDIA GeForce RTX 3090 GPUs). Although the model itself is lightweight, the volume of data for processing is substantially higher than that of individual B- or C-scan views, making GPU memory a critical factor and creating a bottleneck in data loading onto the GPU. Even after down-sampling data by a factor of 10 in the time domain, deployment on less powerful hardware, like the single GPU configuration used in this study, leads to processing times in the range of several minutes, far slower than the few seconds needed by AE and Faster-RCNN. Furthermore, the scans in this study are relatively small compared to those typical in industrial settings for large components, where AE and Faster-RCNN would likely scale better, as 3-DUSSS must process the entire dataset, while other methods operate on compressed 2D views. Additionally, 3-DUSSS faces challenges when encountering variations in thickness, making it more suitable for deployment along scan directions where thickness changes are minimal. However, despite being slower at inference than other methods, 3-DUSSS's capability to generate a complete 3D segmentation map provides a comprehensive visualisation of the ultrasonic scan. This feature not only improves the interpretability of scan results but also allows for the creation of digital twins for reporting, adding practical value to the inspection process.

6.7 Results and Discussion

6.7.1 Level 1 – Operator Assistance

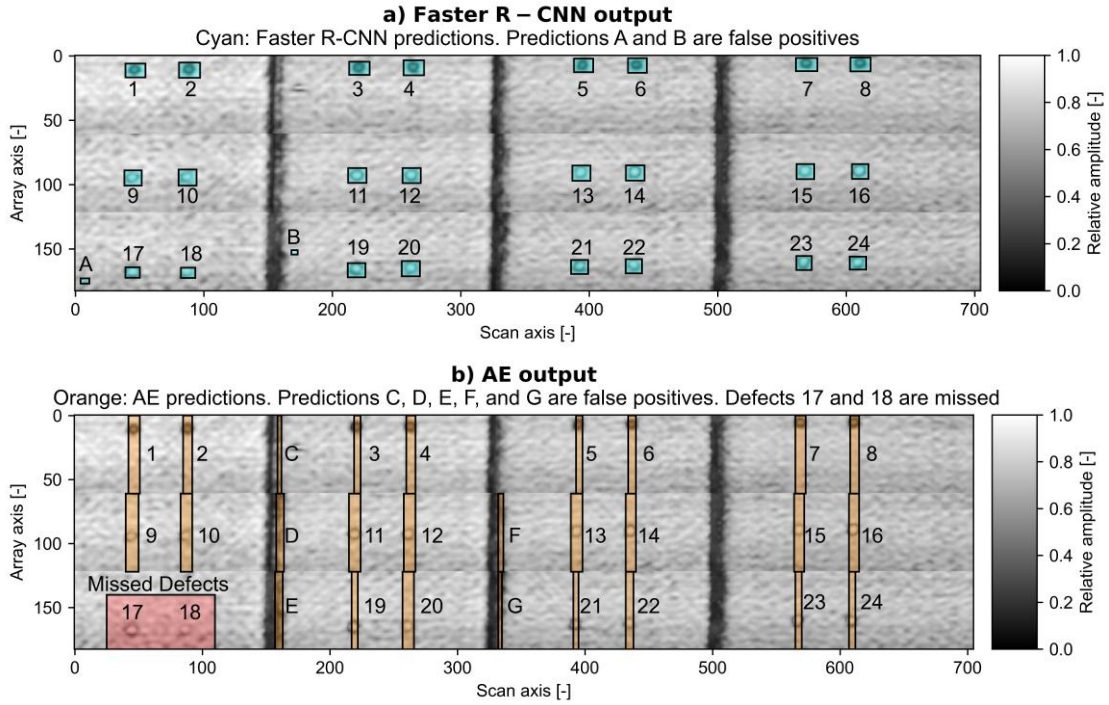
In the Level 1 Operator Assistance level of data analysis, inference parameters for both the Faster R-CNN and AE models are configured to minimise the risk of false negatives. While the ideal performance of an NDE operator or automated system would result in zero false negatives and false positives, achieving this is challenging. In the context of NDE for safety-critical components, the emphasis is heavily on minimising false negatives. Missing critical defects can have severe consequences, while false positives, though not directly threatening to material safety, may lead to higher costs if unnecessary rework is done, components are scrapped, or extended data analysis is conducted by operators to verify whether an indication is a true positive.

For Faster-RCNN, the confidence threshold determines the number of defects identified: a higher threshold results in fewer, but more confident detections, reducing false positives but potentially missing smaller or more subtle defects. On the other hand, a lower confidence threshold increases the number of detections for smaller defects and fainter indications, though this often leads to more false positives. At this automation level, where the final decision rests with the operator and all data is expected to be reviewed, the preference is typically for a lower confidence threshold (i.e., 0.001). This setting helps minimise false negatives while relying on operators to review and filter out false positives, ensuring that potential defects are flagged for further inspection and prioritising safety by reducing the risk of overlooked critical defects.

For the AE model, inference involves setting a threshold for anomaly detection based on observed MSE. A higher threshold flags only severe discrepancies from the MSE associated with pristine B-scans (i.e., significant defects), thus reducing false positives but potentially missing minor defects. A lower threshold, on the other hand, captures a larger number of indications, including minor deviations that may represent pristine B-scans, increasing the risk of false positives. Following a similar approach to the Faster R-CNN model, the AE model at this level of automation is configured to prioritise safety by applying a lower anomaly detection threshold (i.e., median MSE plus 50%).

The detection results for Sample C from both the Faster R-CNN and AE models are illustrated in Figure 63 a) and b). Faster R-CNN was able to capture all defects while producing two false positives. However, the AE model missed two defects, located at the thickest section of the sample near the back wall. A B-scan from that position is shown in Figure 63 c), where it can be observed why defects at such locations complicate detection. The indication is nearly fused with the back wall echo; Therefore, even with an optimal gating approach, a part of the defective signal would also be removed. This presents a challenge, as the defect appears very small on the B-scan level, and the C-scan amplitude response is also considerably weaker compared to other defects. For reference, the raw input data and ground truth annotations for sample C used in this analysis are described in Section 3.3.3.

As a comparison, Figure 63 d) shows a defect of equal size located immediately after the surface echo. While the main defect echo is again partially merged with the surface echo, as seen in the previous scenario, the several recorded ultrasonic repeats of the interface with a defect make detection easier. Therefore, even imperfect gating would leave strong reflections in the data, resulting in easier detections from both AE and Faster R-CNN. Lastly, Figure 63 e) presents a B-scan showing a false positive indication produced by the AE model. In this instance, a change in sample thickness results in many higher amplitude reflections caused by the interaction between the ultrasonic beam and sharp transition in sample geometry. Therefore, while the area captured in the B-scan frame is pristine, these minor indications cause a substantial deviation from the median MSE observed in the rest of the scan, resulting in a false positive flag.



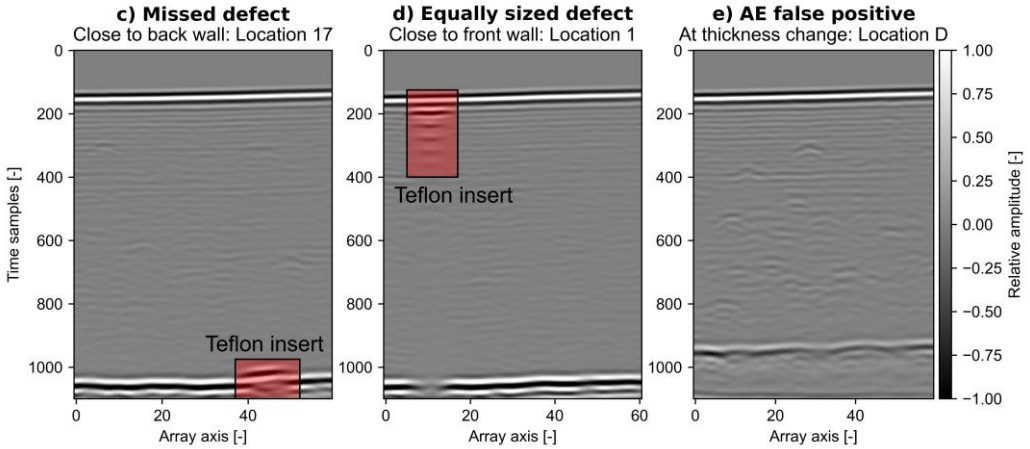


Figure 63 a) Output of the Faster R-CNN model, and b) Output of the AE model on C-scan view of the sample C (cyan/orange bounding boxes); c) B-scan frame containing a missed defect indication close to back wall; d) Equally sized defect close to front wall with ultrasound reverberations aiding the defect detection; e) AE false positive resulting from minor indications received from thickness transition at the location of sample geometrical steps.

In the presented examples, models are prone to generating false positive or false negative indications when calibrated with lower confidence and anomaly thresholds. Unfortunately, this approach yields results unsuitable for higher automation levels, especially due to the risks associated with missing defects, which could compromise the structural integrity of the final product if left unchecked. While false positives degrade the inference performance, they do not pose direct safety risks and can be addressed by NDE operators, albeit at the cost of additional analysis time. Nevertheless, the primary aim of this automation level is to assist with the analysis by providing informed suggestions on areas of interest, with the final decision remaining with the NDE operator who reviews all data. While adjusting model inference parameters could potentially lead to the successful detection of all defects by both models, changing these values on per sample basis is not feasible in the industrial system deployment.

The reasoning behind choosing the AE and Faster R-CNN models as the primary models for this application is their fast inference times, making them suitable for deployment on less powerful hardware. Only the inference times of the models are reported in this work. Faster R-CNN processing for Sample C takes 0.22 ± 0.06 seconds, while the AE model produces results in 2.28 ± 0.12 seconds. Specifically, the inference time for the AE is 1.56 ± 0.01 seconds, with an additional 0.73 ± 0.025 seconds required for padding inputs to match the AE's convolutional structure. On the

other hand, running the 3-DUSSS model on sample C takes 221.34 ± 1.41 seconds. The results of the 3-DUSSS model are overlayed over a C-scan of the sample and presented in Figure 64.

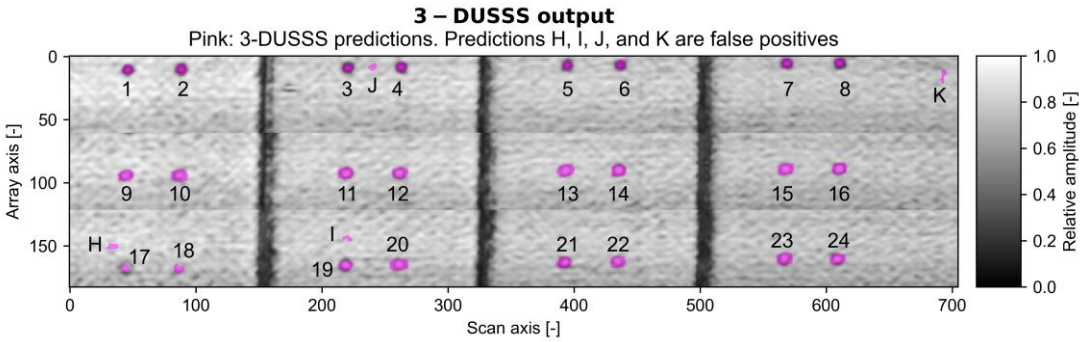


Figure 64 3-DUSSS segmentation output (pink) superimposed on the C-scan image of Sample C.

The detection results for Sample E are illustrated in Figure 65 a) and b). While Faster R-CNN successfully identified all defects with two false positives, the AE model failed to detect a 5.0 × 5.0 mm defect in the stringer section. Upon further inspection, this defect is partially visible in the scans but was not captured in its entirety due to an insufficient overlap between adjacent ultrasonic passes. As a result of this scanning error, the defect appears smaller than its true size, reducing its amplitude response, and preventing it from meeting the detection threshold for the AE model. Although reducing the anomaly threshold further might enable detection of this defect, it would also result in an excessive number of false positives across the scan. This defect is shown in Figure 65 c).

Additionally, while AE successfully identifies defects near the front wall, not all B-scans containing defects are flagged. Since defects typically span several B-scan slices, the MSE error varies across these slices, leading to some B-scans being correctly classified as anomalous while others are not. This approach still serves its purpose, as it provides the operator with a highlighted area of interest, which is valuable for guiding further inspection (although achieving complete detection would be ideal). An example of a partially captured defect is in Figure 65 d). This example highlights the advantage of Faster R-CNN, which leverages the spatial context across the C-scan view, rather than relying solely on individual B-scan slices. For reference, the raw input data and ground truth annotations for Sample E used in this analysis are described in Section 3.3.5.

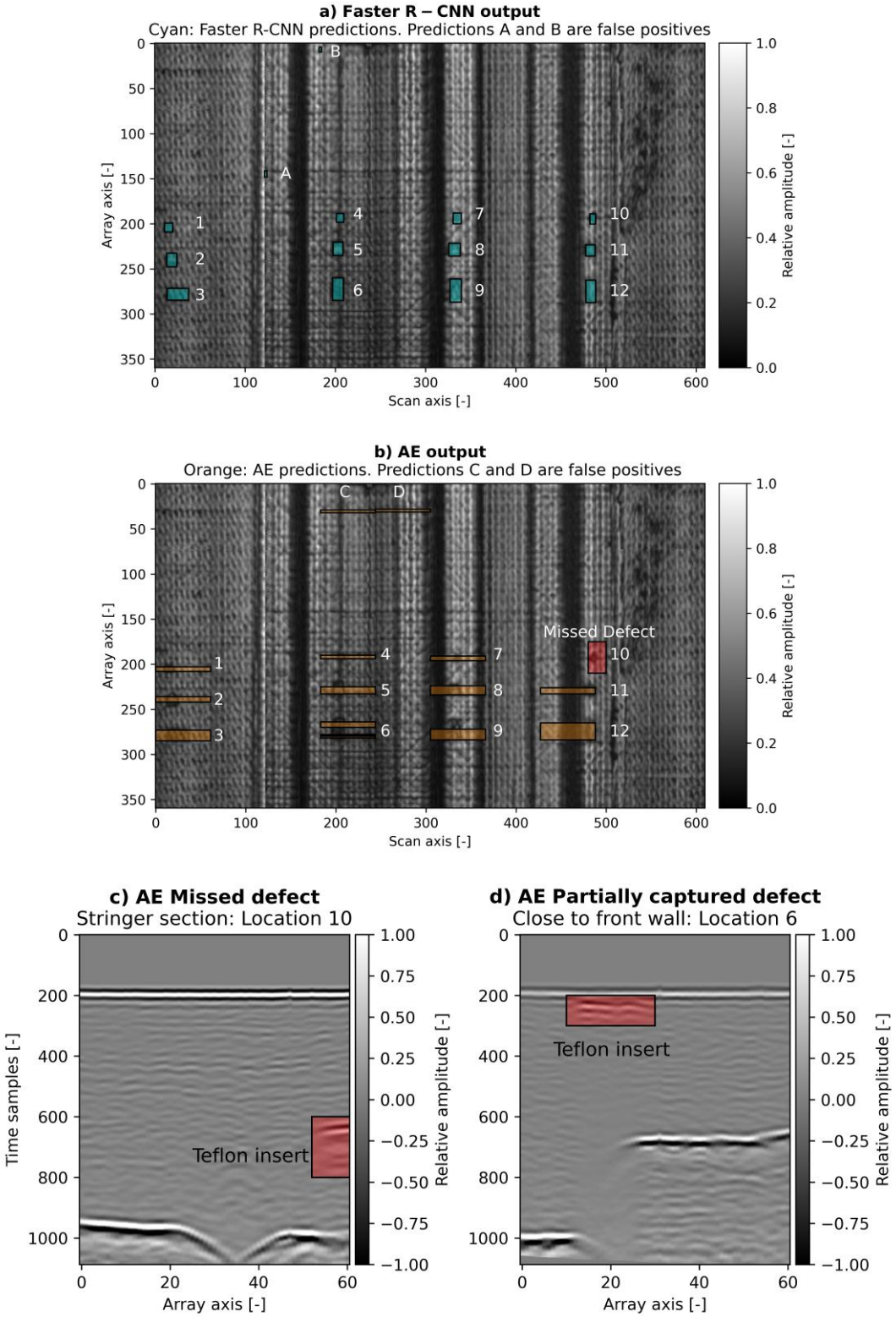


Figure 65 a) Output of the Faster R-CNN model and b) Output of the AE model overlaid on C-scan view of the sample E showing detected/missed defects (cyan and orange/red); c) Missed defect in stringer section; d) Partially captured defect close to the front wall.

