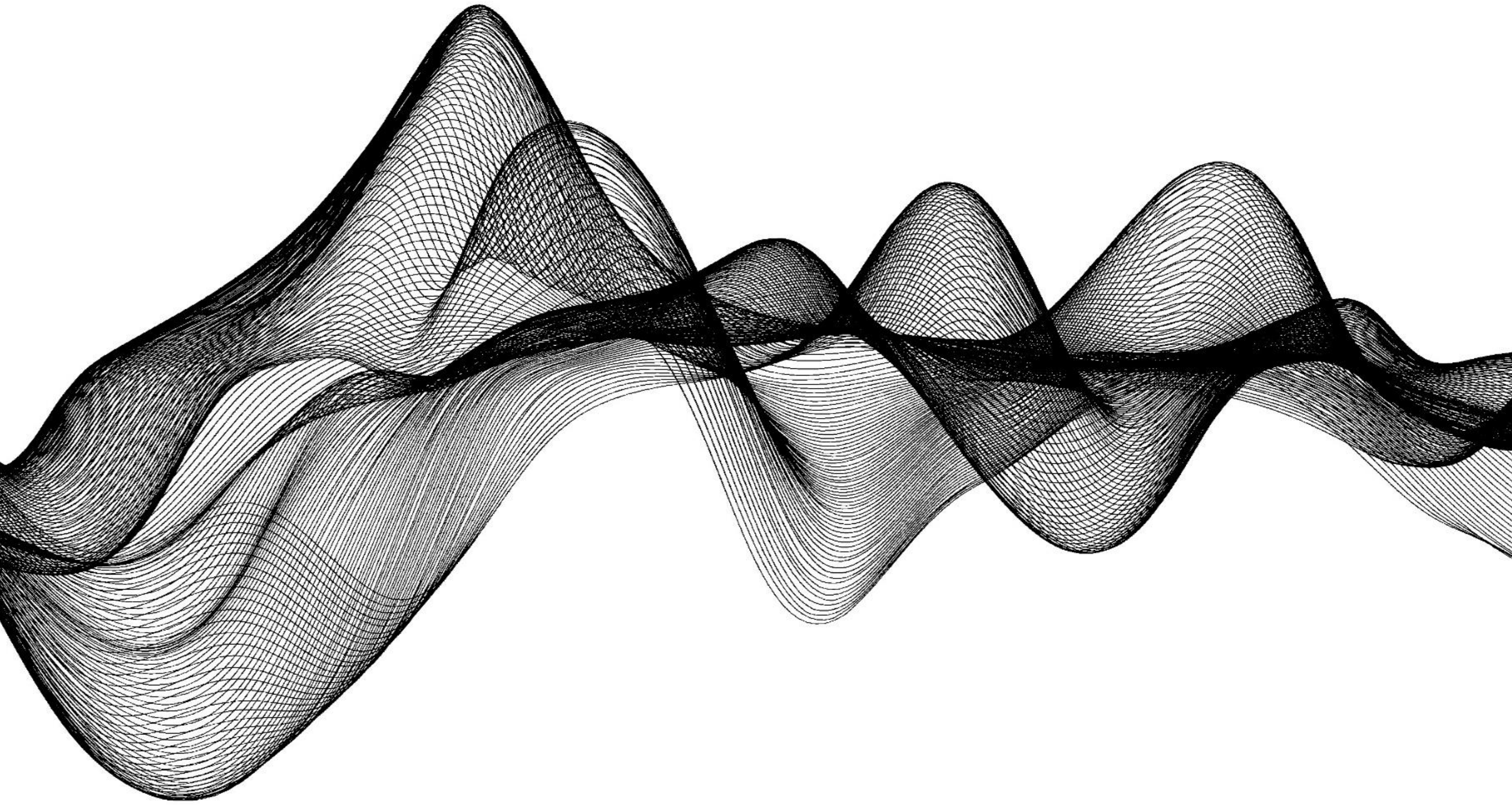
Developing a level-1B qualifiable CNN for in-situ ultrasonic damage classification of aerospace composite structures

S.D.F. Schmidt



TU Delft Faculty of Aerospace Engineering

Cover Image by GarryKillian on Freepik

Developing a level-1B qualifiable CNN for in-situ ultrasonic damage classification of aerospace composite structures

An in-depth evaluation on the end-to-end process of developing a data-driven tool

by

S.D.F. Schmidt

To obtain the degree of Master of Science
at the
Delft University of Technology,
To be defended publicly on: Wednsday, August 30, 2023 at 12:30.

Student number:	4671724	
Project duration:	November 21, 2022 – August 3, 2023	
Thesis committee:	Dr. R. Groves,	TU Delft, Supervisor
	Dr.ir. R. C. Alderliesten,	TU Delft
	Dr. M. L. Baptista,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Abstract

This paper examines the end-to-end development process for a Convolution Neural Network (CNN) based damage classification tool for ultrasonic inspection of aerospace-grade composite structures. The recent advent of Artificial Intelligence (AI) and Machine Learning (ML) has piqued the interest of the aerospace industry since it has the potential to improve performance and alleviate the burden on personnel. The big question in the industry right now is how and where to introduce this technology while assuring the safety and reliability of its implementation. Guidelines drafted by the European Aviation Safety Agency (EASA) for the development of AI showed that maintenance and training were the most accessible points of entry for this technology as it did not have the same stringent requirements that a flying system would have. This paper proposes a research methodology which allows for the cost-effective development of ultrasonic data for the training and testing of data-driven tools. This was partly achieved by using a novel eFlaw technique which has been implemented for the first time in composite structures. The method allows for significant augmentation and generalisation of datasets, resulting in a model with the ability to detect features potentially smaller than one-quarter of a wavelength. This improved performance paves the way for more sensitive low-frequency ultrasonic inspection in thick composites. To evaluate these models, various evaluation techniques were compared and showed that Receiver operator curves and confusion matrix-derived metrics provided comparable results. Explainable methods found that the GradCam and the inspection of feature maps showed the most interpretable results on the features that were being identified. Using the feature maps it was possible to generate a new type of C-scan, called an F-scan (Feature-scan) which provides an inspector with a view of the C-scan from the perspective of a feature map from the model providing an interpretable view of the model's classifications. In addition to these positive results, this thesis provides readers with a cost-effective methodology for developing data-driven tools for maintenance applications within the aerospace industry.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

ASPP Atrous Spatial Pyramid Pooling

ASTM American Society for Testing and Materials

AUC Area Under Curve

AWGN Additive White Gaussian Noise

CAD Computer Aided Design

CEN European Committee of Standards

CNC Computerized Numerical Control

CNN Convolution Neural Network

CSV Comma-Seperated Values

DARPA Defense Advanced Research Projects Agency

DASML Delft Aerospace Structures and Materials laboratory

DCNN Deep Convolution Neural Network

DEMO Dienst Elektronische en Mechanische Ontwikkeling

EASA European Aviation Safety Agency

EU European Union

FBH Flat-Bottom Hole

FEM Finite Element Model

GAN Generative Adversarial Network

GFRP Glass Fibre Reinforced Polymer

GLM Generalised Linear Model

GPU Graphics Processing Unit

GRU Gated Recurrent Unit

LIME Local Interpretable Model-agnostic Explanation

LRP Layer-Wise Relevance Propagation

LSTM Long Short-term Memory

MAPOD Model Assisted Probably of Detection

MCC Matthews correlation coefficient

ML Machine Learning

MSE Mean Square Error

NDT Non-Destructive Testing

NN Neural Network

OLS Ordinary Least Square

PAUT Phased Array Ultrasonic testing

PoD Probability of Detection

RNN Recurrent Neural Network

ROC Receiver Operating Characteristics

SNR Signal-to-Noise Ratio

SSI Structural Similarity Index

SVM Support Vector Machine

VGG Visual Geometry Group

xAI Explainable Artificial Intelligence

1 Introduction

In the last decade, the aviation industry has taken a particular interest in Artificial Intelligence (AI). AI offers the potential to improve the automation, accuracy and precision of various aviation tools and systems. These benefits have led many of the key players within the industry to recognise these benefits and to work together with the authorities to safely introduce these technologies. This collaboration has culminated in the development of a roadmap developed by the European Aviation Safety Agency (EASA) [1]. The roadmap addresses the big-picture challenges and opportunities of AI in aviation, with the core focus being on the trustworthiness of AI. The roadmap also addresses how to prepare for the certification of these advanced systems, as well as identifying what processes, methods and standards still require development before being mature enough for industry. Under phase I of the roadmap, EASA has already released two concept papers for the ”first usable guidance for Level 1 & 2 machine learning applications” which focus on guiding its user when introducing AI technologies into safety-related systems within aviation [2]. It acknowledges that these guidelines are still in their infancy and that they will be enriched with more advanced techniques as time progresses, with the final goal being that these documents serve as a framework which will be consolidated into formal regulatory documentation in the second phase of the roadmap.

The most effective way to develop these concept papers requires the collaboration of the entire industry to verify, validate and advise on the current processes, but also to propose solutions for underdeveloped areas within the development process. Through literature, it was found that there is a gap between the research of the topics and the implementation of these technologies into the industry, with little to no mention of the guidelines proposed by EASA. Part of the issue was identified as there being a lack of an accepted framework to perform research which allows for the effective communication and comparison of research [3, 4]. This thesis develops an AI tool as a medium through which the current state of literature and the EASA guidelines are evaluated. The aim of this process is to propose a framework which can help researchers develop tools in a cost- and time-efficient manner so that more of their efforts can be focused on less mature areas within this topic. This thesis also identifies areas which require development and helps direct future research to close these gaps and advance the qualification of AI more directly.

To research this topic a suitable use case is selected to investigate and demonstrate the proposed framework. The guidelines identify that a data-driven tool designed for maintenance or training will be the most accessible point of entry for a safety-related system since it only requires a continuous safety assessment during operation (provided it was designed for an already certified platform) [2]. Additionally, literature has shown various promising experimental applications for data-driven methods in Non-Destructive Testing (NDT) for aerospace applications [5]. One such application is ultrasonic inspection which is one of the more popular methods due to its ability to identify features within a bulk material [5]. The challenge that inspectors face when using ultrasonic inspection is amplified when inspecting composite materials. Composite materials present a large amount of attenuation in ultrasonic signals and often introduce a trade-off in the probe’s frequency since a low-frequency signal will penetrate deeper into the structure, but be less capable of detecting smaller features and vice versa. Similarly, composite structures can introduce complex reflections within the bulk of the material which make it difficult to interpret specific signals. Given these facts, it becomes evident that a data-driven tool capable of interpreting these complex signals to detect relevant features could prove beneficial. This is made further relevant when considering that over the last 100 years, the aviation industry has steadily adopted composite structures into larger fractions of the total structural weight of aircraft, where at present there are various aircraft whose structural weight consists of over 50% composite materials [6].

Given these developments and challenges, it was chosen to develop a Machine Learning (ML) tool for a low frequency Phased Array Ultrasonic testing (PAUT) wheel probe to aid inspectors in the inspection of thick composites. The scope of this project was to develop a tool end-to-end with reference to both literature and the guidelines and to propose a research framework which includes mature processes and identifies areas which require further development. The scope of this research was restricted to the development of an AI tool, this meant that many of the administrative and project-oriented tasks within the guidelines would be considered out of scope unless directly impacting the development of the model. Within the scope, there were three topics which would be investigated in greater detail, the first was investigating a data augmentation/synthesis technique termed the 'eFlaw' [7] and investigating its capabilities and limitations in data set creation and training. The second was to implement explainable methods to the model to investigate whether it was possible to determine which features, both globally and locally, the model was focusing on to make its decisions. The final topic was to investigate how this tool could be introduced into an industrial setting, and how it would impact the inspection process.

This thesis will be structured to first discuss the relevant literature in chapter 2 which will cover the basic principles of ultrasonic inspection, an overview of the conventional qualification methods for NDT as well as AI for aerospace, experimental damage analogues for experimental samples, model generalisation and overfitting, an overview of machine learning models and the applications in ultrasonic inspection, the performance metrics used to assess the performance of these models, and finally how explainability can be inferred from these models. In chapter 3 the key research points will be discussed, followed by the research questions and an overview of the framework applied in this thesis. In chapter 4 the complete experimental setup and methodology applied will be discussed, this will include the concept of operation for the tool, inspection setup, specimens, data acquisition and preparation, the eFlaw implementation, design of the machine learning models, model training and evaluation, and finally the development of the model into a useable tool through the newly proposed F-scan. This thesis will then discuss the results of the experimental phase in chapter 5 with respect to the three topics of focus and the main research question. With the thesis concluding in chapter 6 and chapter 7 where the conclusions drawn by the thesis and suggestions for future work will be discussed respectively.

2

Literature review

This chapter will serve as an overview of the relevant literature covered for this thesis. It includes new sources as well as sources discussed in a separate more in-depth literature study [3]. To provide a proper overview across the entire scope of this thesis a broad range of topics must be studied. The first is an introduction to Phased Array Ultrasonic testing (PAUT) and how it works, after that this section aims to understand the qualification of ultrasonic inspection within aerospace, as well as the qualification of Artificial Intelligence (AI) within aerospace. Thereafter, the types of damage analogues which are used in research to test these methods will be discussed. This chapter will then discuss machine learning and how data is augmented, and what types of models have already been used in the evaluation of ultrasonic data. The final two sections of this chapter will discuss how the performance of these models is evaluated, as well as the different methods used to extract explanations from the model.

2.1. Ultrasonic inspection

This section will discuss the basic principles of ultrasonic inspection, more specifically, for PAUT. It will be broken down into two subsections, the first discussing the basic principles of PAUT and how it is used. The second subsection will discuss how ultrasonic data is captured and how it is represented.

2.1.1. Phased Array Ultrasonic Testing

The ultrasonic transducer is the building block upon which all ultrasonic inspection is built. Typically constructed from a homogeneous piezoelectric material, these transducers possess the capability to convert electrical energy into mechanical energy, and vice versa [8]. In the past, single-element transducers were employed for these inspection tasks, however, they proved to be time-consuming for technicians who often needed multiple measurements from various angles and locations.

The efficiency of echo transmittance in a transducer can be influenced by the incidence angle requiring special wedges for each inspection setup [9], additionally, technicians also employ various angles to navigate internal geometries, necessitating the use of a wider range of wedges to achieve different internal angles. Figure 2.1 illustrates an example of such a wedge, employed to produce more efficient shear waves within the bulk material. These wedges are specifically designed for standardized angles (e.g., 30, 45, and 70 degrees) for specific materials [9]. However, if a technician requires a wedge for an exotic material or to achieve a non-standard internal angle, a custom wedge must be ordered, leading to increased operating costs and complexity.

PAUT presents a viable solution to these challenges. By employing an array of narrower elements that are arranged in close proximity, PAUT offers enhanced flexibility for inspectors. Each element within the array is linked to a driving/receiving circuit, enabling independent adjustment of delay and application of amplitude weights to the transmitted or received signals. Figure 2.2 visually depicts these weights and delays. Through the manipulation of these parameters, a wide range of beam modifications and corrections can be implemented to cater to the specific inspection requirements at hand. The capabilities of PAUT are demonstrated in Figure 2.3.

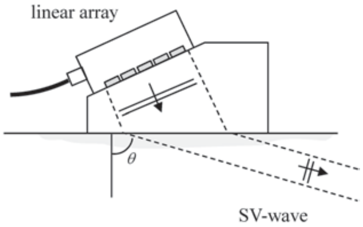


Figure 2.1: Illustration of a low-speed wedge used to generate shear waves internally [8]

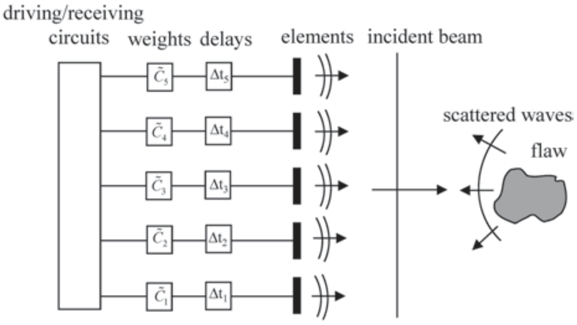


Figure 2.2: Individual elements that make up the array of transducers in PAUT, where each element has the ability to have a delay applied to the signal (Δt_i), or a weight applied to adjust the amplitude of the signal ($\overline{C}_i$) [8].

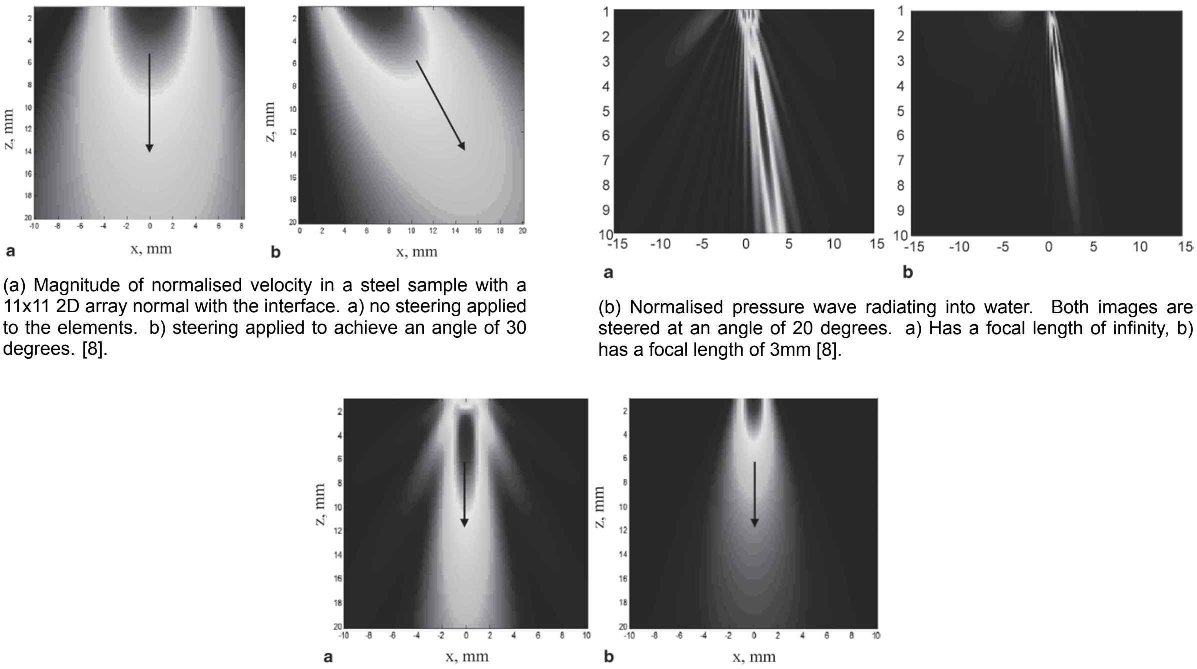
The delay functionality allows for the beam to be both steered and focused by ”triggering” each element individually at different times in a specific sequence. Figure 2.3a illustrates how a beam is steered using the delays of each element. Figure 2.3b illustrates the ability to use the delays to focus a beam to different depths within the material allowing for more energy to be directed to specific areas of interest. One final ability of PAUT is to remove the presence of side lobes in the beams (as can be seen in image a. of Figure 2.3c). Side lobes are an undesirable artefact of the transducer and can cause problems with the quality of the measurements. Specifically, side lobes lower the signal-to-noise ratio which results in more difficult-to-interpret results and poorer image quality in scans [10]. By using the weights and delays, it is possible to apply apodisation to the PAUT and remove the side lobes. The downside to apodisation, however, is that the main beam broadens, thereby decreasing the lateral resolution [10][8]. The ability to use PAUT to steer, focus and correct a beam provides the possibility to gather a lot of data and focus more precisely on specific features of interest providing a vastly superior inspection method to single-element ultrasonic inspection.

Within aerospace, PAUT sees a great deal of use on various structures. It can be used to assess the quality of welds and bulk materials. Such examples include landing gears, fuselage skins, wing boxes, and exotic materials such as GLARE [11]. Additionally, the anisotropic properties of composite have been known to cause a great deal of attenuation and scatter when using single-element probes, the ability for PAUT to be able to steer and focus the beam has made PAUT the dominant choice for composite inspection [12].

2.1.2. Ultrasonic data

When these principles are applied, the ultrasonic waves are received and the ultrasonic data can be obtained in its most fundamental form as an A-scan, which presents itself as a one-dimensional representation of time against the signal amplitude [11]. This single signal alone can be utilized to infer the condition of materials [8]. The two primary methods to capture an A-scan require either employing two probes, one for transmission and the other for reception, positioned on opposite sides of the sample in a ”Pitch-Catch” configuration. The other method, known as ”Pulse-echo,” employs probes placed on the same side of the sample, where either the same element or different elements receive the reflected waves. The Pitch-Catch method is often favoured due to its higher signal-to-noise ratio (SNR), but it necessitates access to both sides of the sample, which can be challenging for many structures in practical scenarios [13]. Consequently, Pulse-echo is the prevailing method used for inspection.

When performing PAUT, a large array of pulse-echo A-scans can be generated to visualize damage



(a) Magnitude of normalised velocity in a steel sample with a 11x11 2D array normal with the interface. a) no steering applied to the elements. b) steering applied to achieve an angle of 30 degrees. [8].

(b) Normalised pressure wave radiating into water. Both images are steered at an angle of 20 degrees. a) Has a focal length of infinity, b) has a focal length of 3mm [8].

(c) Normalised velocity in steel with a normal incidence. a) No steering or focusing applied, shows clear side lobes. b) Hamming window applied, resulting in sidelobe removal (notice how the main beam is slightly wider than in sub-figure a.) [8].

Figure 2.3: Images of the different capabilities that PAUT offers.

in new ways, thanks to the additional dimensions involved [14]. Figure 2.4 illustrates two achievable representations. The B-scan provides a view of the sample's thickness, enabling the measurement of defect depths and sizes, while the C-scan indicates the surface position but lacks depth information. To facilitate the rapid production of both views, a PAUT method was developed using a roller wedge. Coupling this roller with an encoder allows for a large series of scans to be made over an area. Olympus offers an example of such a device [15].

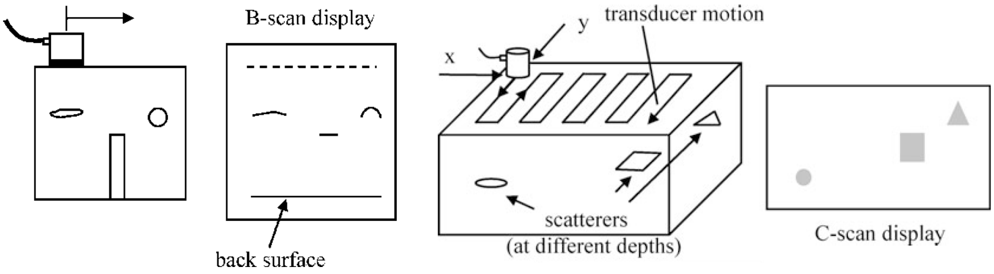


Figure 2.4: Illustration of a B-scan and C-scan [14]. The left images show how the scanning achieves the B-scan image, and the image on the right shows how a C-scan is achieved (requiring full coverage of the sample)

There are additional approaches available for enhancing the quality of captured data, aiming to generate a more distinct image that facilitates the detection of damage. In order to optimize data processing techniques, a substantial amount of data is desirable. The Full Matrix Capture (FMC) method stands out as the most data-intensive capture method. This technique entails transmitting a signal using a single element, while simultaneously receiving the resulting reflections with all elements. Although this process is time-consuming and leads to a slower inspection, it generates N^2 A-scans, where N represents the number of elements [8].

2.2. NDT Qualification

Ultrasonic testing presents a non-trivial inspection technique characterized by its complexity and difficulty. Consequently, any inspection process involving this technology necessitates rigorous verification and validation to attain a satisfactory level of certainty and confidence in the test results. This holds particular significance in industries such as biomedical, nuclear, and aerospace, where stringent qualification processes are essential [16]. In this section, we will explore the typical qualification procedures employed when certifying ultrasonic processes, with a specific focus on aerospace applications. Once these conventions have been discussed, then the leading qualification challenge of certifying AI for use in the aerospace industry can be discussed.

2.2.1. Personel qualification

The qualification process for ultrasonic testing begins with the qualification of personnel, ensuring individuals are capable of understanding and interpreting test results. The European Committee of Standards (CEN) provides standards for personnel qualification in Non-Destructive Testing (NDT) and aerospace [17], such as EN 4179 [18] which is an adapted version of EN473 [19] for use within the aerospace industry. This standard defines four levels of certification for aerospace personnel, ranging from limited certification (Level 1-Limited) with specific test capabilities to the highest level (Level 3) with comprehensive technical and administrative responsibilities [18].

Level 1-Limited personnel can conduct specific NDT tests under guidance, while Level 1 personnel have more responsibilities, including acceptance/rejection testing at the product level. Level 2 personnel must understand technique/method limitations, and aircraft/vehicle maintenance, and develop work instructions which require final approval. Level 3, the highest level, requires a broad knowledge of various NDT methods, manufacturing techniques, and administrative skills.

These certification levels offer context for introducing machine learning in the NDT process. Level 3 personnel may be more risk-averse, ensuring proper operation, while Level 1 personnel can provide data to improve algorithms. Moreover, these levels can be used to assess the feasibility of qualifying AI as an inspector, although ethical and philosophical concerns must be considered.