Inference for FasterRCNN took 0.55 ± 0.08 seconds, while AE produced results in 2.34 ± 0.11 seconds, with additional time for padding resulting in 0.67 ± 0.01 seconds. 3-DUSSS model for this larger sample runs in 379.98 ± 1.21 seconds, which underscores the challenges in the scaling of inference time. In contrast, manual NDE inspection is typically reported to take significantly longer. For example, for a sample approximately double the size, the data interrogation is typically completed in 40 minutes by an operator, although this time is extended by an hour or more when defects are present as a closer examination and sizing of defective areas is required. While direct measurements for human analysis of the specific samples discussed in this work are not available, these figures highlight the time-saving potential of the proposed AI-based methods, which operate on the scale of seconds and minutes compared to tens of minutes or hours for manual inspection. The results of the 3-DUSSS model are overlaid on a C-scan and presented in Figure 66, showing that all defects were successfully detected, with five false positives.

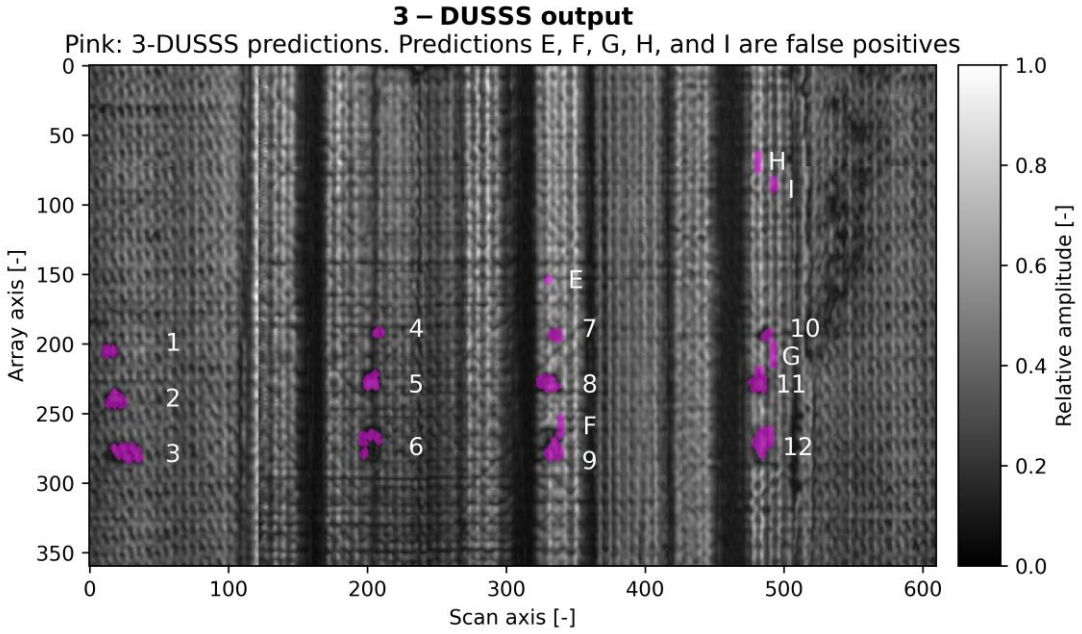


Figure 66 3-DUSSS segmentation output (pink) overlaid on the C-scan view of the sample E.

6.8 Level 2 – Partial automation

Level 2 of automation combines and compares the outputs of models, adding a layer of validation to AI predictions. In sample C, Faster R-CNN and AE agreed on 22 out of 24 defects, as shown in Figure 67. This agreement enhances trust in the system, as these areas are flagged by two independently trained AI models, each trained on

distinct data and ultrasonic views. Meanwhile, the nine areas of disagreement were flagged for human review, streamlining the analysis process. Rather than examining the entire dataset, the operator can now focus on these specific areas of disagreement, efficiently identifying the remaining two defects while filtering out false positives. For sample B, the models reached agreement on 11 out of 12 defects, with the human in the loop triggered to review five areas where the models disagreed, as shown in Figure 68.

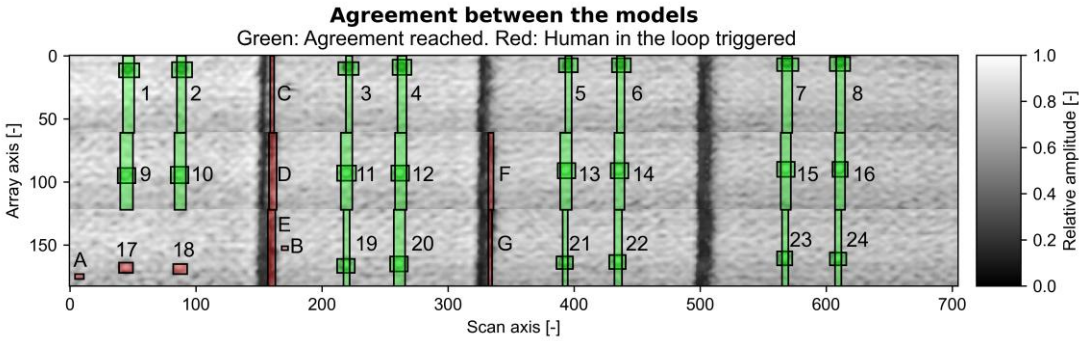


Figure 67 Sample C) Agreement (green) and disagreement (red) between the Faster R-CNN and AE models.

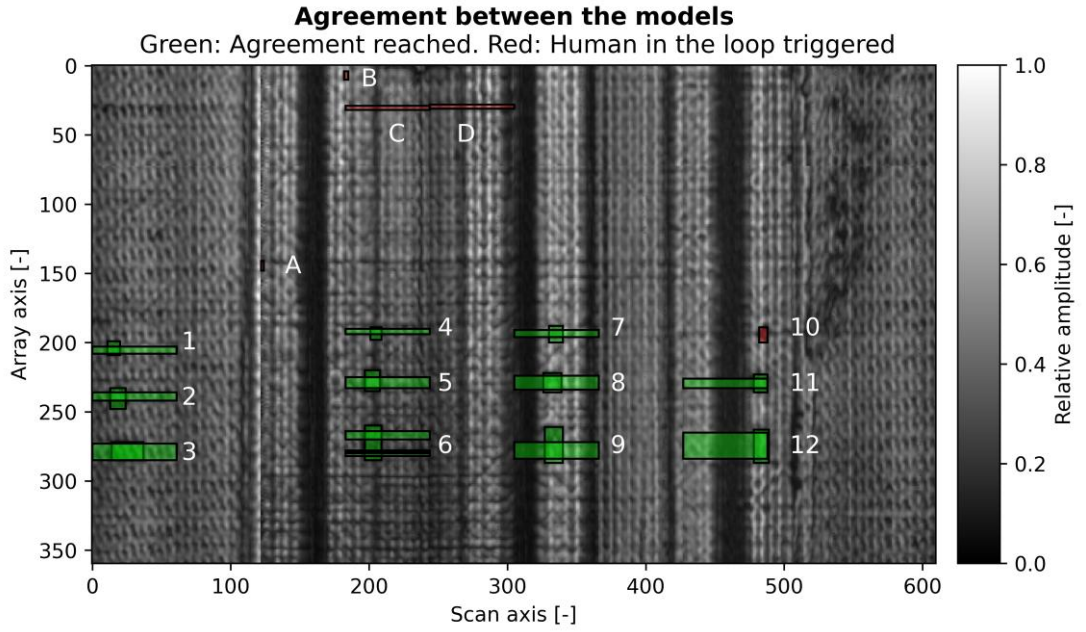
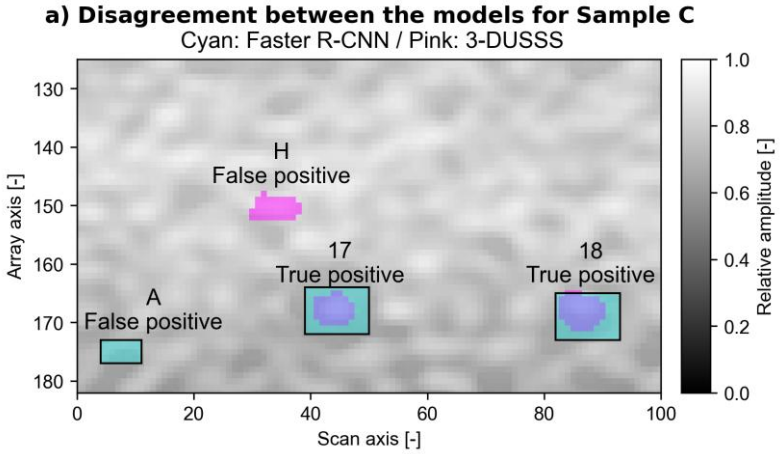


Figure 68 Sample E: Agreement (green) and disagreement (red) between the Faster R-CNN and AE models.

6.8.1 Level 3 – High automation

The multi-model Level 3 automation produced results consistent with Level 2 in terms of agreement between the AE and Faster R-CNN models, with the key difference being that disagreements between the models were resolved by the 3-DUSSS model rather than through a human-in-the-loop mechanism. In Sample C, the 3-DUSSS model confirmed that the two false negative calls by AE were defects, resulting in the successful identification of all 24 defects while discarding other false positive calls. An example of a disagreement in Sample C is shown in Figure 69 a), where two defects near the back wall were detected by the Faster R-CNN but missed by the AE model (refer to Figure 67). In this section, both the Faster R-CNN and 3-DUSSS models identified one false positive each. These detections were rechecked for agreement and rejected as false positives. In contrast, the two detections with coinciding results were confirmed as true positives, resolving the disagreement between the models. Inference time for the 3-DUSSS model was significantly reduced compared to processing the full volume, taking 91.59 ± 0.83 seconds to resolve nine areas of disagreement. While this reduction may seem modest here, it is important to highlight that these experiments were conducted on relatively small scans and reference samples. For larger datasets, typical in industrial applications, this targeted approach would likely result in more substantial time savings.



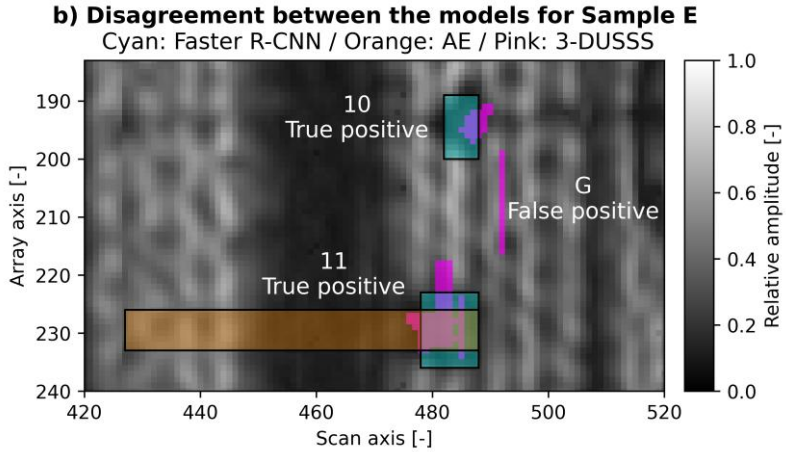


Figure 69 Areas of disagreement between models resolved by 3-DUSSS ; a) Sample C with Faster R-CNN (cyan) and 3-DUSSS (pink) predictions overlaid on the C-scan; b) Sample E with Faster R-CNN (cyan), AE (orange), and 3-DUSSS (pink) predictions overlaid on the C-scan.

Similar results were observed in Sample E, where all defects were correctly identified. An example of model disagreement is shown in Figure 69 b), where a 5.0 × 5.0 mm stringer defect, missed by AE, was confirmed as a true positive by the 3-DUSSS model. As in the previous example, 3-DUSSS produced one false positive, which was rejected since it did not coincide with any other model's detection. The inference time for the 3-DUSSS model to resolve five areas of disagreement was 54.12 ± 0.74 seconds.

Overall, Level 3 automation offers several benefits. By combining the Faster R-CNN, AE, and 3-DUSSS models, all defects in this study were successfully detected, with the 3-DUSSS model resolving areas of disagreement and filtering out false positives. This approach ensures fast, reliable results while enabling the use of less powerful hardware. Additionally, the workflow reduces both analysis time and operator workload, while still allowing operators to review intermediate results, examine areas of disagreement, and intervene if needed, thereby preventing potential performance degradation. This workflow achieves results close to the ideal fully automated process, with minimal impact on analysis time and system failure risk. An overview of the performance metrics for individual models, including the number of false positive and false negative calls, as well as inference times, is provided in Table 14. Recall is defined as the number of true positives divided by the sum of true positives and false

negatives, while precision is the number of true positives divided by the sum of true and false positives. The F1 score is calculated as the harmonic mean of precision and recall

Table 14 Overview of reported performance metrics for different automation levels

Automation Level 1	**Metric**	**System**					
		Anomaly detection AE		**Faster R-CNN**		**3-DUSS**	
		Sample A	Sample B	Sample A	Sample B	Sample A	Sample B
Single model system	Inference [s]	2.28 ± 0.12	2.34 ± 0.11	0.22 ± 0.06	0.55 ± 0.08	221.34 ± 1.41	379.98 ± 1.21
Operator reviews all data and all AI model outputs	False positives [-]	5	2	2	2	4	5
	False negatives [-]	2	1	0	0	0	0
	Precision [-]	0.815	0.846	0.923	0.857	0.857	0.706
	Recall [-]	0.917	0.917	1.000	1.000	1.000	1.000
	F1 [-]	0.863	0.880	0.960	0.923	0.923	0.828

Automation Level 2	**Metric**	**System**	
		Anomaly detection AE | **Faster R-CNN**	
		Sample A	Sample B
Two model system	Inference [s]	2.44 ± 0.18 (2.28 ± 0.12 | 0.22 ± 0.06)	2.89 ± 0.19 (2.34 ± 0.11 | 0.55 ± 0.08)
Human-in-the-loop mechanism triggered for disagreements	False positives [-]	7	4
	False negatives [-]	0	0
	Flagged for Human-in-the-loop mechanism [-]	9 (7 false positives and 2 true positives)	5 (4 false positives and 1 true positive)

Automation Level 3	**Metric**	**System**	
		Anomaly detection AE | **Faster R-CNN** | **3-DUSS**	
		Sample A	Sample B
Three model system	Inference [s]	94.03 ± 1.01 (2.28 ± 0.12 | 0.22 ± 0.06 | 91.59 ± 0.83)	57.01 ± 0.93 (2.34 ± 0.11 | 0.55 ± 0.08 | 54.12 ± 0.74)
Operator moved to supervisory role	False positives [-]	0	0
	False negatives [-]	0	0

While for this specific scenario Faster R-CNN performs best, this does not guarantee that it will always outperform other models across different datasets, defect types, or acquisition conditions. Therefore, the system adopts a layered decision-making

strategy, combining outputs from multiple models to improve robustness. This reflects ensemble learning principles, where agreement among diverse models strengthens confidence in a result. Additionally, since the models operate on fundamentally different data representations, their decisions are independent. The 3-DUSSS model, while slower, is used only in higher-level decision stages to verify results from faster models, improving results without unnecessary processing overhead.