2.2.2. Qualifying an NDT process

In addition to qualifying personnel, the qualification of technology and inspection processes is equally important. In conventional ultrasonic methods, the leading qualification processes involve the Probability of Detection (PoD) of the process, of which two methods exist. The first and most commonly used method is the hit/miss method, which provides a binary value indicating the presence or absence of a discontinuity within the specimen. On the other hand, the â-vs-a model offers more information on the system's performance by predicting the measured damage size versus the true damage size using signal strength [20, 21, 22]. The hit/miss method's popularity is attributed to its simplicity, while the â-vs-a model provides valuable insights into the system's capabilities during inspection. Both forms of qualification play a critical role in ensuring reliable and accurate ultrasonic testing results.

The â-vs-a method, standardized by ASTM in ”ASTM: E3023-21 Standard Practice for Probability of Detection Analysis for â Versus a Data” [23], involves a regression of the measured signal versus the true damage size. The data is often transformed into a combination of [â or log(â)] vs [a or log(a)] in the pursuit of creating a linear regression. Censoring of signal data, where readings may be left-censored due to small cracks or right-censored due to large cracks falling outside of the instrument sensitivity, further complicates the regression problem. Likelihood functions are used to redistribute censored signals and account for lost information [20, 23, 24, 25].

Once a linear model is determined, the Probability of Detection (PoD) curve is created using a specific equation in United States Department of Defence [25]. This curve enables the determination of the minimum target size required for a desired probability and confidence interval. Before applying the â-vs-a method, certain guidelines must be followed, impacting its validity. The model must correspond with the measured data, exhibit uniform variance on either side of the â-vs-a line, have uncorrelated observations, and feature normally distributed errors [25, 21]. These considerations are crucial to ensuring accurate results when applying the â-vs-a method for ultrasonic testing qualification.

The Hit/miss method is simpler in terms of data acquisition as it involves binary values (0 or 1) indicating the presence or absence of a discontinuity. This method has been standardized by ASTM under ASTM: E2862 [26]. The data collected using the Hit/miss method is clustered into two categories.

Unlike the â-vs-a method, the Hit/miss method cannot use the ordinary least squares (OLS) regression model due to the binary nature of the data and the lack of constant variance. Instead, the Hit/miss method requires a generalized linear model (GLM) with an appropriate link function [21]. The most commonly used link function is the logit link function, although other options such as probit, cloglog, and loglog are available [25]. The logit link function is often recommended unless there are strong physical or statistical reasons for using an alternative link function.

Both the â-vs-a method and the Hit/miss method have their own advantages and trade-offs. The â-vs-a method requires a minimum of 40 samples, while the Hit/miss method requires over 60 samples to ensure reliable results [25]. Using fewer samples can lead to less certain results and wider confidence bounds, with the Hit/miss method exhibiting erratic behaviours such as negative slopes, convergence problems, and wider confidence bounds [25]. These sample size requirements can pose a significant challenge when qualifying a process, as a large number of samples is necessary to achieve meaningful results and can be time-consuming and costly.

2.2.3. Qualifying AI for NDT applications within aerospace

As discussed in the preceding sections, the aerospace community is well aware of the challenges and concerns related to the adoption of AI in the industry. To address these issues and introduce AI in a safe and effective manner, various organizations, including the European Aviation Safety Agency (EASA), have developed roadmaps and guidelines.

The first major problem highlighted by EASA and the European Union (EU) is the ethical questions surrounding AI implementation. To create a trustworthy system and foster competitive development within European companies, the EU emphasizes the importance of a ”European ethical approach to AI” [1]. The EU commission has established ethical guidelines for AI, known as the seven pillars, which include accountability, technical robustness and safety, oversight, privacy and data governance, non-discrimination and fairness, transparency, and societal and environmental well-being. These pillars are integral to the trustworthiness of AI systems and formed the foundation for EASA’s roadmap [1].

In February of 2023 EASA released issue 2 of its concept paper on AI Guidelines, where the major addition being the inclusion of guidelines for level 2 AI [2]. Where each of the levels is best described in Table 2.1

Table 2.1: Table from the guidelines describing each of the various levels [2].

Level 1 AI : assistance to human	**Level 2 AI : human/machine teaming**	**Level 3 AI : more autonomous machine**
Level 1A: Human augmentation	Level 2A: Human and AI-based system cooperation	Level 3A: The AI-based system performs decisions and actions, overridable by the human
Level 1B: Human cognitive assistance in decision and action selection	Level 2B: Human and AI-based system collaboration	Level 3B: The AI-based system performs non-overridable decisions and actions

This document outlines the entire end-to-end process for qualifying AI for the aerospace sector, among that process are three main areas which require the most attention, those being:

- Learning Assurance
- AI explainability
- AI Safety Risk Mitigation

The guidelines restrict all AI applications to supervised methods only. Section C.2.1.2 of the guidelines strengthens the point to develop AI for maintenance and training-related topics since there is no ”initial” safety assessment required, with only the continuous safety assessment during operation having to be conducted. This offers a significant amount of flexibility and reduced costs when developing systems for these fields. What this would allow is for the focus of research to be on the first two areas previously mentioned (Learning Assurance, and AI explainability).

2.3. Damage analogues

It is important to understand the current methods employed by researchers to introduce damage into composite structures as well as the limitations and advantages of each method. This section aims to discuss the various different methods and analogues which are used to represent real-world damages which occur during the manufacturing or operation of the structure.

There are two directions for introducing damage into a specimen. Those are synthetic damage and real damage, each with their own advantages and disadvantages. The real damage is induced through a combination of stress, strain, fatigue cycles, manufacturing defects and environmental effects and results in a wide array of damage types ranging from matrix cracking, delaminations, moisture ingress, corrosion, dents and more [27]. Many of these damages are reproducible in laboratory conditions either through impacts setups like a drop or projectile test [28] to induce various types of cracking and delamination, placing samples into fatigue testing machines to progressively introduce damage [29], using tensile and compression testing machines to peel or damage specimen [27]. These methods introduce the closest representation of real-world damage in a controlled environment. A major disadvantage of this method is that the damage often contains some ambiguity when measuring the damage and defining what the true damage size is compared to the measured damage size. This ambiguity can be shown in

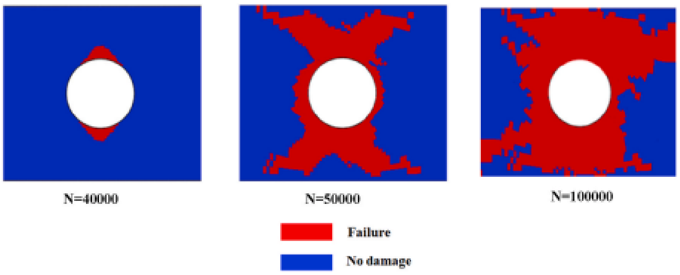


Figure 2.5: C-scan of fatigue damaged notch. Showing the damage progression in the sample over time [29].

The synthetic damage methods attempt to address this problem by providing methods which yield reproducible damage with far less ambiguity on the size and location of the damage, however, they compromise on how representative the damage is to real-world damage. Blain et al. [30] summarize the three main methods to introduce damage in a controlled manner, those being: Flat-Bottom Hole (FBH)s, interlayer insertions, and fibre pullout. This is best illustrated in Figure 2.6.

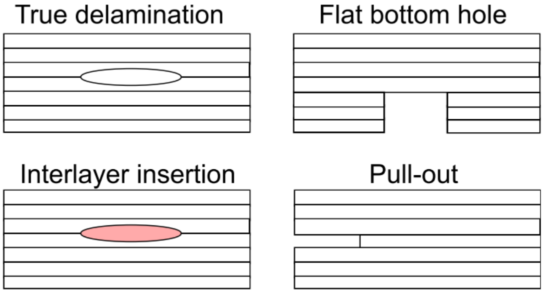


Figure 2.6: Illustration of the different damage analogues compared to conventional delamination [30]

Each of these analogues comes with its own advantages and disadvantages. The FBH, for instance, is a popular analogue as it is the most consistent analogue to reproduce and this makes it an ideal candidate for calibration blocks [31]. Mix [31] discusses the limitations of FBHs and how they make it difficult if not impossible to directly compare the calibration blocks to real damage such as gas porosity, nonmetallic inclusions, etc. This is due to the concave nature of a lot of these natural damages which reflect the beams differently, and the straight-edged nature of the FBH.

Interlayer insertions such as air-filled Teflon layers also produce relatively reproducible damage analogues and must be incorporated during the layup of the composite and cured/consolidated together.

Teflon has been a particularly popular method for shearography and thermography [30, 32]. However, Teflon introduces a new material into the composite which can affect the attenuation and refraction of the beam among other things.

For pull-out, a metal insert is placed at the edge of the composite. After curing the metal insert is removed leaving a void in its place [30]. This method addresses the issue with Teflon but is restricted to the edges of samples (limiting the effective area of the method).

2.4. Generalisation and adressing overfitting

Researchers have encountered a recurring obstacle in the field when utilizing machine learning techniques for nondestructive testing (NDT). This challenge revolves around collecting a sufficient amount of data that is not only used for training the model but also facilitates verification and validation of the model, all while trying to minimize the overfitting and improve the generalisation of the model [33, 34]. The difficulty in gathering sufficient data can primarily be attributed to either cost constraints or the inherent manufacturing complexities involved in introducing the necessary damage within specific regions [7]. The size of the dataset holds great importance as it directly impacts various aspects related to the model. To overcome these limitations, data augmentation techniques play a crucial role in addressing the data scarcity problem for NDT machine learning topics. This section will discuss the impact of overfitting, data augmentation techniques to prevent overfitting, and other methods in the literature that address overfitting.

2.4.1. Overfitting

Overfitting is one of the major impacts that a dataset can have on a Machine Learning (ML) model if the dataset is not diverse or sizeable enough. This issue is frequently encountered in deep learning and has been extensively discussed by Bishop [35]. In essence, ML models can be viewed as highly intricate regression models, and are illustrated best in Figure 2.7 through a straightforward polynomial regression. This illustration effectively demonstrates that an insufficient amount of data can result in a regression model that fails to accurately represent the true underlying function by defining a high-order polynomial which attempts to intersect every point. The figure highlights that the problem can be mitigated by increasing the volume of data. However, it should be noted that generating additional data for Non-Destructive Testing (NDT) is not a straightforward task due to the associated complexity and cost. Consequently, researchers have sought alternative approaches to combat overfitting, leading to the development and utilization of a diverse set of innovative methods. Many of these strategies fall within the realm of data augmentation, which is defined by James et al. [36] as the practice of replicating training data multiple times by randomly distorting each replicate in a natural manner. Other methods look to the model architecture itself and attempt to address the overfitting problem through the design of the model.

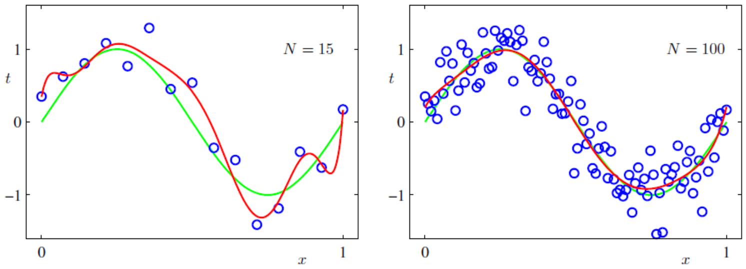


Figure 2.7: Illustration of a polynomial function being overfit to the data, the figure on the right shows that a higher dataset is able to reduce the effects of over-fitting and adhere better to the true function (green line) [35]

2.4.2. Data augmentation methods for ultrasonic data

Data augmentation methods for ultrasonic data depend heavily on what ultrasonic data is being used. Regarding image data, such as C-scans or B-scans, simple rotational and translational modifications can be applied to manipulate the images. Additionally, distortions can be introduced to further expand the dataset [34, 36, 37]. Medak et al. [38] trained their model on a dataset consisting of B-scan images, which were augmented using various transformations, including horizontal flipping, random cropping, translation, contrast adjustment, and colour enhancement. Notably, rotational transformations were not utilized by Medak et al. [38] since B-scans are consistently presented horizontally. Furthermore, Guo et al. [37] and Du et al. [39] employed histogram equalization to enhance the dynamic range of their images, aiming to improve their model's learning capabilities.

When dealing with signal data, such as A-scan data, augmentation methods require a higher level of creativity due to the limited dimensions of the data. Cantero-Chinchilla et al. [4] and Munir et al. [40] employ timeshifting of the A-scan as a means of data augmentation. This augmentation technique simulates the displacement of defects closer to or further away from the transducer. By employing this small adjustment, they were able to increase their dataset by a factor of 5. The data is further augmented by introducing Additive White Gaussian Noise (AWGN) to simulate the temperature rise in resistors within the signal processing devices, which generates a voltage that can be represented by Gaussian noise [40]. To achieve this they augment the Signal-to-Noise Ratio (SNR) of the A-scans, transforming them from low-noise signals to various high-noise signals. Augmenting the data with varying levels of noise benefits the model by enhancing its performance with signals having higher SNRs, thus improving the model's robustness. This technique of applying Gaussian noise was also utilized by Latête et al. [41] where the inclusion of noise resulted in an additional five-fold increase in the dataset.