6.8.2 Conclusions, Limitations, and Future Work

In this chapter, AI-aided data analysis strategies were explored across proposed levels of model aggregation, focusing on the use of multiple AI models to simultaneously process different ultrasonic views. A case study was conducted on two defective CFRP reference samples containing 36 manufactured defects. These samples were inspected using an industrial manipulator and a PAUT roller probe to simulate industrial practices for inspecting large composite components. Integrating multiple models within the NDE data analysis workflow provided flexibility in designing workflows, managing intermediate results, and resolving model disagreements. This approach also facilitated a more robust setup leading to the successful detection of all examined defects. The study revealed that for:

- Level 1 - Operator Assistance: The conservative use of AI models prioritises safety by minimising false negatives, at the cost of increasing false positives. The suggestions provided by the models accelerate data analysis while maintaining minimal risks associated with reliance on automation. Human operators validate all AI outputs and retain full control over decision-making, resulting in only a slight increase in system complexity.
- Level 2 - Partial Automation: Improved results were achieved by comparing outputs from two models and prompting human operators to intervene in areas of disagreement. This comparison acts as an additional validation step for reported detections, aiming to increase trust in the automated process. While this approach speeds up data analysis, it requires a higher degree of trust in the models.
- Level 3 - High Automation: Incorporating the 3-DUSSS model as an arbiter enabled a simultaneous analysis workflow that processes ultrasonic B-scans,

C-scans, and full volumetric data. The deployment of 3-DUSSS to only areas of disagreement greatly reduced inference times and memory requirements, making this strategy deployable on less powerful hardware. The combination of three models achieved near-ideal results while addressing model trust concerns with two layers of validation.

While this research provides an analysis of the performance of different automation levels on fabricated defects of known size and shape, there is an opportunity to explore the system's functionality when applied to naturally occurring defects, such as porosities, to assess the robustness of individual models across a wider range of defect types. Additionally, optimisation of models in terms of hyperparameter tuning, changes in architectures, or training regimes with new and varied data is deemed promising for achieving improved results.

In future work, the developed system will be integrated into a production-level industrial use case to assess its scalability, robustness, and performance in a complex real-world environment, while also addressing integration challenges with existing workflows. Additionally, the expansion of defect detection models to include a broader range of defect types, such as porosities or foreign object inclusions, will be explored.

As with any AI-based system, there remains a potential risk of false positives or false negatives. In safety-critical applications, such risks are typically mitigated through layered inspection strategies. In this study, no false negatives were observed across the tested samples when models were used in conjunction. However, further validation and benchmarking against larger representative datasets would be required for broader deployment.

Building trust in such AI-driven systems requires more than technical performance. It demands transparency, reproducibility, and ongoing validation in real operational settings. The layered, multi-model structure introduced in this work inherently supports trust by enabling cross-validation between complementary models and providing traceable outputs. Trust can be progressively established in production environments through a phased rollout where AI initially supports human inspectors, gradually increasing automation as confidence grows. Maintaining full traceability of decisions, including which model contributed to specific outputs, alongside inclusion

of explainable features (AE residuals, Faster R-CNN bounding boxes, and voxel-level maps from 3-DUSS) in quality reports can improve transparency.

Redundancy through human oversight and override capabilities also serve as a critical safeguard and helps mitigate risks of automation over-reliance highlighted in section 6.3, maintaining operator skills through periodic checks and false flagging. Continuous monitoring of individual model outputs also enables early detection of performance drift or deterioration, which can help with targeted retraining or system calibration. Together, these strategies could form a trust-building framework that complements the technical robustness demonstrated in this chapter.

Chapter 7: Summary and Future Work

7.1 Thesis Purpose and Scope

The purpose of this thesis was to explore how AI technologies can considerably accelerate/enhance the data analysis workflows for NDE data collected via robotically delivered PAUT sensors. The focus was on inspecting CFRP materials employed in the aerospace industry for construction of high-value safety-critical components. Within the scope of this research were:

- Conducting a focused literature review on AI and ML approaches for UT data analysis, particularly for NDE of CFRP materials, to identify research trends and common challenges.
- Acquiring representative datasets of UT scans from various CFRP samples using an automated robotic setup similar to ones used in industry.
- Developing AI-driven workflows to streamline data interpretation with emphasis on process automation.

7.2 Summary of Key Findings

Chapter 1 established the industrial motivation and research context, emphasizing the increasing adoption of CFRPs in the aerospace sector and the vital role of NDE in ensuring the structural integrity of final products. UT was introduced as the primary method for bulk inspection of CFRPs, highlighting its dominant role in the NDE. The integration of robotic systems for deploying UT sensors was discussed, highlighting the resulting improvements in data acquisition throughput and scan consistency. Despite these advancements, data interpretation remains a manual, slow, and labour-intensive process prone to errors and misinterpretation and adversely affecting the production rate, especially when dealing with large datasets where operator fatigue can become a critical factor. AI was identified as a promising solution to these challenges, with capacity to tackle complex tasks and large volumes of data with near-human performance.

Chapter 2 provided the foundational background knowledge necessary for the research conducted in this thesis. It began by explaining the fundamentals of ultrasound, including conventional UT and PAUT, as well as signal processing techniques. The

chapter then highlighted the specific applications of UT in inspecting CFRPs used in the aerospace industry, with a focus on the types of defects that can arise during manufacturing. Following this, basic AI concepts were introduced using examples of linear DL NNs and CNNs, providing a foundation for understanding the AI models used later in the thesis. The chapter also reviewed past academic publications leveraging different formats of UT data in NDE for training and testing AI models, identifying key challenges and summarising the state-of-the-art of AI implementation for UT signal analysis within the field.

Chapter 3 detailed the experimental setup, materials, and equipment used throughout this thesis. It began by introducing the ultrasonic setup, comprising a PAUT roller probe and an ultrasonic controller. The robotic setup was then described, including the industrial manipulator, FT sensor, and the LabVIEW VI environment and control system, which enabled precise and automated data collection. Following this, the chapter provided an overview of the CFRP samples examined in this work. Finally, the hardware specifications of the PCs utilised for AI model training and simulations were outlined.

Chapter 4 explored and examined various approaches for defect detection and localisation within ultrasonic amplitude C-scans. These included traditional signal thresholding based on observed amplitudes, a statistical thresholding method using theoretical mathematical distributions fitted to the data, and the application of several AI object detection models. The supervised training of AI models leveraged transfer learning, using pretrained model weights trained on unrelated datasets to accelerate convergence. In the absence of large experimental datasets, representative training data were generated using the semi-analytical simulation software CIVA. Additionally, a domain adaptation technique was employed by analysing noise profiles from real scans and modelling them into the simulated data to reduce the gap between synthetic and experimental domains. The results demonstrated that object detection models outperformed thresholding methods, highlighting the potential of AI for accurate and efficient defect detection in ultrasonic NDE workflows.

Chapter 5 builds on the findings of the previous chapter by addressing key areas for improvement, particularly in signal gating and leveraging alternative UT data

projections to apply unsupervised learning, reframing defect detection as an anomaly detection problem. To this end, an unsupervised clustering approach using DBSCAN, combined with a peak-finding algorithm, was employed for automated gating, while an autoencoder architecture was used as an anomaly detector. This method enabled precise detection of defects larger than 4.0 mm and efficient removal of front and back walls from the data, regardless of the sample's geometry. Additionally, the impact of human labelling variability on reported performance metrics was examined, emphasising the uncertainties inherent in manual ground truth creation. This chapter demonstrated the viability of unsupervised methods for defect detection, particularly in scenarios where large labelled datasets are unavailable, and displayed the ability of clustering techniques to effectively isolate geometrical features.

Chapter 6 combines the developed supervised and unsupervised approaches presented in this thesis with an additional self-supervised model into a comprehensive AI-driven data analysis workflow for processing ultrasonic data. This chapter addresses a gap in the academic literature, as AI models in NDE are framed primarily as replacements for human operators rather than collaborative tools. Cantero-Chinchilla et al. [22] outline the broader automation process in NDE and its prerequisites, however, detailed strategies specific to AI-driven ultrasonic data analysis remain unexplored. Similar discussions and implementation strategies are happening in other fields such as medicine, where AI tools are increasingly used for drug discovery [301] and are reshaping the technical requirements and training of professionals working with these tools. Likewise, AI is being integrated into medical imaging to assist with various tasks [302]. Other studies have examined the challenges and requirements for implementing AI-based tools in healthcare, as well as their future impact [303]. The proposed approach defines several levels of automation, ranging from basic levels where AI models provide suggestions to the NDE operator, to more advanced workflows in which multiple AI models collaborate to process the data. The interaction between the models, leveraging different ultrasonic views, simulates the manual data examination process performed by an NDE operator. At the same time, it incorporates mechanisms to increase trust in the automated system while minimising false positives and false negatives. Finally, the proposed integration strategy is tested and evaluated on a

complex geometry CFRP sample, highlighting the effectiveness of AI in industrial applications.

7.3 Limitations and Future Work

The primary limitation of applying AI to NDE tasks is the lack of readily available representative datasets for developing and testing. While the field of AI has seen a surge in new research, this progress is enabled by standardised datasets, such as ImageNet and COCO, which are widely used for computer vision tasks. These datasets enable research groups to directly compare their efforts, reducing the barrier to entry posed by dataset creation and labelling. Unfortunately, similar datasets are not available in the field of NDE, particularly for the examination of CFRP materials. Few notable exceptions were mentioned in Chapter 2, but at the time of writing, these datasets have not become standards for testing of new AI research in NDE.

Furthermore, one of the key reasons for the scarcity of publicly available CFRP datasets is their proprietary nature, as they are typically originating from high-value components associated with defence, transportation, aerospace, and energy sectors. Sharing such data often requires following stringent data protection protocols. Even when data sharing is possible, variations in data formats pose additional challenges. These formats are often proprietary and can differ significantly depending on the inspection method used. A notable effort to standardise UT data formats was made by the University of Bristol, which proposed the mfmc format, however, its adoption is still in progress [304].

Because of this, alternative approaches in terms of synthetic data generation or focus on unsupervised methods must be taken to address this issue. While these methods show promising results, it is expected for supervised methods to outperform unsupervised methods. Therefore, future work towards creation of a publicly available UT datasets upon which different AI methods could be tested, and an attempt of exploring the development of a uniform data format would have high impact in this field of research.

Another avenue not explored within this thesis is the classification of defects. Beyond defect localisation, detection, and sizing, an essential aspect of the NDE workflow is the categorisation of defects (distinguishing between different defect types such as

porosities, delaminations, or fibre waviness). This is inherently a supervised learning task, as classification relies on labelled training data. While unsupervised methods can group similar data patterns through clustering or dimensionality reduction (e.g., applying uniform manifold approximation and projection [278] or principal component analysis on the latent space of an autoencoder covered in Chapter 5), these groupings do not correspond to specific defect classes unless they are manually interpreted or later annotated. Therefore, unsupervised methods are insufficient for defect classification on their own, however they may be used as a pre-processing step or in combination with weak supervision approaches [305]. Furthermore, the currently available simulation software, CIVA, does not yet support the simulation of the defects such as porosities, cracks, dry spots, ply wrinkles, or fibre waviness, which would present a more challenging evaluation task. While one might be able to use FEA software to generate more realistic datasets with other types of defects, the associated computational cost would be prohibitively high, as discussed in Chapter 4.

Attempts to artificially induce such defects were made but are not documented within this thesis. While industrial inspection scenarios were successfully replicated, achieving the high level of precision required for manufacturing realistic CFRP defects proved far more difficult. For instance, while drilling FBHs involved some degree of variability, these processes were still more controlled than attempts to create porosities by embedding small glass spheres or purposefully degassing polymer binders. These methods also incur significant material costs, which were financially unfeasible within the scope of this thesis.

The most effective approach would be to incorporate a wide range of naturally occurring defects to rigorously test the performance of the developed models, with the expectation of encountering reduced performance. In the future, if such datasets become available, the work presented in Chapters 4 through 6 could be revisited to evaluate whether the proposed methods remain effective when tested on more diverse and representative datasets.

The samples examined in this thesis were controlled, and scans were performed to minimise variability within the data. However, this level of control may not translate directly to industrial applications. While Chapter 6 demonstrated the applicability of

the developed methods to an industrial wing cover component, other industrial scenarios and components may present increased variability in scan quality and utilise different imaging modalities or equipment. It is worth noting, however, that some industrial setups (such as the one used in Spirit AeroSystems' Belfast factory, which uses a water irrigation system for coupling) can produce higher-quality data than the roller probe used in this study. This suggests that the proposed methods may, in fact, be transferable to industrial environments with minimal adaptation. Nevertheless, the performance of the proposed models must be thoroughly tested in real-world industrial settings, as such environments are likely to pose greater challenges for automating data analysis and defect detection.

Additionally, the regulatory aspect of employing these technologies were beyond the scope of this thesis but could present significant barriers to implementation. A potential pathway to addressing these challenges is the adoption of explainable AI which could enhance transparency in AI-driven systems. However, research in this area remains limited, with only a few academic publications exploring its applicability to NDE.

Lastly, the field of AI is advancing rapidly, making it challenging to stay up to date with every new development, technique, and state of the art model. For instance, in Chapter 4, the YOLOv5 model was employed, but by the time of publication, several newer versions from different research groups had already been released. With each iteration, different architectural components were updated; for example, changes in the feature aggregation strategy in the neck of YOLOv7 [306], or improvements in inference speed in YOLOv10 through the removal of NMS [307]. YOLOv9 focused on lightweight architecture adjustments to better support deployment on edge computing [308]. While these updates could positively impact on the tasks presented in this thesis, most of these improvements are aimed at inference efficiency and deployment flexibility, rather than detection accuracy.

A more substantial architectural shift occurred in YOLOv8 with the introduction of anchor-free bounding box detection. Unlike anchor-based methods, which rely on predefined box sizes and aspect ratios during training, anchor-free approaches predict object locations directly (centre of an object) and then regress the bounding box

dimensions [309]. This change can improve generalisation, particularly in datasets where object sizes vary significantly or are not well represented by predefined anchors used during training. However, in this study, no degradation in performance due to size mismatch was observed when testing on defects larger than those observed during training. Therefore, while anchor-free detection is a promising direction, it is not guaranteed to provide significant gains for the datasets and application context used in this thesis. Future work could explore these newer models within the same framework, particularly where real-time inference or deployment on constrained hardware becomes a primary consideration.

Similarly, hyperparameter optimisation, although computationally intensive, could potentially enhance results further. Additional strategies such as test-time augmentation and inference with an ensemble of models also hold promise for improved performance. Additionally, recent advancements in large language models have made them increasingly multimodal, enabling their application to computer vision tasks. Techniques such as zero-shot or few-shot prompting offer new opportunities for leveraging these models; however, their significant size and prohibitive computational costs remain substantial barriers for research teams outside of large technology companies.

Bibliography

[1] A. Quilter, ‘Composites in Aerospace Applications’, 2001.

[2] R. Slayton and G. Spinardi, ‘Radical innovation in scaling up: Boeing’s Dreamliner and the challenge of socio-technical transitions’, *Technovation*, vol. 47, pp. 47–58, Jan. 2016, doi: 10.1016/j.technovation.2015.08.004.

[3] P. D. Mangalgiri, ‘Composite materials for aerospace applications’, *Bull Mater Sci*, vol. 22, no. 3, pp. 657–664, 1999.

[4] V. Giurgiutiu, *Structural health monitoring of aerospace composites*. in Structural Health Monitoring of Aerospace Composites. Elsevier, 2016, p. 23. doi: 10.1016/B978-0-12-409605-9.00001-5.

[5] J. Bachmann, C. Hidalgo, and S. Bricout, ‘Environmental analysis of innovative sustainable composites with potential use in aviation sector—A life cycle assessment review’, *Sci. China Technol. Sci.*, vol. 60, no. 9, pp. 1301–1317, Sep. 2017, doi: 10.1007/S11431-016-9094-Y.

[6] O. Younossi, M. Kennedy, and J. C. Gräser, *Military Airframe Costs The Effects of Advanced Materials and Manufacturing Processes*. 2001, p. 9. [Online]. Available: http://www.rand.org/

[7] C. Red, ‘777X: Bigger-than-expected carbon fiber impact | CompositesWorld’, 2016, [Online]. Available: https://www.compositesworld.com/articles/777x-bigger-than-expected-carbon-fiber-impact

[8] A. Wilson, *Advances in Technical Nonwovens*. in Advances in Technical Nonwovens. Elsevier Inc., 2016, p. 271. doi: 10.1016/B978-0-08-100575-0.00009-7.

[9] S. Coad, ‘The F-15 strike eagle’, *Adv. Mater. Process.*, vol. 161, no. 4, pp. 45–47, Apr. 2003.

[10] J. P. Halpin *et al.*, ‘T F/A-18 E/F Program Independent Analysis’, *JOHNS HOPKINS APL Tech. Dig.*, vol. 18, no. 1, 1997.

[11] P. J. Charitidis, ‘CRITERIA FOR THE SELECTION OF CARBON FIBER COMPOSITE MATERIALS FOR FIGHTER AIRCRAFT’, *Adv. Mater. Sci. Eng. Int. J. MSEJ*, vol. 5, no. 2, 2018, doi: 10.5121/msej.2018.5401.

[12] CMS Admin, ‘F-22A Raptor Advanced Tactical Fighter, United States of America’, 2020, [Online]. Available: https://www.airforce-technology.com/projects/f22a-raptor/

[13] J. Sloan, ‘Skinning the F-35 fighter | CompositesWorld’, 2009, [Online]. Available: https://www.compositesworld.com/articles/skinning-the-f-35-fighter

[14] S. Hegde, B. Satish Shenoy, and K. N. Chethan, ‘Review on carbon fiber reinforced polymer (CFRP) and their mechanical performance’, *Mater. Today Proc.*, vol. 19, pp. 658–662, Jan. 2019, doi: 10.1016/j.matpr.2019.07.749.

[15] K. I. Tserpes, P. Papanikos, G. Labeas, and S. Pantelakis, ‘Fatigue damage accumulation and residual strength assessment of CFRP laminates’, *Compos. Struct.*, vol. 63, no. 2, pp. 219–230, Feb. 2004, doi: 10.1016/S0263-8223(03)00169-7.

[16] J. Ostrower, ‘Delamination prompts Boeing to inspect 787 fleet’, Flight Global. Accessed: Jul. 15, 2025. [Online]. Available: https://www.flightglobal.com/delamination-prompts-boeing-to-inspect-787-fleet/103901.article

[17] B. Murphy, J. O'Callaghan, M. Fox, L. Ilcewicz, and J. Starnes, 'Overview of the Structures Investigation for the American Airlines Flight 587 Investigation', in *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, American Institute of Aeronautics and Astronautics. doi: 10.2514/6.2005-2251.
[18] 'Non-Destructive Testing (NDT) Market Size, Growth Analysis 2029', [Online]. Available: https://www.fortunebusinessinsights.com/non-destructive-testing-ndt-market-103596
[19] C. Mineo *et al.*, 'Flexible integration of robotics, ultrasonics and metrology for the inspection of aerospace components', *AIP Conf. Proc.*, vol. 1806, no. 1, pp. 020026–020026, Feb. 2017, doi: 10.1063/1.4974567.
[20] M. Bertovic and I. Virkkunen, 'NDE 4.0: New Paradigm for the NDE Inspection Personnel', *Handb. Nondestruct. Eval. 40*, pp. 1–31, 2021, doi: 10.1007/978-3-030-48200-8_9-1.
[21] I. Virkkunen, T. Koskinen, O. Jessen-Juhler, and J. Rinta-aho, 'Augmented Ultrasonic Data for Machine Learning', *J. Nondestruct. Eval.*, vol. 40, no. 1, pp. 1–11, Mar. 2021, doi: 10.1007/S10921-020-00739-5/TABLES/1.
[22] S. Cantero-Chinchilla, P. D. Wilcox, and A. J. Croxford, 'Deep learning in automated ultrasonic NDE -- developments, axioms and opportunities', *ArXiv211206650 Eess*, Dec. 2021, doi: 10.48550/arxiv.2112.06650.
[23] 'EU AI Act: first regulation on artificial intelligence', Topics | European Parliament. Accessed: Jul. 15, 2025. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
[24] 'AI Act | Shaping Europe's digital future'. Accessed: Jul. 15, 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
[25] Clea, 'ENIQ reports', SNETP. Accessed: Jul. 15, 2025. [Online]. Available: https://snetp.eu/eniq-reports/
[26] R. S. Fernandez Orozco, K. Hayes, K. Carpenter, and F. Gayosso, 'Artificial Intelligence and NDE Competencies', in *Handbook of Nondestructive Evaluation 4.0*, Springer, Cham, 2025, pp. 795–852. doi: 10.1007/978-3-031-84477-5_24.
[27] A. Krizhevsky, S. Ilya, and G. E. H. Hinton, 'Imagenet classification with deep convolutional neural networks', presented at the Advances in neural information processing systems 25, 2012.
[28] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.48550/arxiv.1512.03385.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 'ImageNet: A large-scale hierarchical image database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[30] D. Silver *et al.*, 'Mastering the game of Go with deep neural networks and tree search', *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.
[31] D. Silver *et al.*, 'Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm', Dec. 05, 2017, *arXiv*: arXiv:1712.01815. doi: 10.48550/arXiv.1712.01815.

[32] J. De Fauw *et al.*, ‘Clinically applicable deep learning for diagnosis and referral in retinal disease’, *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018, doi: 10.1038/s41591-018-0107-6.
[33] ‘Kaggle Competitions’. Accessed: Jul. 15, 2025. [Online]. Available: https://www.kaggle.com/competitions
[34] ‘Artificial Intelligence (AI) Market Size, Growth, Report By 2032’. Accessed: Oct. 30, 2023. [Online]. Available: https://www.precedenceresearch.com/artificial-intelligence-market
[35] T. G. Leighton, ‘What is ultrasound?’, *Prog. Biophys. Mol. Biol.*, vol. 93, no. 1, pp. 3–83, Jan. 2007, doi: 10.1016/j.pbiomolbio.2006.07.026.
[36] U. Schnars and R. Henrich, ‘Applications of NDT Methods on Composite Structures in Aerospace Industry’, presented at the Conference on Damage in Composite Materials, 2006.
[37] A. Kapadia, ‘National Composites Network Best Practice Guide Non Destructive Testing of Composite Materials’, 2007. [Online]. Available: http://www.twi.co.uk/j32k/index.xtp
[38] S. Kashif Ur Rehman, Z. Ibrahim, S. A. Memon, and M. Jameel, ‘Nondestructive test methods for concrete bridges: A review’, *Constr. Build. Mater.*, vol. 107, pp. 58–86, Mar. 2016, doi: 10.1016/J.CONBUILDMAT.2015.12.011.
[39] D. M. McCann and M. C. Forde, ‘Review of NDT methods in the assessment of concrete and masonry structures’, *NDT E Int.*, vol. 34, no. 2, pp. 71–84, Mar. 2001, doi: 10.1016/S0963-8695(00)00032-3.
[40] W. M. Alobaidi, E. A. Alkuam, H. M. Al-Rizzo, and E. Sandgren, ‘Applications of Ultrasonic Techniques in Oil and Gas Pipeline Industries: A Review’, *Am. J. Oper. Res.*, vol. 05, no. 04, pp. 274–287, 2015, doi: 10.4236/AJOR.2015.54021.
[41] S. A. Titov, R. G. Maev, and A. N. Bogachenkov, ‘Pulse-echo NDT of adhesively bonded joints in automotive assemblies’, *Ultrasonics*, vol. 48, no. 6–7, pp. 537–546, Nov. 2008, doi: 10.1016/J.ULTRAS.2008.07.001.
[42] B. Djordjevic, ‘Nondestructive test technology for the composites’, *10th Int. Conf. Slov. Soc. Non-Destr. Test.*, pp. 259–265, 2009.
[43] V. Dattoma, R. Nobile, F. W. Panella, A. Pirinu, and A. Saponaro, ‘Optimization and comparison of ultrasonic techniques for NDT control of composite material elements’, *Procedia Struct. Integr.*, vol. 12, pp. 9–18, 2018, doi: 10.1016/J.PROSTR.2018.11.111.
[44] 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, 2015, doi: 10.1016/J.PROCIR.2015.07.043.
[45] L. Liu, W. Liu, D. Teng, Y. Xiang, and F.-Z. Xuan, ‘A multiscale residual U-net architecture for super-resolution ultrasonic phased array imaging from full matrix capture data’, *J. Acoust. Soc. Am.*, vol. 154, no. 4, pp. 2044–2054, Oct. 2023, doi: 10.1121/10.0021171.
[46] J. D. Achenbach, *Wave propagation in elastic solids*. in North-Holland series in applied mathematics and mechanics, no. v. 16. Amsterdam New York: North-Holland Pub. Co. American Elsevier Pub. Co, 1973.
[47] B. A. Auld, *Acoustic fields and waves in solids*. Рипол Классик, 1973.
[48] A. Manbachi and R. S. C. Cobbold, ‘Development and Application of Piezoelectric Materials for Ultrasound Generation and Detection’, *Ultrasound*, vol. 19, no. 4, pp. 187–196, Nov. 2011, doi: 10.1258/ult.2011.011027.

[49] R. Smith, ‘Ultrasonic defect sizing in carbon-fibre composites -an initial study’, vol. 36, pp. 595–605, Aug. 1994.
[50] H. J. Gohari, ‘Focusing of ultrasound beams’, Master thesis, 1997. Accessed: Mar. 17, 2025. [Online]. Available: https://www.duo.uio.no/handle/10852/8781
[51] S. Schmid *et al.*, ‘Estimating Young’s moduli based on ultrasound and full-waveform inversion’, *Ultrasonics*, vol. 136, p. 107165, Jan. 2024, doi: 10.1016/j.ultras.2023.107165.
[52] L. W. Schmerr, *Fundamentals of Ultrasonic Nondestructive Evaluation: A Modeling Approach*. in Springer Series in Measurement Science and Technology. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-30463-2.
[53] A. A. Karabutov, N. B. Podymova, and I. O. Belyaev, ‘The influence of porosity on ultrasound attenuation in carbon fiber reinforced plastic composites using the laser-ultrasound spectroscopy’, *Acoust. Phys.*, vol. 59, no. 6, pp. 667–673, Nov. 2013, doi: 10.1134/S1063771013060080.
[54] R. A. Roberts, ‘Computational Prediction of Micro-crack Induced Ultrasound Attenuation in CFRP Composites’, *J. Nondestruct. Eval.*, vol. 33, no. 3, pp. 443–457, Sep. 2014, doi: 10.1007/s10921-014-0240-1.
[55] K. Ono, ‘A Comprehensive Report on Ultrasonic Attenuation of Engineering Materials, Including Metals, Ceramics, Polymers, Fiber-Reinforced Composites, Wood, and Rocks’, *Appl. Sci.*, vol. 10, no. 7, Art. no. 7, Jan. 2020, doi: 10.3390/app10072230.
[56] R. H. Bossi and V. Giurgiutiu, ‘15 - Nondestructive testing of damage in aerospace composites’, in *Polymer Composites in the Aerospace Industry*, P. E. Irving and C. Soutis, Eds., Woodhead Publishing, 2015, pp. 413–448. doi: 10.1016/B978-0-85709-523-7.00015-3.
[57] S. P. Kelly, R. Farlow, and G. Hayward, ‘Applications of through-air ultrasound for rapid NDE scanning in the aerospace industry’, *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 43, no. 4, pp. 581–591, 1996, doi: 10.1109/58.503780.
[58] C. Mineo, ‘Automated NDT inspection for large and complex geometries of composite materials’, 2015, doi: 10.48730/GXQ8-WA04.
[59] S. Pala, Z. Shao, Y. Peng, and L. Lin, ‘Improved Ring-Down Time and Axial Resolution of pMUTs via a Phase-Shift Excitation Scheme’, *Proc. IEEE Int. Conf. Micro Electro Mech. Syst. MEMS*, vol. 2021-January, pp. 390–393, Jan. 2021, doi: 10.1109/MEMS51782.2021.9375227.
[60] I.-Y. Yang *et al.*, ‘Feasibility on fiber orientation detection of unidirectional CFRP composite laminates using one-sided pitch–catch ultrasonic technique’, *Compos. Sci. Technol.*, vol. 69, no. 13, pp. 2042–2047, Oct. 2009, doi: 10.1016/j.compscitech.2009.01.007.
[61] X. L. Han, W. T. Wu, P. Li, and J. Lin, ‘Application of ultrasonic phased array total focusing method in weld inspection using an inclined wedge’, *Proc. 2014 Symp. Piezoelectricity Acoust. Waves Device Appl. SPAWDA 2014*, pp. 114–117, Dec. 2014, doi: 10.1109/SPAWDA.2014.6998539.
[62] R. K. W. Vithanage *et al.*, ‘A Phased Array Ultrasound Roller Probe for Automated in-Process/Interpass Inspection of Multipass Welds’, *IEEE Trans. Ind. Electron.*, vol. 68, no. 12, pp. 12781–12790, Dec. 2021, doi: 10.1109/TIE.2020.3042112.