A more creative approach was taken by Virkkunen et al. [33], who used a concept they referred to as eFlaws. Virkkunen et al. [7] developed this technique to address two problems, the first was to provide a solution to make training NDT specialists easier by providing a solution where inspectors could have a series of real-world damages introduced into their scans at any moment minimizing the number of physical samples required. The second reason was to provide a better solution for Model Assisted Probably of Detection (MAPOD), with some of the problems being that simulated ultrasonic features were sometimes too idealised to be used as a real-world analogue, on top of this, machine learning models did not perform well on real-world data when trained on simulated data. The researchers as a result produced a method which could extract the damage features from a sample, and reintroduce it into the sample in another location. The principle is described in the following flow chart Figure 2.8).

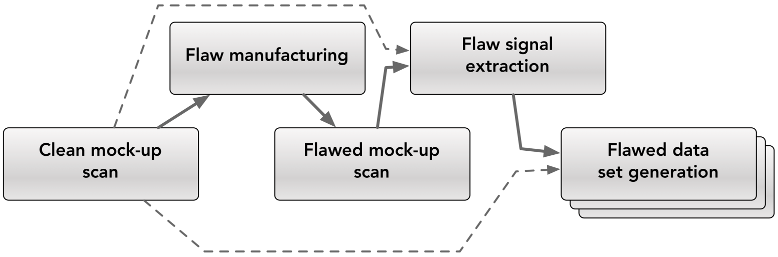


Figure 2.8: Flowchart illustrating the virtual flaw principle [33]

In this process, an undamaged baseline scan of the sample is required. After this, damage can be introduced into the samples via various different methods. The damage is then scanned and the two measured signals can be subtracted from one another to result in the flaw being extracted. With this flaw extracted it is now hypothetically possible to reintroduce the flaw back into the sample at a chosen location. This process, however, is not as trivial as described. It requires a great deal of accuracy in the scans to ensure proper alignment. Secondly, there will always be some noise present in the system, resulting in imperfect subtractions which may require corrections. In the case of Virkkunen et al. [7], they used a simple histogram filter which would zero values while applying an increasing factor related to the intensity of the signal. Additionally, it was pointed out that if the Signal-to-Noise Ratio (SNR) was too low the extracted damage feature would reside in the residual noise and be even more difficult to introduce back into the sample. The impact of these limitations might make it more challenging to create data samples of smaller and more desirable damage for PoD curves.

Koskinen et al. [42] extends the existing eFlaw research by examining the impact of different flaw data on machine learning. They employ an enhanced eFlaw technique that incorporates the ability to modify extracted flaws before reinsertion, with the scaled-down size limited to a maximum reduction of 40% compared to the original data. The study involves six scanned flaws and six simulated flaws (simulated in CIVA2019). By applying the eFlaw technology, the researchers successfully generate datasets comprising up to 7000 samples. However, this raises concerns about the diversity of damage, as the validation set is also based on these eFlaws, thereby questioning the validity of the derived Probability of Detection (PoD) metrics.

Meister et al. [43] conducted a comprehensive literature review to explore the potential application of synthesizing augmented data for enhancing datasets used in fibre layup inspection processes. Among their findings, they highlighted the effectiveness of Generative Adversarial Network (GAN) as a data-driven solution. GANs can be employed either to evaluate the quality of conventional data augmentation methods and identify the most impactful augmentation technique for improving model performance or to generate entirely new data. Creswell et al. [44] define GANs as a burgeoning technique applicable to both semi-supervised and unsupervised learning scenarios. Fundamentally, GANs consist of two models: a generator and a discriminator. The generator's objective is to deceive the discriminator, which is trained to differentiate between synthetic and real data [44]. GANs can serve various purposes, with the primary and widely pursued objective being training the generator to produce increasingly realistic data that is indistinguishable from real data. This capability can prove highly valuable for industries struggling to obtain sufficient training data. Additionally, GANs can facilitate the creation of effective discriminators, which can be utilized in areas such as cybersecurity to identify falsified data. In the aerospace domain, GANs can be employed for generating new training data or, as highlighted by Meister et al. [43], for verifying the integrity of existing training data. By flagging mislabeled samples, the discriminator helps reduce model errors and enhances overall performance.

2.4.3. Model solutions to combat under/overfitting

Within Neural Network (NN)s and Convolution Neural Network (CNN)s there exist a host of new methods to improve the training performance and generalisation of machine learning models. The first of these methods is dropout regularization, first introduced by Hinton et al. [45]. This method directly addresses overfitting by preventing co-adaptations on the training data. Co-adaptation is defined by Hahn and Choi [46] as the process by which two or more nodes behave as if they are a single node or group. To minimize or prevent co-adaptation the method randomly omits neurons in the network during training and has been described as a method of performing ”model averaging with neural networks”. Hinton et al. [45]'s research shows that dropout can be used during finetuning to overcome overfitting and allow the model to converge during training.

Ioffe and Szegedy [47] present another solution which addresses both generalisation and training performance. One of the factors which impacts this is the internal covariate shift during training which they describe as being the ”change in the distribution of network activations due to the change in network parameters during training” [47]. What they observed is that in certain cases this covariate shift is high, and results in a longer convergence during training but also poorer generalisation. The process in essence centers and normalizes the feature maps of the model and smooths the loss function of the model (due to the reduced covariate shift).

These two methods are among some of the most impactful model-agnostic methods which can be implemented. Of course, there are other solutions and factors such as adaptive learning rates [48], activation functions [49], and general hyperparameter tuning [50]. All of these factors should be considered during the design and training of the model.

2.4.4. dataset sizes

Previously it was discussed how to effectively use the data that is available via augmentation and how model architecture can be adjusted to improve generalization, however, these points did not cover the typical size a dataset should be for ultrasonic applications. In their study, Hestness et al. [51] effectively illustrate the importance of dataset size in the learning process, as depicted in Figure 2.9. The figure demonstrates a power-law relationship between the dataset size and the generalization error. Increasing the dataset size yields the most significant benefits within the power-law region before reaching a point of irreducible error. This irreducible error arises from imperfect generalization, often caused by mislabeled data.

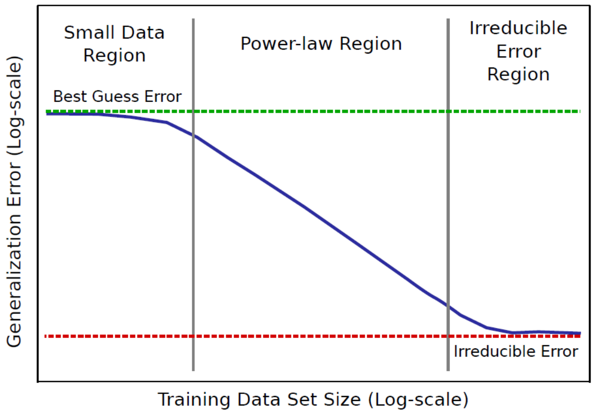


Figure 2.9: Power-law of learning curves [51]

Cho et al. [52] conducted research in the biomedical industry, which shares rigorous standards akin to those in the aerospace sector, to explore dataset sizes in ML applications for NDT. Specifically, they focused on examining the influence of dataset size on the effectiveness of deep learning systems for medical image analysis using CNNs. The results, depicted in Figure 2.10, demonstrated a rapid effect of the power-law relationship. With a dataset of approximately 200 samples, the system achieved an accuracy of about 95%. Subsequently, at 1000 samples, the accuracy improved to 97.25%, and at 4092 samples, it soared to an impressive 99.5% accuracy.

To model the learning curve, Cho et al. [52] employed Equation 2.1.

$$\mathbf{y} = f(\mathbf{x}; \mathbf{b}) = 100 + b_1 \cdot \mathbf{x}^{b_2} \tag{2.1}$$

Where x and y are the vectors which represent the dataset size and the resulting accuracy. b_1 and b_2 in this instance represent the learning rate and the decay rate respectively Figueroa et al. [53]. This method can provide an indication of the order of magnitude required for certain degrees of accuracy. This curve allowed Cho et al. [52] to establish an upper limit on the accuracy, representing the irreducible error. In their study, this irreducible error was assumed to be 0%, as the limit tended towards infinity.

It is important to highlight that Cho et al. [52]'s investigation in the biomedical industry, although not directly related to NDT, provides valuable insights into the relationship between dataset size and the efficacy of deep learning systems. Such research findings can be beneficial for developing effective ML applications in the field of non-destructive testing as well.

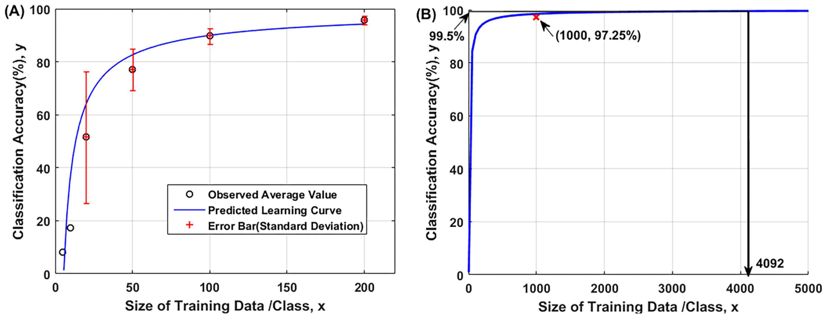


Figure 2.10: A: Predicted learning curve for different dataset sizes; B: Results for larger datasets [52]

In Table 2.2, an overview of dataset sizes used by researchers investigating machine learning (ML) in ultrasonic applications is presented. The smaller datasets were employed for data-driven methods with lower complexities, such as Support Vector Machine (SVM) or Hidden Markov models. These statistical methods aim to model a system and are known to require significantly smaller datasets compared to the more intricate Convolutional Neural Networks (CNN), which demand larger datasets [43].

At the higher end of the dataset sizes, some studies approached or even exceeded the 100,000-sample mark. Achieving such volumes required extensive data augmentation. For instance, Guo et al.

[37] augmented their dataset from 1400 recorded samples to a total of 99,800 samples. Similarly, Siljama et al. [54] utilized novel eFlaw technology, conventional augmentation techniques, and 16 different physical flaws within welds. By leveraging undamaged welds as a canvas, they generated an extensive array of samples, reaching an astonishing 500,000 samples. In the mid-range of dataset sizes, the adoption of CNN yielded favourable performance, with accuracy levels reaching as high as 99.2% in certain cases [40]. However, it is worth noting that the performance was significantly influenced by the type of defect and Signal-to-Noise Ratio (SNR).

These findings collectively illustrate the wide range of dataset sizes utilized in ML applications for ultrasonic inspection in non-destructive testing, highlighting the varying complexities and the importance of dataset size in achieving desired performance levels.

Table 2.2: Compiled dataset sizes from different papers, illustrating the distribution of size and their associated models used [3].

Sample size	**Type of Learning**	**Source**
282	Support Vector Machine	Virupakshappa and Oruklu [55]
395	Hidden Markov model	Virupakshappa and Oruklu [56]
1078	CNN	Virupakshappa et al. [57]
1400	CNN	Garifulla et al. [58]
3600	CNN	Munir et al. [40]
4174	CNN	Medak et al. [38]
4545	CNN	Khan et al. [59]
7000	DCNN	Koskinen et al. [42]
20,000	DCNN	Virkkunen et al. [60]
80,000	SVM vs CNN	Melville et al. [61]
99’800	CNN	Guo et al. [37]
500,000	DCNN	Siljama et al. [54]

From the existing literature, there is a preliminary indication that dataset sizes on the order of 10^3 or greater are required to develop an acceptable model in NDT applications using machine learning. This observation aligns with the findings of Meister et al. [43], who conducted a literature study on dataset sizes in image-based machine learning applications for NDT.

However, it is crucial to emphasize that the dataset size requirement can vary significantly depending on the specific ML architecture employed and the desired performance criteria. More complex models, such as Convolutional Neural Networks (CNN), often demand larger datasets to effectively learn intricate patterns and features. The appropriate dataset size for an NDT application will be influenced by factors such as the complexity of the ML model, the complexity of the inspection task, the quality of the data, and the desired level of accuracy.

2.5. Machine learning

In this section, the current state-of-the-art in ML and NDT, particularly in the context of Ultrasonic testing will be explored. The section will encompass various aspects, including the algorithms utilized, dataset generation methods, input-output specifications for each approach, and the assessment of algorithm performance.