[63] T. Stratoudaki, M. Clark, and P. D. Wilcox, ‘Laser induced ultrasonic phased array using full matrix capture data acquisition and total focusing method’, *Opt. Express*, vol. 24, pp. 329–348, 2016, doi: 10.1364/OE.24.021921.
[64] E. Duernberger, C. MacLeod, D. Lines, C. Loukas, and M. Vasilev, ‘Adaptive optimisation of multi-aperture ultrasonic phased array imaging for increased inspection speeds of wind turbine blade composite panels’, *NDT E Int.*, vol. 132, p. 102725, Dec. 2022, doi: 10.1016/j.ndteint.2022.102725.
[65] S. C. Wooh and Y. Shi, ‘Optimum beam steering of linear phased arrays’, *Wave Motion*, vol. 29, no. 3, pp. 245–265, 1999, doi: 10.1016/S0165-2125(98)00039-0.
[66] 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 Int.*, vol. 38, no. 8, pp. 701–711, Dec. 2005, doi: 10.1016/J.NDTEINT.2005.04.002.
[67] P. D. Wilcox, ‘Ultrasonic arrays in NDE: Beyond the B-scan’, *AIP Conf. Proc.*, vol. 1511, no. 1, pp. 33–33, Jan. 2013, doi: 10.1063/1.4789029.
[68] R. Spencer, R. Sunderman, and E. Todorov, ‘FMC/TFM experimental comparisons’, *AIP Conf. Proc.*, vol. 1949, no. 1, pp. 020015–020015, Apr. 2018, doi: 10.1063/1.5031512.
[69] P. Shen, Y. Wu, Z. Luo, Z. Wu, J. Jing, and H. Zhang, ‘Advanced Orthogonal Frequency and Phase Modulated Waveform for Ultrasonic Phased Array TFM Detection in CFRP Composites’, *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–10, 2024, doi: 10.1109/TIM.2024.3378307.
[70] J.-C. Grager, H. Mooshofer, and C. U. Grosse, ‘Evaluation of the imaging performance of a CFRP-adapted TFM algorithm’.
[71] V. Pauli and J. Leinonen, ‘Technology survey on NDT of carbon-fiber composites’, 2012. Accessed: Oct. 31, 2023. [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/54515/vaara%20leinonen%20B%208%202012.pdf?sequence=1
[72] 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, pp. 503–507, 2000.
[73] A. Benammar, R. Drai, and A. Guessoum, ‘Detection of delamination defects in CFRP materials using ultrasonic signal processing’, *Ultrasonics*, vol. 48, no. 8, pp. 731–738, Dec. 2008, doi: 10.1016/j.ultras.2008.04.005.
[74] D. Nicholas, D. K. Nassiri, P. Garbutt, and C. R. Hill, ‘Tissue characterization from ultrasound B-scan data’, *Ultrasound Med. Biol.*, vol. 12, no. 2, pp. 135–143, Feb. 1986, doi: 10.1016/0301-5629(86)90018-9.
[75] T. Hasiotis, E. Badogiannis, and N. G. Tsouvalis, ‘Application of Ultrasonic C-Scan Techniques for Tracing Defects in Laminated Composite Materials’, *Stroj. Vestn. – J. Mech. Eng.*, vol. 2011, no. 03, pp. 192–203, Mar. 2011, doi: 10.5545/sv-jme.2010.170.
[76] T. D’Orazio, M. Leo, A. Distante, C. Guaragnella, V. Pianese, and G. Cavaccini, ‘Automatic ultrasonic inspection for internal defect detection in composite materials’, *NDT E Int.*, vol. 41, no. 2, pp. 145–154, Mar. 2008, doi: 10.1016/j.ndteint.2007.08.001.
[77] G. Piao *et al.*, ‘Phased array ultrasonic imaging and characterization of adhesive bonding between thermoplastic composites aided by machine learning’,

Nondestruct. Test. Eval., vol. 38, no. 3, pp. 500–518, May 2023, doi: 10.1080/10589759.2022.2134365.
[78] Y. Chen *et al.*, 'Manufacturing Technology of Lightweight Fiber-Reinforced Composite Structures in Aerospace: Current Situation and toward Intellectualization', *Aerospace*, vol. 10, no. 3, Art. no. 3, Mar. 2023, doi: 10.3390/aerospace10030206.
[79] C. Soutis, 'Carbon fiber reinforced plastics in aircraft construction', *Mater. Sci. Eng. A*, vol. 412, no. 1, pp. 171–176, Dec. 2005, doi: 10.1016/j.msea.2005.08.064.
[80] M. Elkington, Bloom ,D., Ward ,C., Chatzimichali ,A., and K. and Potter, 'Hand layup: understanding the manual process', *Adv. Manuf. Polym. Compos. Sci.*, vol. 1, no. 3, pp. 138–151, Jul. 2015, doi: 10.1080/20550340.2015.1114801.
[81] O. A. Ekuase, N. Anjum, V. O. Eze, and O. I. Okoli, 'A Review on the Out-of-Autoclave Process for Composite Manufacturing', *J. Compos. Sci.*, vol. 6, no. 6, Art. no. 6, Jun. 2022, doi: 10.3390/jcs6060172.
[82] A. P. Mouritz, *Introduction to aerospace materials*. in Woodhead publishing in materials. Oxford (GB): Woodhead publ, 2012.
[83] O. Ley and V. Godinez, 'Non-destructive evaluation (NDE) of aerospace composites: Application of infrared (IR) thermography', in *Non-Destructive Evaluation (NDE) of Polymer Matrix Composites: Techniques and Applications*, Elsevier Ltd, 2013, pp. 309–334. doi: 10.1533/9780857093554.3.309.
[84] F. Heinecke and C. Willberg, 'Manufacturing-Induced Imperfections in Composite Parts Manufactured via Automated Fiber Placement', *J. Compos. Sci.*, vol. 3, no. 2, Art. no. 2, Jun. 2019, doi: 10.3390/jcs3020056.
[85] 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', *Compos. Struct.*, vol. 134, pp. 845–853, Dec. 2015, doi: 10.1016/j.compstruct.2015.08.119.
[86] S. Gholizadeh, 'A review of non-destructive testing methods of composite materials', *Procedia Struct. Integr.*, vol. 1, pp. 50–57, Jan. 2016, doi: 10.1016/j.prostr.2016.02.008.
[87] 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.
[88] I. Solodov, A. Dillenz, and M. Kreutzbruck, 'A new mode of acoustic NDT via resonant air-coupled emission', *J. Appl. Phys.*, vol. 121, no. 24, p. 245101, Jun. 2017, doi: 10.1063/1.4985286.
[89] Z. Zhang, M. Liu, Q. Li, and M. Png, 'Baseline-free defect evaluation of complex-microstructure composites using frequency-dependent ultrasound reflections', *Compos. Part Appl. Sci. Manuf.*, vol. 139, pp. 106090–106090, Dec. 2020, doi: 10.1016/J.COMPOSITESA.2020.106090.
[90] Z. Zhang, M. Liu, Q. Li, and Y. Ang, 'Visualized characterization of diversified defects in thick aerospace composites using ultrasonic B-scan', *Compos. Commun.*, vol. 22, pp. 100435–100435, Dec. 2020, doi: 10.1016/J.COCO.2020.100435.

[91] H. Taheri and A. A. Hassen, ‘Nondestructive Ultrasonic Inspection of Composite Materials: A Comparative Advantage of Phased Array Ultrasonic’, *Appl. Sci.*, vol. 9, no. 8, Art. no. 8, Jan. 2019, doi: 10.3390/app9081628.
[92] A. M. Kokurov and D. E. Subbotin, ‘Ultrasonic detection of manufacturing defects in multilayer composite structures’, *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1023, no. 1, p. 012013, Jan. 2021, doi: 10.1088/1757-899X/1023/1/012013.
[93] M. Vasilev *et al.*, ‘Sensor-Enabled Multi-Robot System for Automated Welding and In-Process Ultrasonic NDE’, *Sens. 2021 Vol 21 Page 5077*, vol. 21, no. 15, pp. 5077–5077, Jul. 2021, doi: 10.3390/S21155077.
[94] C. Mineo, S. G. Pierce, P. I. Nicholson, and I. Cooper, ‘Robotic path planning for non-destructive testing – A custom MATLAB toolbox approach’, *Robot. Comput.-Integr. Manuf.*, vol. 37, pp. 1–12, Feb. 2016, doi: 10.1016/j.rcim.2015.05.003.
[95] C. Mineo, S. G. Pierce, B. Wright, P. I. Nicholson, and I. Cooper, ‘Robotic path planning for non-destructive testing of complex shaped surfaces’, *AIP Conf. Proc.*, vol. 1650, no. 1, pp. 1977–1987, Mar. 2015, doi: 10.1063/1.4914825.
[96] P. Gardner *et al.*, ‘Machine learning at the interface of structural health monitoring and non-destructive evaluation’, *Philos. Trans. R. Soc. A*, vol. 378, no. 2182, Oct. 2020, doi: 10.1098/RSTA.2019.0581.
[97] K. Senthil, A. Arockiarajan, R. Palaninathan, B. Santhosh, and K. M. Usha, ‘Defects in composite structures: Its effects and prediction methods – A comprehensive review’, *Compos. Struct.*, vol. 106, pp. 139–149, Dec. 2013, doi: 10.1016/j.compstruct.2013.06.008.
[98] A. Krishnamoorthy, J. Lilly Mercy, K. S. M. Vineeth, and M. K. Salugu, ‘Delamination Analysis of Carbon Fiber Reinforced Plastic (CFRP) Composite plates by Thermo graphic technique’, *Mater. Today Proc.*, vol. 2, no. 4, pp. 3132–3139, Jan. 2015, doi: 10.1016/j.matpr.2015.07.101.
[99] W. J. Cantwell and J. Morton, ‘The impact resistance of composite materials — a review’, *Composites*, vol. 22, no. 5, pp. 347–362, Sep. 1991, doi: 10.1016/0010-4361(91)90549-V.
[100] K. Dransfield, C. Baillie, and Y.-W. Mai, ‘Improving the delamination resistance of CFRP by stitching—a review’, *Compos. Sci. Technol.*, vol. 50, no. 3, pp. 305–317, Jan. 1994, doi: 10.1016/0266-3538(94)90019-1.
[101] M. J. Suriani, H. Z. Rapi, R. A. Ilyas, M. Petrů, and S. M. Sapuan, ‘Delamination and Manufacturing Defects in Natural Fiber-Reinforced Hybrid Composite: A Review’, *Polymers*, vol. 13, no. 8, Art. no. 8, Jan. 2021, doi: 10.3390/polym13081323.
[102] Y. Li, B. Wang, and L. Zhou, ‘Study on the effect of delamination defects on the mechanical properties of CFRP composites’, *Eng. Fail. Anal.*, vol. 153, p. 107576, Nov. 2023, doi: 10.1016/j.engfailanal.2023.107576.
[103] T. Huang and M. Bobyr, ‘A Review of Delamination Damage of Composite Materials’, *J. Compos. Sci.*, vol. 7, no. 11, Art. no. 11, Nov. 2023, doi: 10.3390/jcs7110468.
[104] J. Reiner and R. Vaziri, ‘8.4 Structural Analysis of Composites With Finite Element Codes: An Overview of Commonly Used Computational Methods’, in *Comprehensive Composite Materials II*, P. W. R. Beaumont and C. H. Zweben, Eds., Oxford: Elsevier, 2018, pp. 61–84. doi: 10.1016/B978-0-12-803581-8.10050-5.

[105] M. McElroy, W. Jackson, R. Olsson, P. Hellström, S. Tsampas, and M. Pankow, 'Interaction of delaminations and matrix cracks in a CFRP plate, Part I: A test method for model validation', *Compos. Part Appl. Sci. Manuf.*, vol. 103, pp. 314–326, Dec. 2017, doi: 10.1016/j.compositesa.2017.09.011.
[106] 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', *Compos. Part Appl. Sci. Manuf.*, vol. 164, p. 107259, Jan. 2023, doi: 10.1016/j.compositesa.2022.107259.
[107] H. Koushyar, S. Alavi-Soltani, B. Minaie, and M. Violette, 'Effects of variation in autoclave pressure, temperature, and vacuum-application time on porosity and mechanical properties of a carbon fiber/epoxy composite*', *J. Compos. Mater.*, vol. 46, no. 16, pp. 1985–2004, Aug. 2012, doi: 10.1177/0021998311429618.
[108] D. Das, S. K. Pradhan, R. K. Nayak, B. K. Nanda, and B. C. Routara, 'Influence of curing time on properties of CFRP composites: A case study', *Mater. Today Proc.*, vol. 26, pp. 344–349, Jan. 2020, doi: 10.1016/j.matpr.2019.12.028.
[109] A. Stamopoulos, K. Tserpes, P. Prucha, and D. Vavrik, 'Evaluation of porosity effects on the mechanical properties of carbon fiber-reinforced plastic unidirectional laminates by X-ray computed tomography and mechanical testing', *J. Compos. Mater.*, vol. 50, no. 15, pp. 2087–2098, Jun. 2016, doi: 10.1177/0021998315602049.
[110] O. Baysallı, A. Cambaz, and Y. F. Görgülü, 'Effects of Porosity on CFRP Repair Performance with Aerospace Applications', *J. Aviat.*, vol. 8, no. 1, Art. no. 1, Feb. 2024, doi: 10.30518/jav.1378148.
[111] I. A. Hakim, S. L. Donaldson, N. G. Meyendorf, and C. E. Browning, 'Porosity Effects on Interlaminar Fracture Behavior in Carbon Fiber-Reinforced Polymer Composites', *Mater. Sci. Appl.*, vol. 8, no. 2, Art. no. 2, Feb. 2017, doi: 10.4236/msa.2017.82011.
[112] 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', *Compos. Part B Eng.*, vol. 78, pp. 86–94, Sep. 2015, doi: 10.1016/j.compositesb.2015.03.048.
[113] R. A. Nargis, D. P. Pulipati, and D. A. Jack, 'Automated Foreign Object Detection for Carbon Fiber Laminates Using High-Resolution Ultrasound Testing', *Materials*, vol. 17, no. 10, Art. no. 10, Jan. 2024, doi: 10.3390/ma17102381.
[114] M. Thor, M. G. R. Sause, and R. M. Hinterhölzl, 'Mechanisms of Origin and Classification of Out-of-Plane Fiber Waviness in Composite Materials—A Review', *J. Compos. Sci.*, vol. 4, no. 3, Art. no. 3, Sep. 2020, doi: 10.3390/jcs4030130.
[115] P. Kulkarni, K. D. Mali, and S. Singh, 'An overview of the formation of fibre waviness and its effect on the mechanical performance of fibre reinforced polymer composites', *Compos. Part Appl. Sci. Manuf.*, vol. 137, p. 106013, Oct. 2020, doi: 10.1016/j.compositesa.2020.106013.
[116] S. Hörrmann, A. Adumitroaie, C. Viechtbauer, and M. Schagerl, 'The effect of fiber waviness on the fatigue life of CFRP materials', *Int. J. Fatigue*, vol. 90, pp. 139–147, Sep. 2016, doi: 10.1016/j.ijfatigue.2016.04.029.
[117] C. Wu *et al.*, 'Influences of in-plane and out-of-plane fiber waviness on mechanical properties of carbon fiber composite laminate', *J. Reinf. Plast.*

Compos., vol. 37, no. 13, pp. 877–891, Jul. 2018, doi: 10.1177/0731684418765981.
[118] R. C. V. V. Gomes *et al.*, 'Qualitative and Quantitative Assessment of Out-of-Plane Waviness in Carbon-Fibre Reinforced Plastics: Comparing Different Non-Destructive Evaluation Modalities', May 17, 2025, *Social Science Research Network, Rochester, NY*: 5258464. doi: 10.2139/ssrn.5258464.
[119] 'All topics'. Accessed: Jul. 15, 2025. [Online]. Available: https://www.ncsc.gov.uk/section/advice-guidance/all-topics
[120] 'Machine learning principles'. Accessed: Jul. 15, 2025. [Online]. Available: https://www.ncsc.gov.uk/collection/machine-learning-principles
[121] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[122] A. Paszke *et al.*, 'Automatic differentiation in PyTorch', Oct. 2017, Accessed: Jul. 08, 2024. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ
[123] M. Abadi *et al.*, 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems'.
[124] J. Gui *et al.*, 'A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9052–9071, Dec. 2024, doi: 10.1109/TPAMI.2024.3415112.
[125] A. Baevski, M. Auli, and A. Mohamed, 'Effectiveness of self-supervised pre-training for speech recognition', May 18, 2020, *arXiv*: arXiv:1911.03912. doi: 10.48550/arXiv.1911.03912.
[126] M. Ravanelli *et al.*, 'Multi-Task Self-Supervised Learning for Robust Speech Recognition', in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6989–6993. doi: 10.1109/ICASSP40776.2020.9053569.
[127] S. Ramesh *et al.*, 'Dissecting self-supervised learning methods for surgical computer vision', *Med. Image Anal.*, vol. 88, p. 102844, Aug. 2023, doi: 10.1016/j.media.2023.102844.
[128] Z. Wang, 'Self-supervised Learning in Computer Vision: A Review', in *Proceedings of the 12th International Conference on Computer Engineering and Networks*, Q. Liu, X. Liu, J. Cheng, T. Shen, and Y. Tian, Eds., Singapore: Springer Nature, 2022, pp. 1112–1121. doi: 10.1007/978-981-19-6901-0_116.
[129] A. K. Shakya, G. Pillai, and S. Chakrabarty, 'Reinforcement learning algorithms: A brief survey', *Expert Syst. Appl.*, vol. 231, p. 120495, Nov. 2023, doi: 10.1016/j.eswa.2023.120495.
[130] E. F. Morales, R. Murrieta-Cid, I. Becerra, and M. A. Esquivel-Basaldua, 'A survey on deep learning and deep reinforcement learning in robotics with a tutorial on deep reinforcement learning', *Intell. Serv. Robot.*, vol. 14, no. 5, pp. 773–805, Nov. 2021, doi: 10.1007/s11370-021-00398-z.
[131] Z. Cao, S. Xu, H. Peng, D. Yang, and R. Zidek, 'Confidence-Aware Reinforcement Learning for Self-Driving Cars', *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7419–7430, Jul. 2022, doi: 10.1109/TITS.2021.3069497.
[132] W. S. McCulloch and W. Pitts, 'A logical calculus of the ideas immanent in nervous activity', *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.

[133] F. Rosenblatt, ‘The perceptron: A probabilistic model for information storage and organization in the brain.’, *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.
[134] K. Fukushima, ‘Cognitron: A self-organizing multilayered neural network’, *Biol. Cybern.*, vol. 20, no. 3, pp. 121–136, Sep. 1975, doi: 10.1007/BF00342633.
[135] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, ‘Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)’, Feb. 22, 2016, *arXiv*: arXiv:1511.07289. doi: 10.48550/arXiv.1511.07289.
[136] D. Hendrycks and K. Gimpel, ‘Gaussian Error Linear Units (GELUs)’, Jun. 2016, doi: 10.48550/arxiv.1606.08415.
[137] S. Ruder, ‘An overview of gradient descent optimization algorithms’, Jun. 15, 2017, *arXiv*: arXiv:1609.04747. doi: 10.48550/arXiv.1609.04747.
[138] J. Kiefer and J. Wolfowitz, ‘Stochastic Estimation of the Maximum of a Regression Function’, *Ann. Math. Stat.*, vol. 23, no. 3, pp. 462–466, Sep. 1952, doi: 10.1214/aoms/1177729392.
[139] J. Duchi, E. Hazan, and Y. Singer, ‘Adaptive Subgradient Methods for Online Learning and Stochastic Optimization’, *J. Mach. Learn. Res.*, vol. 12, no. 61, pp. 2121–2159, 2011.
[140] D. P. Kingma and J. Ba, ‘Adam: A Method for Stochastic Optimization’, 2014, doi: 10.48550/ARXIV.1412.6980.
[141] I. Loshchilov and F. Hutter, ‘Decoupled Weight Decay Regularization’, Jan. 04, 2019, *arXiv*: arXiv:1711.05101. doi: 10.48550/arXiv.1711.05101.
[142] S. Ioffe and C. Szegedy, ‘Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift’, *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, Feb. 2015, doi: 10.48550/arxiv.1502.03167.
[143] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, ‘Gradient-based learning applied to document recognition’, *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
[144] K. Simonyan and A. Zisserman, ‘VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION’, 2015. [Online]. Available: http://www.robots.ox.ac.uk/
[145] A. Dosovitskiy *et al.*, ‘An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale’, Oct. 2020, doi: 10.48550/arxiv.2010.11929.
[146] S. J. Song, H. J. Kim, and H. Cho, ‘Development of an intelligent system for ultrasonic flaw classification in weldments’, *Nucl. Eng. Des.*, vol. 212, no. 1–3, pp. 307–320, Mar. 2002, doi: 10.1016/S0029-5493(01)00495-2.
[147] 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.
[148] N. Amiri, G. H. Farrahi, K. R. Kashyzadeh, and M. Chizari, ‘Applications of ultrasonic testing and machine learning methods to predict the static & fatigue behavior of spot-welded joints’, *J. Manuf. Process.*, vol. 52, pp. 26–34, Apr. 2020, doi: 10.1016/J.JMAPRO.2020.01.047.
[149] E. Ghafarallahi, G. H. Farrahi, and N. Amiri, ‘Acoustic simulation of ultrasonic testing and neural network used for diameter prediction of three-sheet spot welded joints’, *J. Manuf. Process.*, vol. 64, pp. 1507–1516, Apr. 2021, doi: 10.1016/J.JMAPRO.2021.03.012.

[150] S. Niu and V. Srivastava, ‘Simulation trained CNN for accurate embedded crack length, location, and orientation prediction from ultrasound measurements’, *Int. J. Solids Struct.*, vol. 242, May 2022, doi: 10.1016/J.IJSOLSTR.2022.111521.

[151] M. S. Alavijeh, R. Scott, F. Seviaryn, and R. Gr. Maev, ‘Using machine learning to automate ultrasound-based classification of butt-fused joints in medium-density polyethylene gas pipesa)’, *J. Acoust. Soc. Am.*, vol. 150, no. 1, pp. 561–561, Jul. 2021, doi: 10.1121/10.0005656.

[152] F. C. Cruz, E. F. Simas Filho, M. C. S. Albuquerque, I. C. Silva, C. T. T. Farias, and L. L. Gouvêa, ‘Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing’, *Ultrasonics*, vol. 73, pp. 1–8, Jan. 2017, doi: 10.1016/J.ULTRAS.2016.08.017.

[153] N. Munir, H.-J. Kim, S.-J. Song, and S.-S. Kang, ‘Investigation of deep neural network with drop out for ultrasonic flaw classification in weldments †’, *J. Mech. Sci. Technol.*, vol. 32, no. 7, pp. 3073–3080, 2018, doi: 10.1007/s12206-018-0610-1.

[154] 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.

[155] N. Munir, J. Park, H. J. Kim, S. J. Song, and S. S. Kang, ‘Performance enhancement of convolutional neural network for ultrasonic flaw classification by adopting autoencoder’, *NDT E Int.*, vol. 111, pp. 102218–102218, Apr. 2020, doi: 10.1016/J.NDTEINT.2020.102218.

[156] J. G. Kim, C. Jang, and S. S. Kang, ‘Classification of ultrasonic signals of thermally aged cast austenitic stainless steel (CASS) using machine learning (ML) models’, *Nucl. Eng. Technol.*, vol. 54, no. 4, pp. 1167–1174, Apr. 2022, doi: 10.1016/J.NET.2021.09.033.

[157] W. Xu, X. Li, and J. Zhang, ‘Multi-feature fusion imaging via machine learning for laser ultrasonic based defect detection in selective laser melting part’, *Opt. Laser Technol.*, vol. 150, pp. 107918–107918, 2022, doi: 10.1016/j.optlastec.2022.107918.

[158] 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.

[159] P. Zacharis, G. West, G. Dobie, and T. Lardner & Anthony Gachagan, ‘Data-Driven Analysis of Ultrasonic Inspection Data of Pressure Tubes’, *Nucl. Technol.*, vol. 202, no. 3, pp. 153–160, 2018, doi: 10.1080/00295450.2017.1421803.

[160] Y. Guo *et al.*, ‘Fully Convolutional Neural Network With GRU for 3D Braided Composite Material Flaw Detection’, *IEEE Access*, vol. 7, pp. 151180–151188, 2019, doi: 10.1109/ACCESS.2019.2946447.

[161] Y. Yan, D. Liu, B. Gao, G. Y. Tian, and Z. C. Cai, ‘A Deep Learning-Based Ultrasonic Pattern Recognition Method for Inspecting Girth Weld Cracking of Gas Pipeline’, *IEEE Sens. J.*, vol. 20, no. 14, pp. 7997–8006, Jul. 2020, doi: 10.1109/JSEN.2020.2982680.

[162] L. F. M. Rodrigues *et al.*, 'Carburization level identification in industrial HP pipes using ultrasonic evaluation and machine learning', *Ultrasonics*, vol. 94, pp. 145–151, Apr. 2019, doi: 10.1016/J.ULTRAS.2018.10.005.
[163] S.-H. Park, S. Choi, and K.-Y. Jhang, 'Porosity Evaluation of Additively Manufactured Components Using Deep Learning-based Ultrasonic Nondestructive Testing', *Int. J. Precis. Eng. Manuf.-Green Technol.*, vol. 9, pp. 395–407, 2021, doi: 10.1007/s40684-021-00319-6.
[164] S. H. Park, J. Y. Hong, T. Ha, S. Choi, and K. Y. Jhang, 'Deep Learning-Based Ultrasonic Testing to Evaluate the Porosity of Additively Manufactured Parts with Rough Surfaces', *Met. 2021 Vol 11 Page 290*, vol. 11, no. 2, pp. 290–290, Feb. 2021, doi: 10.3390/MET11020290.
[165] Y. Jin *et al.*, 'Numerically Trained Ultrasound AI for Monitoring Tool Degradation', *Adv. Intell. Syst.*, pp. 2100215–2100215, May 2022, doi: 10.1002/AISY.202100215.
[166] J. C. Aldrin and D. S. Forsyth, 'Demonstration of using signal feature extraction and deep learning neural networks with ultrasonic data for detecting challenging discontinuities in composite panels', *AIP Conf. Proc.*, vol. 2102, pp. 230004–230004, May 2019, doi: 10.1063/1.5099716/FORMAT/PDF.
[167] K. Sudheera, N. M. Nandhitha, V. Bhavagna, V. Sai, and N. V. Kumar, 'Deep Learning Techniques for Flaw Characterization in Weld Pieces from Ultrasonic Signals', *Russ. J. Nondestruct. Test.*, vol. 56, no. 10, pp. 820–830, 2020, doi: 10.1134/S1061830920100083.
[168] 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 Int.*, vol. 135, p. 102804, Apr. 2023, doi: 10.1016/j.ndteint.2023.102804.
[169] V. Samaitis, B. Yilmaz, and E. Jasiuniene, 'Adhesive bond quality classification using machine learning algorithms based on ultrasonic pulse-echo immersion data', *J. Sound Vib.*, vol. 546, p. 117457, Mar. 2023, doi: 10.1016/j.jsv.2022.117457.
[170] Y. Duan *et al.*, 'Automatic Air-Coupled Ultrasound Detection of Impact Damages in Fiber-Reinforced Composites Based on One-Dimension Deep Learning Models', *J. Nondestruct. Eval.*, vol. 42, no. 3, p. 79, Aug. 2023, doi: 10.1007/s10921-023-00988-0.
[171] Z. Wang, F. Shi, and F. Zou, 'Deep learning based ultrasonic reconstruction of rough surface morphology', *Ultrasonics*, vol. 138, p. 107265, Mar. 2024, doi: 10.1016/j.ultras.2024.107265.
[172] D. Smagulova, V. Samaitis, and E. Jasiuniene, 'Machine learning based approach for automatic defect detection and classification in adhesive joints', *NDT E Int.*, vol. 148, p. 103221, Dec. 2024, doi: 10.1016/j.ndteint.2024.103221.
[173] K. Cho *et al.*, 'Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation', Sep. 03, 2014, *arXiv*: arXiv:1406.1078. doi: 10.48550/arXiv.1406.1078.
[174] T. Koskinen, I. Virkkunen, · Oskar Siljama, and · Oskari Jessen-Juhler, 'The Effect of Different Flaw Data to Machine Learning Powered Ultrasonic Inspection', *J. Nondestruct. Eval.*, vol. 40, pp. 24–24, 2021, doi: 10.1007/s10921-021-00757-x.
[175] 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, pp. 1–13, Sep. 2021, doi: 10.1007/S10921-021-00796-4/FIGURES/6.
[176] D. Medak, L. Posilovic, M. Subasic, M. Budimir, and S. Loncaric, 'Automated Defect Detection from Ultrasonic Images Using Deep Learning', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 68, no. 10, pp. 3126–3134, Oct. 2021, doi: 10.1109/TUFFC.2021.3081750.
[177] D. Medak, L. Posilovic, M. Subasic, M. Budimir, and S. Loncaric, 'Deep Learning-Based Defect Detection from Sequences of Ultrasonic B-Scans', *IEEE Sens. J.*, vol. 22, no. 3, pp. 2456–2463, Feb. 2022, doi: 10.1109/JSEN.2021.3134452.
[178] 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.
[179] L. Posilovic, D. Medak, M. Subasic, T. Petkovic, M. Budimir, and S. Loncaric, 'Flaw detection from ultrasonic images using YOLO and SSD', *Int. Symp. Image Signal Process. Anal. ISPA*, vol. 2019-September, pp. 163–168, Sep. 2019, doi: 10.1109/ISPA.2019.8868929.
[180] L. Posilović, D. Medak, F. Milković, M. Subašić, M. Budimir, and S. Lončarić, 'Deep learning-based anomaly detection from ultrasonic images', *Ultrasonics*, vol. 124, pp. 106737–106737, Aug. 2022, doi: 10.1016/J.ULTRAS.2022.106737.
[181] F. Milković, B. Filipović, M. Subašić, T. Petković, S. Lončarić, and M. Budimir, 'Ultrasound Anomaly Detection Based on Variational Autoencoders', in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Sep. 2021, pp. 225–229. doi: 10.1109/ISPA52656.2021.9552041.
[182] H. Chen and J. Tao, 'Utilizing improved YOLOv8 based on SPD-BRSA-AFPN for ultrasonic phased array non-destructive testing', *Ultrasonics*, vol. 142, p. 107382, Aug. 2024, doi: 10.1016/j.ultras.2024.107382.
[183] J. Mendikute *et al.*, 'Defect detection in wind turbine blades applying Convolutional Neural Networks to Ultrasonic Testing', *NDT E Int.*, vol. 154, p. 103359, Sep. 2025, doi: 10.1016/j.ndteint.2025.103359.
[184] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, 'GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training', Nov. 13, 2018, *arXiv*: arXiv:1805.06725. doi: 10.48550/arXiv.1805.06725.
[185] R. Sunkara and T. Luo, 'No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects', Aug. 07, 2022, *arXiv*: arXiv:2208.03641. doi: 10.48550/arXiv.2208.03641.
[186] C. Wang and C. Zhong, 'Adaptive Feature Pyramid Networks for Object Detection', *IEEE Access*, vol. 9, pp. 107024–107032, 2021, doi: 10.1109/ACCESS.2021.3100369.
[187] C. Li *et al.*, 'Intelligent damage recognition of composite materials based on deep learning and ultrasonic testing', *AIP Adv.*, vol. 11, no. 12, pp. 125227–125227, Dec. 2021, doi: 10.1063/5.0063615.
[188] B. Filipović, F. Milković, M. Subašić, S. Lončarić, T. Petković, and M. Budimir, 'Automated Ultrasonic Testing of Materials based on C-scan Flaw Classification', in *2021 12th International Symposium on Image and Signal*

Processing and Analysis (ISPA), Sep. 2021, pp. 230–234. doi: 10.1109/ISPA52656.2021.9552056.