2.5.1. Deep Learning architectures

Machine learning has the ability to revolutionize the industry by enabling the development of highly adaptable automated solutions. Deep learning, a subset of machine learning, uses multi-layered architectures to extract complex features from data [4]. In a study of the cutting-edge advancements in deep learning applications for ultrasonic data, Cantero-Chinchilla et al. [4] was able to identify a large number of applications. These purposes include, but are not limited to:

- Classification
- Data Characteristics
- Regression
- Computer vision
- Object detection
- Prediction

- Image labelling
- Denoising
- Feature extraction
- Image segmentation

Within the literature, three different architectures showed the most promise, those being the CNN, Recurrent Neural Network (RNN) and Autoencoders.

Among the architectures explored in the research, the CNN emerged as one of the most widely used models. Its versatility extends to applications in computer vision, object detection, and classification [62]. Notably, CNNs excel in handling structured multidimensional input data, owing to their convolution layers that convolve kernels with the input data to perform diverse transformations [63]. The training of CNNs involves tuning the parameters of each kernel, allowing them to establish strong correlations with neighbouring pixels, making them particularly effective for image processing [35].

Another prominent architecture in the studies was the RNN. Designed for time-series data, RNNs are frequently utilized in applications like speech recognition, finance, and ultrasonic inspection, particularly for A-Scan data. They incorporate loops in their hidden layers to retain information over time [64]. However, RNNs suffer from limitations due to their limited memory, which can lead to the vanishing gradient problem when dealing with large gaps between relevant information [64]. To address this issue, the Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU) were developed, using special memory cells to identify long-term dependencies within the data [64].

The research also identified AutoEncoders, GANs, and Diffusion models as networks widely used in generative applications. AutoEncoders and GANs have proven effective for data synthesis, denoising, and SNR [65, 4]. These applications are particularly valuable in NDT, where obtaining large datasets can be challenging. GANs, first proposed by Goodfellow et al. [66], involve two networks competing to improve performance. Subsequently, diffusion models, proposed by Sohl-Dickstein et al. [67], utilize Non-equilibrium Thermodynamics for unsupervised learning. Notably, the Stable Diffusion model [68] and Dall-E model [69] have gained popularity for high-fidelity image generation using latent diffusion from text prompts.

Although various other architectures and combinations exist, the ones discussed above represent a significant portion of the methods employed in literature when exploring applications in NDT and ultrasonic detection. These diverse architectures offer promising avenues for advancing autonomous ultrasonic detection and non-destructive testing.

2.5.2. Machine Learning models for Ultrasonic testing in Literature

This subsection will discuss notable progress made in the direct applications of ML in ultrasonic inspection with a focus on which models were used.

In their study, Munir et al. [40] employed a relatively straightforward CNN architecture consisting of two convolution layers, a max pool, and three fully connected layers. Their objective was to examine the influence of different levels of Signal-to-Noise Ratio (SNR) on model performance. As expected, the model achieved a high average accuracy at high SNRs while unsurprisingly performing worse the lower the SNR became.

In their research, Virupakshappa et al. [57] explored the possibility of using wavelet packet decomposition as a pre-processing technique before feeding signals into a CNN. They aimed to evaluate two different models: a modified version of the LeNet architecture [70] and a 1D CNN adapted from Zhang et al. [71]. Where the LeNet architecture performed exceptionally well compared to their 1D CNN, with the small dataset being identified as a major limitation to the results.

In their study, Medak et al. [38] focused on utilizing a CNN to classify and highlight defects in ultrasonic B-scans. They explored several models, including YOLO [72], ResNet [73], and EfficientDet [74]. The researchers tested three different versions of each model, except for the YOLO model, which was specifically YOLOv3. After evaluation, their findings revealed that EfficientDet-D0 achieved the highest accuracy among all the models tested. This result highlights the potential of EfficientDet-D0 as a promising candidate for defect detection and classification within ultrasonic B-scans.

In their research, Khan et al. [59] aimed to automate the detection of delaminations in composite beams using deep learning methods. They induced a low-frequency chirp through a shaker to obtain a time series response representing the acceleration of the beam's tip. While this setup was not an ultrasonic test, the output response was similar to that of an A-scan. To improve training efficiency, Khan et al. [59] employed transfer learning, using four established models: AlexNet, GoogleNet, SqueezeNet, and VGG16. The results showed that both AlexNet and GoogleNet performed effectively in identifying damage, with AlexNet slightly outperforming the other models. Transfer learning allowed the researchers to leverage pre-trained models from the ImageNet Large Scale Visual Recognition Challenge, which contains images of various categories such as cats, dogs, trucks, and cars. Although not directly relevant to NDT damage detection, the results demonstrated promising outcomes, indicating that transfer learning remained a viable and effective approach for their application.

The studies conducted by Virkkunen et al. [60], Koskinen et al. [42], and Siljama et al. [54] all incorporated their developed eFlaw technology in combination with a modified machine-learning model based on the VGG16 architecture. These research papers demonstrated exceptional performance on their validation and test datasets. Particularly, Virkkunen et al. [60] showed such strong results that they had no misses, prompting them to artificially insert ”miss” data at a crack length of 0 in order to properly create PoD plots.

Criticism has been raised regarding the validity of these results. The concern lies in the fact that the range of damage sizes used (from 1mm to 9mm) was entirely comprised of augmented eFlaws sourced from only three initial fatigue flaws. The authors acknowledged this risk and attempted to label the data appropriately to identify potential overfitting of the model on virtual flaws. Nevertheless, a critical assessment of how well these results truly represent real-world scenarios appears to be lacking.

As a result, the conclusion drawn by these studies, suggesting that deep convolutional neural networks can achieve human-level performance, and the significant role eFlaws play in this process, may be overstated. More comprehensive and diverse testing on real-world datasets would be essential to substantiate such claims adequately.

In their research within the biomedical industry, Guo et al. [37] utilized deep learning to segment thyroid nodules in ultrasonic images. They employed a modified version of the DeepLab v3+ model, which excels at extracting features and performing precise image segmentation. The model produces high-contrast images that effectively highlight the distinctive features of thyroid nodules.

A noteworthy aspect of this CNN is the application of AutoEncoding at both the beginning and end of the model. This step compresses the input data and extracts relevant features, reducing them into a 3x3 low-dimensional subspace. Subsequently, these features are passed through an Atrous Spatial Pyramid Pooling (ASPP) module, which simultaneously applies multiple convolutions at varying rates. This ensures that features are captured at different scales, enabling a comprehensive representation of the image’s characteristics.

While this solution shows significant promise, the authors emphasised that it heavily relies on a supervised process with accurately labelled ground truth data. Ensuring low error rates in the model requires meticulous labelling. Further investigation and validation of real-world datasets are necessary to determine their effectiveness and potential applicability in various contexts.

This section highlights the complexity and diversity within the field of machine learning, particularly in the context of damage classification using CNNs. The review indicates that CNNs hold great potential for this application with a variety of different methods and approaches.

2.6. Performance metrics

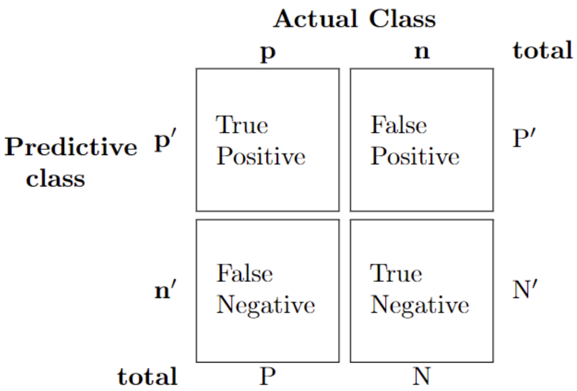
From the literature on machine learning applications in ultrasonics, an observation was made similar to that of Cantero-Chinchilla et al. [4], who highlighted the absence of a model-agnostic, universally-accepted method for quantifying model performance for ultrasonic inspection. This limitation hinders the accurate comparison of different models’ performance. This section aims to address this issue by investigating evaluation methods for assessing model performance, as it directly influences confidence in the model’s predictions. This section will explore general performance evaluation methods used to evaluate models.

A clear overview of various performance metrics was covered by Gong [75]. Their focus was on classification models and how to assess their performance.

2.6.1. Confusion Matrices

The first method discussed by Gong [75] is the confusion matrix, a commonly used evaluation tool for classification models. When evaluating the performance of a model in producing a Hit/Miss Probability of Detection (PoD), the classification becomes straightforward as the model categorizes data into either ”Hit” or ”Miss” classes, resulting in a 2x2 confusion matrix. This matrix is depicted in Figure 2.11, where the horizontal axis represents the actual class of the item, and the vertical axis represents the predicted class by the model. Utilizing this matrix provides a more comprehensive representation of the types of errors made by the model.

Drawing a parallel to hypothesis testing, classifying a False Negative by the model is equivalent to a Type I error, while classifying a False Positive is equivalent to a Type II error [36]. To simplify notation, acronyms are used to denote each outcome, with True Positive denoted as **TP**, True Negative as **TN**, False Positive as **FP**, and False Negative as **FN**. By analyzing these outcomes through the confusion matrix, researchers can gain valuable insights into the model's performance and its ability to correctly identify true positives and true negatives while minimizing false positives and false negatives.



		Actual Class		
		p	n	total
Predictive class	p′	True Positive	False Positive	P′
	n′	False Negative	True Negative	N′
	total	P	N	

Figure 2.11: Illustration of confusion matrix for a 2x2 classification model [75]

From this confusion matrix, there are four main metrics which can be derived. Those being the: Accuracy (Equation 2.2), Recall (Equation 2.3), Specificity(Equation 2.3), and Precision (Equation 2.4).[75]

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

$$\text{Recall} = \frac{TP}{P} = \frac{TP}{TP + FN}; \quad \text{Specificity} = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{2.3}$$

$$\text{Precision} = \frac{TP}{P'} = \frac{TP}{TP + FP} \tag{2.4}$$

Precision and recall, however, exhibit an inverse relationship. This makes tuning a model or assessing overall performance a common challenge in classification problems. For instance, when improving precision, it often comes at the expense of reduced recall and vice versa. To address this, a metric called the Fβ score has been introduced, providing a way to balance both precision and recall when tuning the model or its thresholds by maximising the score.

The Fβ score is sensitive to class imbalance, meaning that if the testing dataset has a different class distribution ratio, the results may become incomparable. Researchers, such as Siblini et al. [76], have attempted to overcome this limitation by suggesting the use of a fixed ratio for the class distribution, ensuring more comparable results across different datasets.

The Fβ score is defined as follows:

$$F_\beta = \left(1 + \beta^2\right) * \frac{\text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \tag{2.5}$$

Where β is a parameter that controls the relative importance of precision and recall. When $\beta = 1$, it becomes the F1 score, which balances precision and recall equally. However, adjusting the value of β allows researchers to emphasize precision ($\beta > 1$) or recall ($\beta < 1$) based on the specific requirements of the problem at hand.

Despite its usefulness in balancing precision and recall, the Fβ score still has limitations in providing a complete performance description, particularly due to its sensitivity to class imbalance and the omission of true negatives in its calculation. Researchers should consider these factors when utilizing the Fβ score for model evaluation and tuning.

the Matthews Correlation Coefficient (MCC) is another measure that was developed to address the limitations of other evaluation metrics, such as the F1 score. Chicco and Jurman [77] highlights the advantages of the MCC over the F1 score, as the MCC takes into account all four categories of the confusion matrix, providing a more comprehensive assessment of model performance.

The MCC is calculated as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.6}$$

The MCC produces a value in the range of [-1, 1], where a perfect performance is represented by 1, indicating that the model's predictions align perfectly with the actual values. A value of 0 suggests completely random predictions, while -1 indicates a complete disagreement between the model's predictions and the actual values [77].

2.6.2. Reciever Operating Curves

The Receiver Operating Characteristics (ROC) curve is another widely used metric for assessing the performance of binary classifiers, and it is commonly employed in the PoD process to display the ROC for a specific defect size [78]. Its origins trace back to the Second World War when it was developed to tune radar operators' thresholds for identifying aircraft. The ROC curve allowed operators to evaluate the trade-off between the competing losses of a higher false positive rate and true positive rate at different thresholds, aiding in decision-making for optimal classification performance [79].

To plot an ROC curve, the False Positive Rate is plotted on the x-axis, while the true positive rate is plotted on the y-axis. By adjusting the threshold, a new point on the ROC plot is generated. Figure 2.12 provides a clear illustration of this process [79]. The ROC curve showcases the model's performance at various thresholds, providing valuable insights into the trade-offs between sensitivity and specificity.