[189] X. Zhang and J. Saniie, 'Material Texture Recognition using Ultrasonic Images with Transformer Neural Networks', in *2021 IEEE International Conference on Electro Information Technology (EIT)*, May 2021, pp. 1–5. doi: 10.1109/EIT51626.2021.9491908.

[190] P. Trouvé-Peloux, B. Abeloos, A. Ben Fekih, C. Trottier, and J.-M. Roche, 'Benefit of Neural Network for the Optimization of Defect Detection on Composite Material Using Ultrasonic Non Destructive Testing', presented at the 2021 48th Annual Review of Progress in Quantitative Nondestructive Evaluation, American Society of Mechanical Engineers Digital Collection, Jan. 2022. doi: 10.1115/QNDE2021-75925.

[191] A. Yunker, R. Lake, R. Kettimuthu, and Z. Kral, 'Comparative Study on Deep Learning Methods for Defect Identification and Classification in Composite Aerostructure Material', presented at the 2023 50th Annual Review of Progress in Quantitative Nondestructive Evaluation, American Society of Mechanical Engineers Digital Collection, Jul. 2023. doi: 10.1115/QNDE2023-108602.

[192] A. Gulsen, B. Kolukisa, A. T. Ozdemir, B. Bakir-Gungor, and V. C. Gungor, 'Defect classification of composite materials using transfer learning methods', *Nondestruct. Test. Eval.*, vol. 0, no. 0, pp. 1–17, doi: 10.1080/10589759.2024.2422527.

[193] D. Nguyen, P. Davidson, K. DeMille, and V. Ranatunga, 'Machine learning based segmentation of delamination patterns from sparse ultrasound data of barely visible impact damage in composites', *J. Compos. Mater.*, vol. 59, no. 3, pp. 321–330, Feb. 2025, doi: 10.1177/00219983241292779.

[194] J. Ye, S. Ito, and N. Toyama, 'Computerized Ultrasonic Imaging Inspection: From Shallow to Deep Learning', *Sensors*, vol. 18, no. 11, 2018, doi: 10.3390/s18113820.

[195] 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.

[196] J. Ye and N. Toyama, 'Automatic defect detection for ultrasonic wave propagation imaging method using spatio-temporal convolution neural networks', *Struct. Health Monit.*, vol. 0, no. 0, pp. 1–18, 2022, doi: 10.1177/14759217211073503.

[197] M. Słónski, K. Schabowicz, and E. Krawczyk, 'Detection of Flaws in Concrete Using Ultrasonic Tomography and Convolutional Neural Networks', *Materials*, vol. 13, no. 7, pp. 1557–1557, 2020, doi: 10.3390/ma13071557.

[198] 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. Ferroelectr. Freq. Control*, vol. 68, no. 5, pp. 1854–1865, May 2021, doi: 10.1109/TUFFC.2020.3045847.

[199] 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 Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 4, pp. 1485–1496, Apr. 2022, doi: 10.1109/TUFFC.2022.3151397.

[200] R. J. Pyle, R. R. Hughes, A. A. S. Ali, and P. D. Wilcox, 'Uncertainty Quantification for Deep Learning in Ultrasonic Crack Characterization', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 7, pp. 2339–2351, Jul. 2022, doi: 10.1109/TUFFC.2022.3176926.

[201] R. J. Pyle, R. R. Hughes, and P. D. Wilcox, 'Interpretable and Explainable Machine Learning for Ultrasonic Defect Sizing', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 70, no. 4, pp. 277–290, Apr. 2023, doi: 10.1109/TUFFC.2023.3248968.

[202] 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 Int.*, vol. 141, p. 102978, Jan. 2024, doi: 10.1016/j.ndteint.2023.102978.

[203] H. Hervé-Côte, F. Dupont-Marillia, and P. Bélanger, 'Automatic flaw detection in sectoral scans using machine learning', *Ultrasonics*, p. 107316, Apr. 2024, doi: 10.1016/j.ultras.2024.107316.

[204] S. Zhang and Y. Zhang, 'Automated weld defect segmentation from phased array ultrasonic data based on U-net architecture', *NDT E Int.*, vol. 146, p. 103165, Sep. 2024, doi: 10.1016/j.ndteint.2024.103165.

[205] W. Cao *et al.*, 'The detection of PAUT pseudo defects in ultra-thick stainless-steel welds with a multimodal deep learning model', *Measurement*, vol. 241, p. 115662, Feb. 2025, doi: 10.1016/j.measurement.2024.115662.

[206] U. Kamath, J. Liu, and J. Whitaker, 'Transfer Learning: Domain Adaptation', in *Deep Learning for NLP and Speech Recognition*, U. Kamath, J. Liu, and J. Whitaker, Eds., Cham: Springer International Publishing, 2019, pp. 495–535. doi: 10.1007/978-3-030-14596-5_11.

[207] Y. Gong, H. Shao, J. Luo, and Z. Li, 'A deep transfer learning model for inclusion defect detection of aeronautics composite materials', *Compos. Struct.*, vol. 252, pp. 112681–112681, Nov. 2020, doi: 10.1016/J.COMPSTRUCT.2020.112681.

[208] Y. Gong, J. Luo, H. Shao, K. He, and W. Zeng, 'Automatic Defect Detection for Small Metal Cylindrical Shell Using Transfer Learning and Logistic Regression', *J. Nondestruct. Eval.*, vol. 39, no. 1, pp. 1–13, Mar. 2020, doi: 10.1007/S10921-020-0668-4/FIGURES/14.

[209] N. Saeed, N. King, Z. Said, and M. A. Omar, 'Automatic defects detection in CFRP thermograms, using convolutional neural networks and transfer learning', *Infrared Phys. Technol.*, vol. 102, p. 103048, Nov. 2019, doi: 10.1016/j.infrared.2019.103048.

[210] A. Krizhevsky, 'Learning Multiple Layers of Features from Tiny Images', 2009. Accessed: Nov. 07, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086

[211] S. Kumaresan, K. S. J. Aultrin, S. S. Kumar, and M. D. Anand, 'Transfer Learning with CNN for Classification of Weld Defect', *IEEE Access*, vol. 9, pp. 95097–95108, 2021, doi: 10.1109/ACCESS.2021.3093487.

[212] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, 'A New Image Recognition and Classification Method Combining Transfer Learning Algorithm and MobileNet Model for Welding Defects', *IEEE Access*, vol. 8, pp. 119951–119960, 2020, doi: 10.1109/ACCESS.2020.3005450.

[213] 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.
[214] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang, and B. Hu, 'Liver Fibrosis Classification Based on Transfer Learning and FCNet for Ultrasound Images', *IEEE Access*, vol. 5, pp. 5804–5810, 2017, doi: 10.1109/ACCESS.2017.2689058.
[215] A. Hijab, M. A. Rushdi, M. M. Gomaa, and A. Eldeib, 'Breast Cancer Classification in Ultrasound Images using Transfer Learning', *Int. Conf. Adv. Biomed. Eng. ICABME*, vol. 2019-October, Oct. 2019, doi: 10.1109/ICABME47164.2019.8940291.
[216] Y. Na *et al.*, 'Advances of Machine Learning in Phased Array Ultrasonic Non-Destructive Testing: A Review', *AI*, vol. 6, no. 6, Art. no. 6, Jun. 2025, doi: 10.3390/ai6060124.
[217] J. L. Tai, M. T. H. Sultan, A. Łukaszewicz, J. Józwik, Z. Oksiuta, and F. S. Shahar, 'Recent Trends in Non-Destructive Testing Approaches for Composite Materials: A Review of Successful Implementations', *Materials*, vol. 18, no. 13, p. 3146, Jul. 2025, doi: 10.3390/ma18133146.
[218] 'Machine Learning Glossary', Google for Developers. Accessed: Oct. 02, 2023. [Online]. Available: https://developers.google.com/machine-learning/glossary
[219] J. H. Kurz, A. Jüngert, S. Dugan, G. Dobmann, and C. Boller, 'Reliability considerations of NDT by probability of detection (POD) determination using ultrasound phased array', *Eng. Fail. Anal.*, vol. 35, pp. 609–617, Dec. 2013, doi: 10.1016/J.ENGFAILANAL.2013.06.008.
[220] D. Mery, 'Aluminum Casting Inspection using Deep Object Detection Methods and Simulated Ellipsoidal Defects', *Mach. Vis. Appl.*, vol. 32, no. 3, p. 72, Apr. 2021, doi: 10.1007/s00138-021-01195-5.
[221] Olympus-ims, 'RollerFORM: Phased Array Wheel Probe manual', 2023, [Online]. Available: https://www.olympus-ims.com/en/rollerform/
[222] 'MicoPulse 6PA | Phased Array Ultrasonic Technology | Peak NDT', [Online]. Available: https://www.peakndt.com/products/micropulse-6pa/
[223] K. Ono, 'Ultrasonic Attenuation of Carbon-Fiber Reinforced Composites', *J. Compos. Sci.*, vol. 7, no. 11, Art. no. 11, Nov. 2023, doi: 10.3390/jcs7110479.
[224] KUKA Robotics, 'KUKA KR90 R3100 extra HA specification manual', 2023, [Online]. Available: https://www.kuka.com/-/media/kuka-downloads/imported/8350ff3ca11642998dbdc81dcc2ed44c/0000208694_en.pdf
[225] R. Zimermann *et al.*, 'Collaborative Robotic Wire + Arc Additive Manufacture and Sensor-Enabled In-Process Ultrasonic Non-Destructive Evaluation', *Sens. 2022 Vol 22 Page 4203*, vol. 22, no. 11, pp. 4203–4203, May 2022, doi: 10.3390/S22114203.
[226] Schunk, 'SCHUNK Force Torque sensors manual', 2023, [Online]. Available: https://schunk.com/us/en/automation-technology/force/torque-sensors/ft/ftn-gamma-si-130-10/p/EPIM_ID-30865
[227] C. R. Harris *et al.*, 'Array programming with NumPy', *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[228] P. Blain *et al.*, 'Artificial defects in CFRP composite structure for thermography and shearography nondestructive inspection',

https://doi.org/10.1117/12.2271701, vol. 10449, no. 13, pp. 562–571, Jun. 2017, doi: 10.1117/12.2271701.
[229] 'EXTENDE, Experts in Non Destructive Testing Simulation with CIVA Software', [Online]. Available: https://www.extende.com/
[230] P. Huthwaite, 'Accelerated finite element elastodynamic simulations using the GPU', *J. Comput. Phys.*, vol. 257, pp. 687–707, Jan. 2014, doi: 10.1016/J.JCP.2013.10.017.
[231] P. D. Wilcox *et al.*, 'Fusion of multi-view ultrasonic data for increased detection performance in non-destructive evaluation', *Proc. R. Soc. A*, vol. 476, no. 2243, Nov. 2020, doi: 10.1098/RSPA.2020.0086.
[232] R. Girshick, 'Fast R-CNN', 2015, [Online]. Available: https://github.com/rbgirshick/
[233] S. Ren, K. He, R. Girshick, and J. Sun, 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2015, doi: 10.48550/arxiv.1506.01497.
[234] M. Tan, R. Pang, and Q. V. Le, 'EfficientDet: Scalable and Efficient Object Detection', *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 10778–10787, Nov. 2019, doi: 10.48550/arxiv.1911.09070.
[235] J. Redmon and A. Farhadi, 'YOLOv3: An Incremental Improvement', Apr. 2018, doi: 10.48550/arxiv.1804.02767.
[236] J. Redmon and A. Farhadi, 'YOLO9000: Better, Faster, Stronger', *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017*, vol. 2017-January, pp. 6517–6525, Dec. 2016, doi: 10.48550/arxiv.1612.08242.
[237] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, 'YOLOv4: Optimal Speed and Accuracy of Object Detection', Apr. 2020, doi: 10.48550/arxiv.2004.10934.
[238] A. Hauffe, F. Hähnel, and K. Wolf, 'Comparison of algorithms to quantify the damaged area in CFRP ultrasonic scans', *Compos. Struct.*, vol. 235, pp. 111791–111791, Mar. 2020, doi: 10.1016/J.COMPSTRUCT.2019.111791.
[239] X. Li, Y. Wang, P. Ni, H. Hu, and Y. Song, 'Flaw sizing using ultrasonic C-scan imaging with dynamic thresholds', *Insight Non-Destr. Test. Cond. Monit.*, vol. 59, no. 11, pp. 603–608, Nov. 2017, doi: 10.1784/INSI.2017.59.11.603.
[240] S. Barut, V. Bissauge, G. Ithurralde, and W. Claassens, 'Computer-aided analysis of ultrasound data to speed-up the release of aerospace CFRP components', presented at the 18th World Conference on Nondestructive Testing, Durban, South Africa: e-Journal of Nondestructive Testing Vol. 17(7), Apr. 2012.
[241] Y. Song, J. A. Turner, Z. Peng, C. Chen, and X. Li, 'Enhanced Ultrasonic Flaw Detection Using an Ultrahigh Gain and Time-Dependent Threshold', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 65, no. 7, pp. 1214–1225, Jul. 2018, doi: 10.1109/TUFFC.2018.2827464.
[242] A. Dogandžić, N. Eua-Anant, and A. D. Dogandži´c, 'Defect Detection in Correlated Noise', *AIP Conf. Proc.*, vol. 700, no. 1, pp. 628–628, Apr. 2004, doi: 10.1063/1.1711680.
[243] A. Wronkowicz, A. Katunin, and K. Dragan, 'Ultrasonic C-Scan Image Processing Using Multilevel Thresholding for Damage Evaluation in Aircraft Vertical Stabilizer', *Int. J. Image Graph. Signal Process.*, 2015.