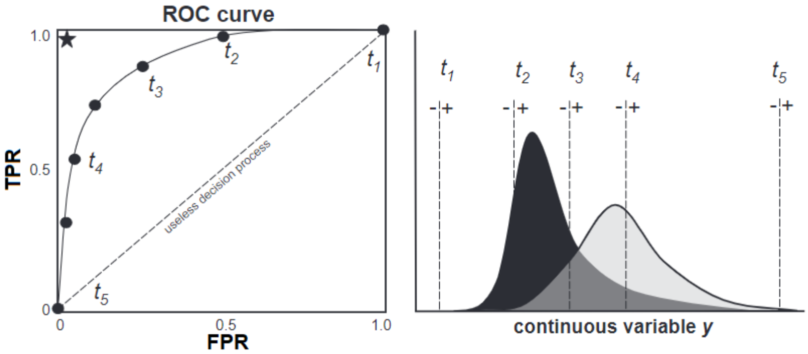


Figure 2.12: Illustration on how the ROC and threshold relate to the probability distribution of the two classifiers [79]. Where FPR and TPR stand for the False Positive Rate and True positive rate respectively. t_i represents the respective threshold applied.

Depending on the underlying probability distributions of the two classes, different ROC curves will be plotted for a range of thresholds, allowing for a better comparison between different models. By comparing the ROC curves of various classifiers, researchers can make informed decisions on which model offers better performance.

Brown and Davis [79] and Streiner and Cairney [80], along with other researchers, have delved into in-depth discussions on interpreting ROC results and inferring confidence in classifier performance. For instance, comparing the Area Under Curve (AUC) of the ROC for different classifiers can help determine which one has a larger area and, consequently, better performance.

In summary, the ROC curve is a valuable tool for evaluating binary classifiers, providing a comprehensive view of their performance across different thresholds and enabling better-informed decisions when selecting the most suitable model.

2.6.3. Metrics in literature

The literature highlights various methods and metrics used to evaluate the performance of machine learning models in the context of ultrasonic inspection and non-destructive testing (NDT). Different researchers have adopted different approaches to assess the effectiveness of their models, leading to a wide range of metrics being used. However, it becomes evident that some studies lack comprehensive evaluation methods, potentially leading to difficulty in comparing results between different papers.

Medak et al. [38] utilized precision and recall as their main metrics, plotting them on a Cartesian plane for different thresholds to compare models. Additionally, they employed bounding boxes to calculate the Intersection-over-Union of damage features against ground truths, providing an additional metric for comparison. However, using ground truths through bounding boxes can be problematic if they are inconsistently drawn, potentially impacting the accuracy of the evaluation.

On the other hand, Munir et al. [40] adopted a model-agnostic approach by using the Mean Average Precision to evaluate performance across different datasets. While this provides a single metric for all data, it may not account for variations in results with high standard deviations but similar means, which could be crucial for certification purposes. Similarly, Guo et al. [37] employed a segmentation model to highlight nodules and used a segmentation accuracy index to evaluate performance. They also used model-agnostic methods such as loss and accuracy. The study by Melville et al. [61] only used accuracy to assess their CNN's performance, which may not provide a comprehensive picture of model effectiveness.

Khan et al. [59] took a more comprehensive approach by using accuracy, loss, and the area under the ROC curve as performance indicators. They also presented confusion matrices for different models, offering valuable data for model validation. While Virkkunen et al. [60] used loss and accuracy as performance metrics for their model and created a POD curve from their test data, it appears that there is a lack of consistency in the metrics employed in ultrasonic machine learning literature. This inconsistency makes it difficult to directly compare results from one paper to another.

Siblini et al. [76] and others have proposed various additional evaluation methods, building on the discussed F1 score, indicating a growing interest in addressing the challenges of evaluating machine learning models in NDT applications. More literature addressing this inconsistency and adopting standardized evaluation methods could lead to improved comparability and a deeper understanding of model performance in ultrasonic inspection.

2.7. Explainability

Due to the evergrowing complexity of AI models it has become more and more important to ensure that the performance of the AI aligns with the defined scope of the model, and if it does not then the goal shifts to understanding why it isnot performing as intended and whether there is a way improve the performance. This requirement has led to the advent of Explainable Artificial Intelligence (xAI) [81] and many authoritative bodies (such as the EU and EASA) require these data-driven models to offer some transparency for the end user [82]. This section aims to discuss the methods which can be applied to these models as to provide some insight and explanation into the inference of a model.

The first notable effort into xAI research was initiated by Defense Advanced Research Projects Agency (DARPA) which sought to develop new processes that result in models which are able to offer explanations for the model's output [81]. Gunning et al. [83] worked with DARPA on this topic and produced a psychological model on explanation. Some of the notable findings were a list of core lessons learned, those being [83]:

1. Explanation only assists user performance if the task is difficult enough to require explanation.
2. Users preferred and trusted a system that offered an explanation over a system which only provided decisions.
3. If the interpretation of the explanation requires a high cognitive load it may hinder user performance.
4. An incorrect explanation is considered extremely valuable for edge-case investigations.

There are many different approaches to extracting an explanation from a model, with many methods attempting to create an explanation in the same form as either the input or output space of the model. This generally makes the comparison of the explanation with the data more intuitive and seamless. One popular dataset used in literature is the MNIST dataset [84] due to its small size and clearly labeled data. Tong et al. [85] made use of this dataset and applied the previously mentioned principle on visualizing explanations. By using graph spectral regularisation they were able to visualize the neuron structure of a NN in order to understand the latent space of the model and identify clusters of neurons responsible for each class. Their results showed promise for models with a high number of classes, however little utility in cases such as a binary classification model.

Doorenmalen and Menkovski [86] similarly used the MNIST dataset to investigate explainability. Their explanations were in the input space of the model so it was possible to directly identify features and relations in their explanations. The premise of their explanation method was to use a weakly supervised generative model as a ”proxy” model. This model would be used to generate images that were able to probe the actual model in question. By identifying the features in the image which contribute to the classification, it ispossible to gradually alter the input image towards the target class and track the point at which the model begins to classify the input differently. Having this progression of a slowly evolving input image can provide an overview of the degree of change that the input image has to undergo before the classification changes, and can provide an indication on the sensitivity of a model. This method however is limited to shallow models since an SVM was used in the process and is limited in its depth.

Where these explainable methods truly become useful are in the methods that offer model agnostic explanations with few limitations. The first of these methods is potentially one of the most prevalent model agnostic methods, known as Local Interpretable Model-agnostic Explanation (LIME) and was first proposed by Ribeiro et al. [87]. In order to be truly model agnostic it must not take into account any of the content within the black box which is the model, which means it must rely on cause and effect. This method modifies a single input over many iterations and assesses the effects that these changes have on the resulting output. This allows LIME, in the case of an image classifier, to provide the end user with a mask that shows the relevant features responsible for the classification. Ribeiro et al. [87] also emphasizes one of the driving considerations for this method was to provide an interpretable representation that is easily understandable to humans and connects back to Gunning et al. [83]'s third core lesson. This method however is computationally intense as it relies on many iterations when modifying the data, and has some pitfalls when dealing with nonlinear data.

Within the image classification field, there exist other methods such as the integrated gradients, first proposed by Sundararajan et al. [88]. Their solution aimed to address the pitfalls of other attribution methods at the time such as another well know attribution method known as Layer-Wise Relevance Propagation (LRP) [89]. One of these advantages is that the integrated gradients method does not suffer from any limitations from batch normalization unlike LRP. This method uses a baseline such as a black image, or a Gaussian noise image. The principle of this method is that the gradients are calculated along a path where the input image (x) linear transforms into the baseline (x'). These gradients are all integrated resulting in the methods name. This integration process is actually not solvable so a numerical idealization is proposed below [88] where m defines the number of interpolations steps between x and x'.

$$\begin{aligned}\text{IG}_{\,i}(x) &= \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha \\ &\approx \left(x_i - x_i'\right) \times \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i} \times \frac{1}{m}\end{aligned}$$

Looking at more specific explainable applications for CNN's there is the GradCam method proposed by Selvaraju et al. [90]. This method is used to determine the location of particular features that contribute to the classification of the image. It achieves this by assigning weights to the feature maps of a specific convolution layer (normally the final convolution layer) and from this it is able to produce a heat map for each of the classes in the model. The weighting is assigned by using the gradients for a specific class to evaluate the feature maps. The benefit of this method is that it can operate independently of the model architecture so long as there are convolution layers present.

Where LIME altered parts of an image slightly and observed the impact, occlusion sensitivity [91] takes a more polarising approach by completely masking areas of the input image and evaluating the impact it has on the feature maps of the model. The only parameters that require tuning are the convolution layer at which the feature maps are being assessed, and the size of the patches used during the process. Its simplicity makes it a powerful CNN agnostic explainable method.

For the more curious there exists even more hands-on approaches to explainability, this involves looking at the activation maps manually to review how the overall filter layer is behaving to specific inputs. This method can be effective in providing a preliminary overview of how the model is operating.

3
Research objectives and breakdown

3.1. Identifying key points of research

The literature in chapter 2 revealed many concepts and developments within the field of AI, NDT, and the wider aerospace industry. Ultrasonic inspection has reached a mature point where it is currently being used throughout the industry. However, the ever-growing number of aircraft and the requirement for skilled technicians to maintain these aircraft have placed a big focus on improving the workflow and processes to meet these demands without overexerting the technicians currently available.

The AI guidelines currently being drafted by EASA provide a clear overview of the expectations of AI development within the industry. Notably, there is a focus on supervised methods, at both level 1 and level 2 automation. The guidelines focus on learning assurance, explainability and safety/risk mitigation as the key areas of development.

Maintenance and training offer the best point of entry for the introduction of AI for use on aircraft. The reason for this is that EASA has deemed it not necessary to conduct an in-depth initial safety assessment on maintenance or training tools during development, stating that it is only necessary for a continuous safety assessment to be done during the service life of these tools. However, there are some conditions to this rule and that is that an initial safety requirement is not needed ONLY if the hardware/technology that the AI is being applied to is already certified. Being able to avoid safety assessments during the development of these tools for maintenance would allow for faster development cycles which would further our understanding and methods for the development of AI within aviation.

Another observation from these guidelines is the ambiguity around evaluating how generalized a model is (in other words how well the model would adapt to the real world and all its edge cases). This ambiguity is highlighted as a major point that requires further research towards a potentially general method of assessing the generalization of these models.

Finally, the guidelines include, in the appendix, minor examples for the end-to-end development of an AI tool for each domain within the guidelines. The maintenance case study is rather shallow in its detail and only covers the early stage of defining the project, but not how one would tackle the development end-to-end as an overview (such as annex 2.1). A real-world application of these guidelines could serve as a better case study and improve the quality of these guidelines.

When comparing these areas of focus to current literature it is evident that there is a lack of comparable and reproducible methods in the application of AI in NDT topics. There appears to be a lack of consistency in evaluation methods and documentation of methodology which makes it difficult to compare with other methods. There was also a lack of transparency on some of these methods, for instance, the lack of explanation on the differentiation of training and testing datasets and how isolated each data set is from one another. These problems make it difficult or impossible to recreate these results. Taking all these points into account when comparing them to the guidelines proposed it becomes apparent that there is no clear process in place to develop an automated system for use in NDT (though it is unsurprising considering these guidelines were first released in 2022) and there is a lack of alignment with the three core areas of focus for the guidelines. Instead, literature has shown a general focus on producing models with the highest accuracy rather than one that guarantees performance and quantifiability in an industrial setting.

What was found in the literature was that, for phased array ultrasonic data, the CNN offers promising performance, versatility, and a large amount of multi-disciplinary literature on the topic. This type of architecture is currently one of the most mainstream architectures which conveniently also performs well with ultrasonic data. Of the various CNN models which exist, the VGGnet appeared to offer the best simplicity-to-performance ratio of all the models. This is relevant since the EASA guidelines advise for simple models. The trade-off of course is that VGG is not very efficient in terms of training since it still has many parameters compared to more novel CNN models currently being used.

Training these models is also a big point of research, with various sources implementing different solutions to improve the generalisation and training performance of their models. In summary, the two most important factors which dictate this is the model architecture and the dataset quality in terms of size and distribution. Different solutions have been proposed for the model architecture including drop-out and normalisation layers. Regarding datasets, a strong emphasis has been made on the size of the dataset, and various augmentation techniques including a novel eFlaw technique (yet to be implemented in composites). Other literature also stated the importance of having a lot of data at the edge cases of the model, in this instance, it was the impact that small and less detectable features have on the model's performance.

As mentioned, the lack of comparability was also due to some literature not including metrics which could be useful for a reader, such as training performance, training regiment or simply not including other metrics outside of the accuracy. For classification models, there is already a large range of metrics which provide a clearer image of the performance of a model, and would ideally be seen in future research.

Similarly, few of these papers attempt to unpack their models and understand the features which they are identifying and assess their generalisation. Currently, there are a large number of explainable methods in the literature which have the potential of providing more information on model performance and the features being identified.

The overall literature has shown considerable development in the topic, with each paper tackling one of the three core development points from the guidelines. However, from this literature and the more comprehensive literature review conducted by Schmidt [3], it is clear there has to be an improvement and alignment in the research process for AI topics in the NDT community.

3.2. Research questions

From the literature review, a clear vision for a research topic is developed. The research topic aims to consolidate various mature methods together to investigate whether it is possible to already develop an AI tool for use in aircraft maintenance while making reference to the EASA guidelines. The main research question of this topic is stated as follows:

How would a qualifiable maintenance machine learning model, which is capable of classifying damage in ultrasonic scans of aerospace-grade composite panels, be developed?

Within this topic, there are three main sub-questions that are focussed on. These subquestions are as follows:

1. What are the impacts of the eFlaw concept on the process?
2. What are the types of features the model identifies?
3. How can the output of the model be interpreted and used in a practical ultrasonic non-destructive testing application, and what are the implications for safety and reliability in the aerospace industry?

Where each of these questions focuses on the three development points in the guidelines. Subquestion 1. focuses on learning assurance with its potential to further generalise the data. Subquestion 2. focuses on AI explainability, and subquestion 3. focuses on AI safety and risk mitigation by investigating an effective way to communicate the results to the user for effective decision-making.

The expectation of this thesis is not to develop a cutting-edge model with the highest accuracy achievable, but rather to develop a transparent process which other researchers can use to quickly develop their own tools, which they themselves can optimise for whichever metric they want to maximize.

3.3. Research overview for certifiable AI for NDT applications in aerosapce

In this section, a high-level overview of the research plan is discussed. One of the goals of this thesis is to provide a template for research which works towards contributing to the AI guidelines drafted by EASA, with the goal of aligning research to the common goal of developing solutions which bring the industry closer to qualifying maintenance AI tools for industry.

For all of the deliverables in this overview, there are two fundamental concepts which must always be considered. Those are Transparency and generalization. The guidelines require everything to be traceable and reproducible, as a result, all methods should adhere to these requirements. As for the generalization point, it's important to consider edge cases and model performance for the less controlled environments within aircraft maintenance compared to production. Below is an itemized list of the deliverables expected in the research methodology:

1. *Concept of Operations:* This deliverable expects an overview of the AI-based system. This will include the vision of the operation, its interfaces and end-users. The concept of operations in the guidelines is more thorough, but for the purposes of research are not completely required. Instead providing the concept of operations provides the reader with a clear understanding of what this tool should be capable of and who/what it will be interfacing with. This includes a clear definition of what level of automation this model will be achieving.

2. *Data acquisition:* As the name would suggest, this deliverable explains the process of capturing all the data that will be used for training. It is important to include setups, specimens and more importantly any relevant distributions that might exist within the data. This deliverable should provide the reader with a clear understanding of why the specific methods used for data acquisition were made, particularly with a focus on learning assurance.

3. *Data management:* This is a continuation of learning assurance and transparency as it provides a clear overview of how the data was managed throughout the project. This includes version control and data integrity, but also augmentation techniques and how the data will be separated between the testing and training data sets.

4. *Model design and training:* The model architecture should be discussed. This will include a full plot of the model, a description of the input and output of the model and any relevant features within the model. After that, the training regiment, hyperparameters and hardware that were used to train the model are discussed.

5. *Model evaluation:* Another area to emphasise is the evaluation metrics being used to evaluate the model performance. For classification models, it is encouraged to include at the very least the confusion matrices, the accuracy, recall, specificity, precision or F_{B} scores. Transparency on the thresholds of the model is also encouraged.

6. *Explainability:* A method of inferring trust in the model is through explainable methods. It is important that explainability is always considered for the model to ensure that the model is performing for the correct reasons.

7. *Safety and control:* Finally, the safety and control of the tool must be considered. This ties the tool back to the concept of operation as it considers what the end user will see, but also what information should be communicated and how decision-making should be structured.

This overview provides a general set of deliverables/considerations to take into account when designing tools for aviation maintenance with the goal of introducing them into the field. The sections to follow will be an application of these points to develop a tool with the potential of seeing service in the field. The remainder of this thesis will illustrate these points by making use of PAUT of thick composite panels to develop a ML damage classification model to assist technicians in locating areas of interest.

4
Experimental setup and methodology

In this chapter, both the setup and methodology will be discussed for the experiments required for this research topic. The experimental phase can be broken up into two categories, data collection, and model development. The data collection portion of this chapter will discuss the complete inspection setup, which includes the fixtures, PAUT tool, and inspection procedure. the specimens that were manufactured for inspection, followed by the damage introduction into the panels which includes explaining how the damage was distributed throughout the panel. The model development portion of this chapter will discuss how the data was processed and prepared for training, the different models that were selected for this research, and the procedure that was used for training the aforementioned models.

4.1. Concept of operation

In this section, the concept of operations will be described. This will involve the vision of the system, what the tool should be capable of, and who will be interacting with the tool. This will be followed by an explanation of the level of automation this tool will achieve. The overall vision of this tool is to assist the inspector during the use of PAUT wheel probes. These probes can produce a large number of scans and data that will have to be evaluated by the inspector using a series of gates and tests. To aid in this process it is proposed to develop a model which can detect discontinuities in composite panels. This tool should be capable of detecting whether a B-scan either contains damage or not and appropriately labelling each scan as such. Outside of that, the inspector will still hold the responsibility of evaluating the scans and providing a final decision on where damage is present. The end users for this tool will be level 2 personnel as they have the necessary knowledge of inspection techniques to be aware of the limitations of ultrasonic inspection, and as a result, will be more qualified to disagree with the model's result. This also provides the added benefit that the level 2 inspector will be able to perform inspections more efficiently with this tool if perfectly integrated into the inspector's workflow.

Since there are no qualified AI systems currently in place within maintenance, this system shall be restricted to an automation level of 1-B. This is defined in the EASA AI guidelines and states that it shall provide human cognitive assistance in decision and action selection [2].

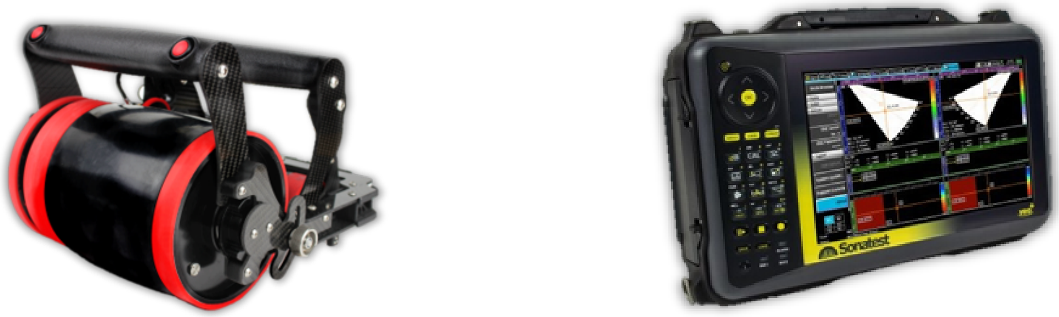
4.2. Inspection setup

In this section, the setup and equipment used to gather the ultrasonic data to create the dataset will be explained.

The primary goal of this setup was to create a rigid fixture that would hold the samples rigidly in place while also providing a reliable guide for the ultrasonic roller probe to accurately scan the same location of a sample. The reason for this rigidity and reliability was primarily to cater to the E-flaw concept which would require scans that are positionally identical to one another for comparison.

4.2.1. Phased Array Ultrasonic Roller Probe

The inspection tool was the first place to help define the setup. Figure 4.1 shows the inspection equipment used. This PAUT setup was the most recent roller probe addition to the NDT laboratory. The roller probe itself consists of fifty 1.9 mm wide elements with a pitch of 2 mm, resulting in a scan width of approximately 100 mm. The probe is equipped with an encoder set to create a scan every 1 mm and the capture method used is the linear phased array scan (more details on the probe and settings can be found in Appendix A). At this encoder resolution **FMC!** (**FMC!**) was not feasible due to the time it takes to process the captured data.



(a) Sonatest's large low-frequency roller probe. [92]

(b) Sonatest's VEO+ phased array flaw detector. [93]

Figure 4.1: Images of the inspection equipment used to inspect the samples and produce ultrasonic data. More information on the technical setup of the probe and flaw detector can be found in Appendix A

4.2.2. Fixture

In the introduction to section 4.2 it was briefly mentioned that a rigid fixture was required for the panels which also included a consistent guide for the roller probe to create reproducible scans. To achieve this the NDT laboratory has a set of AluFix high-precision machined aluminium fixture blocks designed for the purpose of creating temporary high-precision fixtures [94]. Figure 4.2 shows the final fixture with a panel already fixed into it and the PAUT probe on the sample. The fixture was designed to constrain the ends but allow the underside of the panel to be suspended in the air to emulate the boundary conditions of the panel as it would be in a real structure. This setup requires panels to have a length of 500 mm and a width of 250 mm and allows for two lanes with a length of 415 mm to be scanned in a repeatable manner.

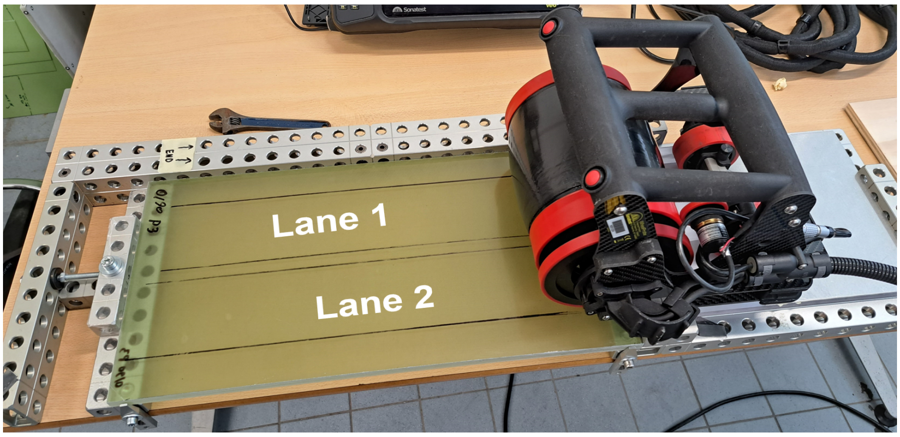


Figure 4.2: Image of the inspection setup assembled using high-precision aluminum blocks to assemble the fixture. It includes three points of contact to fix the panel to the rest of the fixture. This setup allows two lanes of the panel to be scanned (requiring the panel to first be rotated before the second lane may be scanned) and maximizes the usable area of the panel.

4.2.3. Ensuring repeatablity

To implement the eFlaw concept discussed in section 2.4 two positionally identical scans are required to ensure that the signals are subtracted as best as possible. Despite the rigid fixture, it was still not a trivial task to create a repeatable scan due to other factors. This section will discuss how the coupling agent application, starting position of the roller probe, starting position on the panel, pressure and speed during scanning, and the mounting of the panels affect the quality of the scans.

The first factor was the coupling agent, originally water was sprayed onto the surface, but it was quickly found that the coverage was not consistent or repeatable on the surface of the panels. The cause was due to the water forming small pools in random areas, resulting in variations in amplitude and minor phase shifts on the scans. To overcome this inconsistency, a gel-based coupling agent was applied with a flexible spatula to create a consistent layer across the entire scanning surface. Using the spatula to redistribute the coupling gel before every scan ensured consistent coverage on the probe.

The second factor was the starting position of the roller probe. In the case of this probe it was found that there were some systemic errors within the probe itself which resulted in reproducible artifacts within the scans and can be seen in Figure 4.3. This error was accounted for by marking a starting point on the roller and etching a starting point on the panel. As part of the inspection process, the roller would be aligned with the markings before every scan.

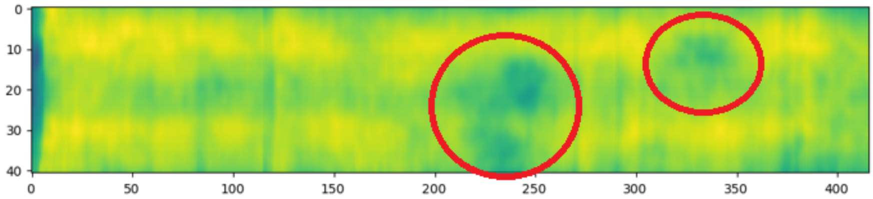


Figure 4.3: Ultrasonic scan of an undamaged panel. The red circles indicate the artifacts which are formed as a result of what is believed to be debris on the inner surface of the wheel probe causing variations in the acoustic impedance.

The third factor involved the starting point of the roller and the impact it had on the initial scanning quality. On the far left side of Figure 4.3 a dark region can be seen, this region was caused if the roller probe initially did not have enough coupling agent on it and would cause the first few millimeters of the scan to have poor coupling to the sample. This was made worse during the alignment of the probe if it was rotated with the probe on the surface of the sample in a counterclockwise direction as it would further remove coupling agent from the roller's surface which was next in the circumference to roll on the surface. The solution was to only rotate the probe clockwise when aligning the probe, and to also ensure proper coverage of the coupling agent before beginning a scan.