[244] B. C. F. de Oliveira, P. Nienheysen, C. R. Baldo, A. A. Gonçalves, and R. H. Schmitt, 'Improved impact damage characterisation in CFRP samples using the fusion of optical lock-in thermography and optical square-pulse shearography images', *NDT E Int.*, vol. 111, pp. 102215–102215, Apr. 2020, doi: 10.1016/J.NDTEINT.2020.102215.

[245] A. Osman, V. Kaftandjian, U. Hassler, M. Rehak, and R. Hanke, 'Steps Toward Automated 3D Evaluation of Ultrasound Data', 2010, [Online]. Available: https://www.researchgate.net/publication/268296913

[246] H.-F. Ng, 'Automatic thresholding for defect detection', *Pattern Recognit. Lett.*, vol. 27, no. 14, pp. 1644–1649, Oct. 2006, doi: 10.1016/j.patrec.2006.03.009.

[247] C. Li *et al.*, 'Intelligent damage recognition of composite materials based on deep learning and ultrasonic testing', *AIP Adv.*, vol. 11, no. 12, p. 125227, Dec. 2021, doi: 10.1063/5.0063615.

[248] T. Y. Lin *et al.*, 'Microsoft COCO: Common Objects in Context', *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 8693 LNCS, no. PART 5, pp. 740–755, May 2014, doi: 10.48550/arxiv.1405.0312.

[249] C. U. Grosse *et al.*, 'Comparison of NDT Techniques to Evaluate CFRP-Results Obtained in a MAIzfp Round Robin Test', 2016, [Online]. Available: http://creativecommons.org/licenses/by-nd/3.0/

[250] F. Bolelli, S. Allegretti, L. Baraldi, and C. Grana, 'Spaghetti Labeling: Directed Acyclic Graphs for Block-Based Connected Components Labeling'.

[251] D. Freedman and P. Diaconis, 'On the histogram as a density estimator:L2 theory', *Z. Für Wahrscheinlichkeitstheorie Verwandte Geb.*, vol. 57, no. 4, pp. 453–476, Dec. 1981, doi: 10.1007/BF01025868/METRICS.

[252] P. Virtanen *et al.*, 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[253] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, 'Feature Pyramid Networks for Object Detection', Dec. 2016, doi: 10.48550/arxiv.1612.03144.

[254] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You Only Look Once: Unified, Real-Time Object Detection', *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 779–788, Jun. 2015, doi: 10.48550/arxiv.1506.02640.

[255] G. Jocher *et al.*, 'YOLOv5 SOTA Realtime Instance Segmentation', Nov. 2022, doi: 10.5281/ZENODO.7347926.

[256] C. Y. Wang, H. Y. Mark Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, 'CSPNet: A New Backbone that can Enhance Learning Capability of CNN', *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop*, vol. 2020-June, pp. 1571–1580, Nov. 2019, doi: 10.48550/arxiv.1911.11929.

[257] K. He, X. Zhang, S. Ren, and J. Sun, 'Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition', *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 8691 LNCS, no. PART 3, pp. 346–361, Jun. 2014, doi: 10.1007/978-3-319-10578-9_23.

[258] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, 'Path Aggregation Network for Instance Segmentation', *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8759–8768, Mar. 2018, doi: 10.48550/arxiv.1803.01534.

[259] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, 'Focal Loss for Dense Object Detection', *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 2999–3007, Dec. 2017, doi: 10.1109/ICCV.2017.324.
[260] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, 'End-to-End Object Detection with Transformers', *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 12346 LNCS, pp. 213–229, May 2020, doi: 10.48550/arxiv.2005.12872.
[261] J. Maurício, I. Domingues, and J. Bernardino, 'Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review', *Appl. Sci.*, vol. 13, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/app13095521.
[262] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, 'A cost focused framework for optimizing collection and annotation of ultrasound datasets', *Biomed. Signal Process. Control*, vol. 92, p. 106048, Jun. 2024, doi: 10.1016/j.bspc.2024.106048.
[263] S. McKnight *et al.*, 'Advancing Carbon Fiber Composite Inspection: Deep Learning-Enabled Defect Localization and Sizing via 3-D U-Net Segmentation of Ultrasonic Data', *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 71, no. 9, pp. 1106–1119, Sep. 2024, doi: 10.1109/TUFFC.2024.3408314.
[264] Y. Wu and X. Zhu, 'Rail Defect Detection Using Ultrasonic A-Scan Data and Deep Autoencoder', *https://doi.org/10.1177/03611981221150923*, pp. 036119812211509–036119812211509, Jan. 2023, doi: 10.1177/03611981221150923.
[265] I. Kraljevski, F. Duckhorn, M. Barth, C. Tschoepe, F. Schubert, and M. Wolff, 'Autoencoder-based Ultrasonic NDT of Adhesive Bonds', *Proc. IEEE Sens.*, vol. 2021-October, 2021, doi: 10.1109/SENSORS47087.2021.9639864.
[266] J. M. Ha, H. M. Seung, and W. Choi, 'Autoencoder-based detection of near-surface defects in ultrasonic testing', *Ultrasonics*, vol. 119, pp. 106637–106637, Feb. 2022, doi: 10.1016/J.ULTRAS.2021.106637.
[267] D. P. Kingma and M. Welling, 'Auto-Encoding Variational Bayes', Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.
[268] O. Ronneberger, P. Fischer, and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation', May 18, 2015, *arXiv*: arXiv:1505.04597. doi: 10.48550/arXiv.1505.04597.
[269] Y. Wu, X. Zhu, and J. Baillargeon, 'Deep Autoencoder for Ultrasound-Based Rail Flaw Detection', *Proc. 2022 Jt. Rail Conf. JRC 2022*, Jun. 2022, doi: 10.1115/JRC2022-79554.
[270] S. Barut and N. Dominguez, 'NDT Diagnosis Automation: a Key to Efficient Production in the Aeronautic Industry', *E-J. Nondestruct. Test.*, vol. 21, no. 07, Jul. 2016, Accessed: Aug. 05, 2023. [Online]. Available: https://www.ndt.net/search/docs.php3?id=19184&msgID=0&rootID=0
[271] J. C. Aldrin, C. Coughlin, D. S. Forsyth, and J. T. Welter, 'Progress on the development of automated data analysis algorithms and software for ultrasonic inspection of composites', *AIP Conf. Proc.*, vol. 1581, no. 1, pp. 1920–1927, Feb. 2014, doi: 10.1063/1.4865058.
[272] 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 Int.*, vol. 122, p. 102478, Sep. 2021, doi: 10.1016/j.ndteint.2021.102478.
[273] K. Lee and V. Estivill-Castro, 'Feature extraction and gating techniques for ultrasonic shaft signal classification', *Appl. Soft Comput.*, vol. 7, no. 1, pp. 156–165, Jan. 2007, doi: 10.1016/j.asoc.2005.05.003.
[274] 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.
[275] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, 'Improved Techniques for Training GANs', Jun. 10, 2016, *arXiv*: arXiv:1606.03498. doi: 10.48550/arXiv.1606.03498.
[276] A. Radford, L. Metz, and S. Chintala, 'Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks', *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, Nov. 2015, doi: 10.48550/arxiv.1511.06434.
[277] A. Odena, V. Dumoulin, and C. Olah, 'Deconvolution and Checkerboard Artifacts', *Distill*, vol. 1, no. 10, p. e3, Oct. 2016, doi: 10.23915/distill.00003.
[278] L. McInnes, J. Healy, and J. Melville, 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', Sep. 18, 2020, *arXiv*: arXiv:1802.03426. doi: 10.48550/arXiv.1802.03426.
[279] P. Dahal, 'Learning Embedding Space for Clustering From Deep Representations', in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 3747–3755. doi: 10.1109/BigData.2018.8622629.
[280] Madhumita and S. Paul, 'Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping', *Comput. Biol. Med.*, vol. 148, p. 105832, Sep. 2022, doi: 10.1016/j.compbiomed.2022.105832.
[281] D. Chardin, C. Gille, T. Pourcher, O. Humbert, and M. Barlaud, 'Learning a confidence score and the latent space of a new supervised autoencoder for diagnosis and prognosis in clinical metabolomic studies', *BMC Bioinformatics*, vol. 23, no. 1, p. 361, Sep. 2022, doi: 10.1186/s12859-022-04900-x.
[282] C. Breen, F. Guild, and M. Pavier, 'Impact of thick CFRP laminates: the effect of impact velocity', *Compos. Part Appl. Sci. Manuf.*, vol. 36, no. 2, pp. 205–211, Feb. 2005, doi: 10.1016/j.compositesa.2004.06.005.
[283] S. Nilsson, A. Bredberg, and L. E. Asp, 'Effects of CFRP laminate thickness on bending after impact strength', *Plast. Rubber Compos.*, vol. 38, no. 2–4, pp. 61–66, May 2009, doi: 10.1179/174328909X387801.
[284] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'.
[285] I. Ndiour, N. Ahuja, U. Genc, and O. Tickoo, 'FRE: A Fast Method For Anomaly Detection And Segmentation', Nov. 22, 2022, *arXiv*: arXiv:2211.12650. doi: 10.48550/arXiv.2211.12650.
[286] L. Posilović, D. Medak, M. Subašić, M. Budimir, and S. Lončarić, 'Generating ultrasonic images indistinguishable from real images using Generative Adversarial Networks', *Ultrasonics*, vol. 119, pp. 106610–106610, Feb. 2022, doi: 10.1016/J.ULTRAS.2021.106610.

[287] W. Zhou *et al.*, ‘Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images’, *Nat. Commun.*, vol. 12, no. 1, p. 1259, Feb. 2021, doi: 10.1038/s41467-021-21466-z.
[288] M. Bertovic, ‘A human factors perspective on the use of automated aids in the evaluation of NDT data’, *AIP Conf. Proc.*, vol. 1706, no. 1, pp. 020003–020003, Feb. 2016, doi: 10.1063/1.4940449.
[289] M. Vagia, A. A. Transeth, and S. A. Fjerdingen, ‘A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed?’, *Appl. Ergon.*, vol. 53, pp. 190–202, Mar. 2016, doi: 10.1016/j.apergo.2015.09.013.
[290] M. R. Endsley and E. O. Kiris, ‘The Out-of-the-Loop Performance Problem and Level of Control in Automation’, *Hum. Factors*, vol. 37, no. 2, pp. 381–394, Jun. 1995, doi: 10.1518/001872095779064555.
[291] R. Parasuraman, M. Mouloua, and R. Molloy, ‘Effects of Adaptive Task Allocation on Monitoring of Automated Systems’, *Hum. Factors*, vol. 38, no. 4, pp. 665–679, Dec. 1996, doi: 10.1518/001872096778827279.
[292] L. Mui, M. Mohtashemi, and A. Halberstadt, ‘Notions of reputation in multi-agents systems: a review’, in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, in AAMAS ’02. New York, NY, USA: Association for Computing Machinery, Jul. 2002, pp. 280–287. doi: 10.1145/544741.544807.
[293] M. Yin, J. Wortman Vaughan, and H. Wallach, ‘Understanding the Effect of Accuracy on Trust in Machine Learning Models’, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, in CHI ’19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12. doi: 10.1145/3290605.3300509.
[294] G. Bravo-Rocca, P. Liu, J. Guitart, A. Dholakia, D. Ellison, and M. Hodak, ‘Human-in-the-loop online multi-agent approach to increase trustworthiness in ML models through trust scores and data augmentation’, May 02, 2022, *arXiv*: arXiv:2204.14255. doi: 10.48550/arXiv.2204.14255.
[295] A. Kore, *Designing Human-Centric AI Experiences: Applied UX Design for Artificial Intelligence*. in Design Thinking. Berkeley, CA: Apress, 2022. doi: 10.1007/978-1-4842-8088-1.
[296] M. T. Ribeiro, S. Singh, and C. Guestrin, ‘“Why Should I Trust You?”: Explaining the Predictions of Any Classifier’, Aug. 09, 2016, *arXiv*: arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938.
[297] S. Lundberg and S.-I. Lee, ‘A Unified Approach to Interpreting Model Predictions’, Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.
[298] 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.
[299] S. McKnight *et al.*, ‘3-DUSSS: 3-Dimensional Ultrasonic Self Supervised Segmentation’, Nov. 12, 2024, *arXiv*: arXiv:2411.07835. doi: 10.48550/arXiv.2411.07835.
[300] S. McKnight, ‘Deep learning for ultrasonic non-destructive evaluation of aerospace composites’.

[301] S. R. Khan, D. Al Rijjal, A. Piro, and M. B. Wheeler, ‘Integration of AI and traditional medicine in drug discovery’, *Drug Discov. Today*, vol. 26, no. 4, pp. 982–992, Apr. 2021, doi: 10.1016/j.drudis.2021.01.008.

[302] R. Najjar, ‘Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging’, *Diagnostics*, vol. 13, no. 17, Art. no. 17, Jan. 2023, doi: 10.3390/diagnostics13172760.

[303] A. Mansourzadeh and S. Rasouli, ‘The Future of Medical Education: A Review of the Opportunities and Challenges of Artificial Intelligence Integration’, *Med. Educ. Bull.*, vol. 5, no. 2, pp. 973–982, Dec. 2024, doi: 10.22034/meb.2024.491888.1102.

[304] ‘Multi-Frame Matrix Capture Common File Format | Research | University of Bristol’. Accessed: Jul. 18, 2025. [Online]. Available: https://www.bristol.ac.uk/research/groups/ndt/projects/multi-frame-matrix/

[305] Z. Ren, S. Wang, and Y. Zhang, ‘Weakly supervised machine learning’, *CAAI Trans. Intell. Technol.*, vol. 8, no. 3, pp. 549–580, 2023, doi: 10.1049/cit2.12216.

[306] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, ‘YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors’, Jul. 06, 2022, *arXiv*: arXiv:2207.02696. doi: 10.48550/arXiv.2207.02696.

[307] A. Wang *et al.*, ‘YOLOv10: Real-Time End-to-End Object Detection’, Oct. 30, 2024, *arXiv*: arXiv:2405.14458. doi: 10.48550/arXiv.2405.14458.

[308] Z. Qiu, Z. Zhao, S. Chen, J. Zeng, Y. Huang, and B. Xiang, ‘Application of an Improved YOLOv5 Algorithm in Real-Time Detection of Foreign Objects by Ground Penetrating Radar’, *Remote Sens. 2022 Vol 14 Page 1895*, vol. 14, no. 8, pp. 1895–1895, Apr. 2022, doi: 10.3390/RS14081895.

[309] M. Sohan, T. Sai Ram, and Ch. V. Rami Reddy, ‘A Review on YOLOv8 and Its Advancements’, in *Data Intelligence and Cognitive Informatics*, I. J. Jacob, S. Piramuthu, and P. Falkowski-Gilski, Eds., Singapore: Springer Nature, 2024, pp. 529–545. doi: 10.1007/978-981-99-7962-2_39.