The fourth factor related to the pressure and speed applied to the roller probe during scanning. The effects of this can be seen in Figure 4.3 by strong vertical lines shown at x equals 80 and x equals 120. This error is caused when there is either a small jerk which results in a change of speed but also in pressure. The remedy to this problem was one of practice, which involved developing a wide stance to allow for the entire distance to be covered in one single motion. Secondly, the amount of pressure applied should not be too high as it could cause fatigue very quickly. The weight of the probe itself is almost sufficient to create adequate contact with the surface and only required minor pressure.

The final factor relates to mounting the panels to the fixture. When rescanning a panel it is important that it is constrained to the fixture the same way it was constrained initially. This way the boundary conditions on the panel in terms of clamping force and position remain consistent. As a result, it was kept track of how tight each panel was tightened and if required how many shims were used at each point of contact to ensure proper mounting.

These five factors played the largest roles in ensuring that the scans between an undamaged sample and the same sample after the damage was introduced would align as best as possible, and it will be discussed and shown to what extent these methods proved successful in the subsequent sections.

4.3. Inspection specimens

For this research, a series of composite panels were required for inspection. In this section the selected materials will be discussed, followed by the layup selection, manufacturing process, and the final panels.

4.3.1. Materials

Selecting the appropriate materials required a series of trade-offs and considerations based on availability, budget, inspection method, and equipment. Ultimately the material selected for use was a Glass Fibre Reinforced Polymer (GFRP) composed of HexForce 7581 [95] fiberglass as the reinforcement. This reinforcement is a certified aerospace-grade reinforcement and would serve as a suitable analog for similar aerospace structures. The matrix used was a resin system composed of Epikote 04908 as the resin and Epikure 04908 as the curing agent [96]. This selection was made for various reasons, the first being that GFRP has the benefit that it is partially transparent, an important quality to have when evaluating the initial manufacturing quality of the samples, and later when classifying the data for training(discussed in section 4.5). GFRP is also a common material used within aviation as it is a cost-effective alternative to its carbon fibre alternative. Finally, this material was available for use, free of charge, at the Delft Aerospace Structures and Materials laboratory (DASML). This was an important trade-off within the budget of this thesis since thick composite panels (greater than 10 millimeters) had to be manufactured. This thickness requirement came from the PAUT tool selected for this project and the low-frequency signal that it inspects with. More information on the PAUT tool can be found in section 4.2.

4.3.2. Layups

Two layups were selected to further generalize the data that would be used for training, validation and testing. These layups were a uniform [0/90] layup and quasi-isotropic layup. The selected woven fabric was a 8 Harness satin weave, with a thickness of 0.23mm. Since the PAUT tool used for inspection was a low frequency probe with large incident waves with a wavelength of 5mm, it was chosen to make 10mm thick panels to allow for the incident and back-wall reflection to be easily identified. At 10mm thickness a total number of 44 layers would be required per specimen. As a result, the resulting layups are listed below:

$$\text{0/90 layup: } [[(0,90)]_{22}]_S$$

$$\text{quasi-isotropic layup: } [[(0,90)/(\pm45)_2/(0,90)]_5/(0,90)/(\pm45)]_S$$

These layups were decided conservatively as fibre volume fractions were not accounted for. The result is that the panel would be slightly thicker than the intended 10mm, but this was deemed acceptable since the additional thickness would only further separate the incident and back wall reflections.

4.3.3. Manufacturing process

The manufacturing of these specimens took place at the DASML composites laboratory. A total of 150 plies were cut using the lab's GERBER cutter, an automated single-ply cutting machine. All plies were cut to a dimension of $300x550mm$ where these dimensions account for a margin to cut the panels to their final dimension of $250x500mm$. The process used for creating the panels was vacuum infusion and Figure 4.4 contains illustrations showing the layout of the infusion setup. Once the panel was infused the resin was left to cure at room temperature for 24 hours, or until the resin in the tubes was fully cured and could break the tubing once bent. After this, the panel would be demoulded, and the mould could be cleaned and prepared for the next specimen.

After all the panels had been manufactured, it was possible to post-process the panels to the correct dimensions. Using a guided circular diamond saw each panel was cut to size with an accuracy of ±1 mm. At this point, the panels were ready to be placed in the fixture and ultrasonically inspected. Figure 4.5 shows Figure 4.4 in practice with one of the panels being infused. This figure also shows the resin front and includes time stamps of when the resin front reached the flowmesh until when the resin front eventually passes the layup (with an approximate time of infusion at 5 minutes).

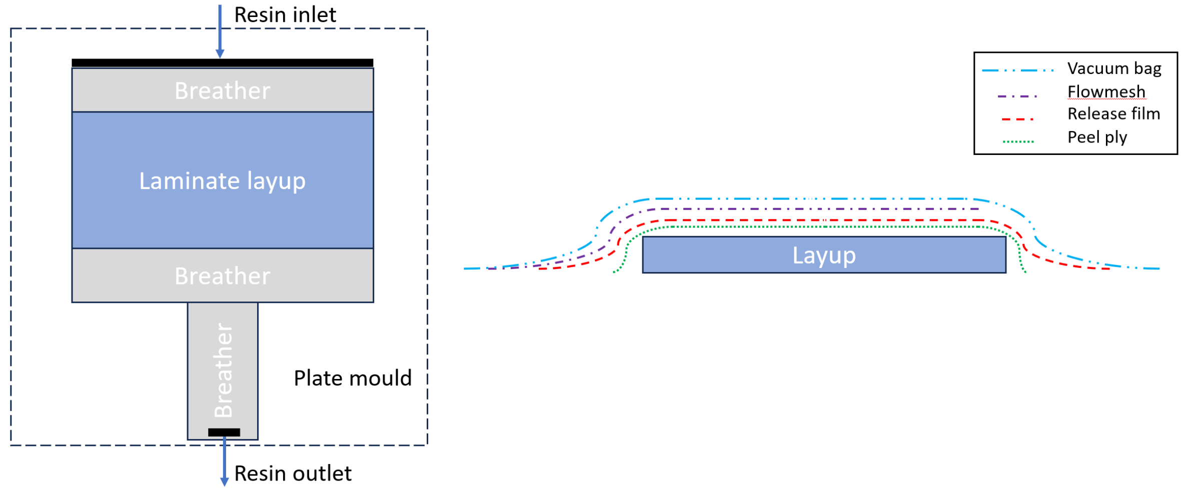


Figure 4.4: Illustration showing the infusion setup for the specimen creation. LEFT: Top view of the infusion setup, showing the flat aluminium plate mould where the layup is built on, the inlet and outlet points of for the resin, and finally the areas where the breather is placed to encourage resin flow. RIGHT: Side view of the infusion setup showing all of the layers placed ontop of the layup for the infusion process. In the right figure the flow of the resin is from left to right (a relevant point for flow mesh which is the only asymmetric layer in the figure).

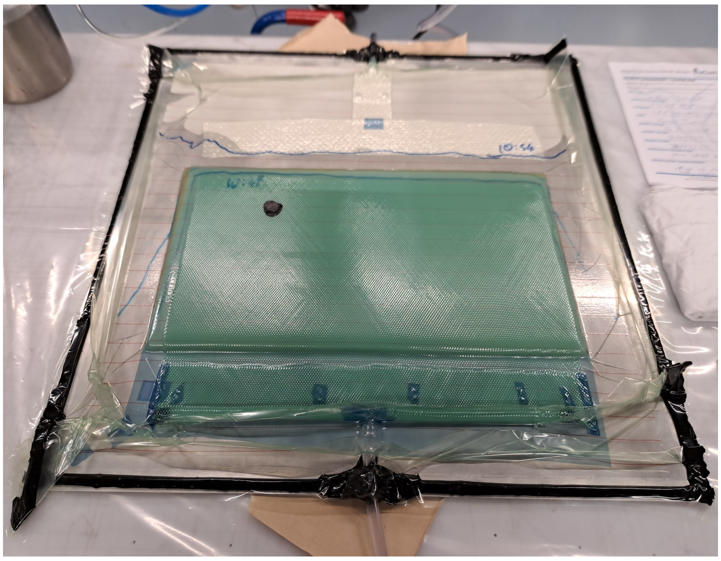
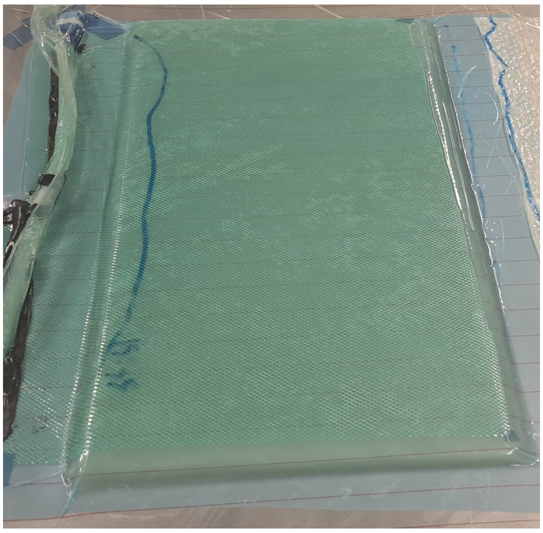


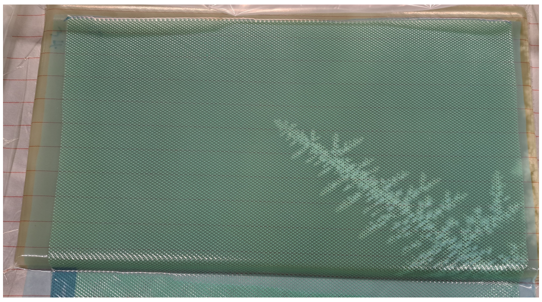
Figure 4.5: Image of the infusion process. Note the resin front can be seen in this image with two-time stamps of the resin front. The first is at the end of the flowmesh and the next one is after the panel. The total time for an infusion of the layup was approximately 5 minutes.

4.3.4. Final panels

In total four panels were produced. However, two of the four panels experienced a leak during infusion overnight which resulted in either the entire or parts of the panel being lost. These effects can be best seen in Figure 4.6, where the nature of the manufacturing errors is shown and described. Panels 2 and 3 experienced no issues during manufacturing and were considered perfect specimens for the training portion of the thesis.



(a) Panel 1: a large amount of air entered the system while the resin still had a relatively low viscosity resulting in air spreading throughout the panel. The result is a panel that was deemed unusable for any purpose within the scope of this thesis.



(b) Panel 4: a pin-sized hole which opened in a pleat overnight during curing resulted in a lightning-strike shaped ingress of air, the afflicted area would not be usable as an undamaged portion, and would thus be excluded from the training and validation data.

Figure 4.6: Images of the two panels which saw their vacuums either partially or fully fail.

Due to the infusion process and setup, the thickness variation along the panel was a difficult parameter to correct for. This was caused by the phenomena where resin pressure decreases from the inlet toward the outlet, the result is a difference in thickness along the width of the panel which was observed [97]. Various methods exist to correct this, but either require advanced equipment or an iterative fine-tuning of the vacuum pressure. With the limited production number, this was not practically feasible, however, it was found that reducing the vacuum pressure to 70% vacuum after gelation provided fairly consistent thickness values. This is shown in Table 4.1 where this was applied to panel 3 and resulted in only a 0.07 mm variation in thickness.

Table 4.1: Table showing the final thickness dimensions and layups of each of the manufactured panels. Each value was measured using callipers on each of the corners of the panel. This provides an initial indication on the geometry of the panel.

	Layup	**Corner 1**	**Corner 2**	**Corner 3**	**Corner 4**	**Max. Difference**	**Avg. Thickness**
	-	**mm**	**mm**	**mm**	**mm**	**mm**	**mm**
Panel 1	0/90	11.73	11.82	11.82	11.82	0.09	11.80
Panel 2	Quasi	11.45	11.6	11.51	11.53	0.15	11.52
Panel 3	0/90	11.94	12.01	11.95	12.01	0.07	11.98
Panel 4	Quasi	12.13	11.98	11.89	11.94	0.24	11.99

The end result of this process was four GFRP panels being manufactured, with the first panel consisting of too many voids to be considered usable for this thesis. The remaining three panels are fully usable, with the exception of panel 4 which includes a region where air entered the laminate during curing. This section will not be used for training but can be used as an edge case to test the model's ability to generalise and detect voids that are not representative of the training data.

4.4. Damage distribution

It is important that a suitable distribution of damage is selected so that both the training and assessment of the model can be done effectively. This became additionally challenging due to panel 1 having had a poor infusion and reducing the number of usable panels from four down to three. This section will discuss the factors to consider when selecting the types of damage and distribution of damage within the panel.

4.4.1. Types of damage

In section 2.3 the different types of damage which can be introduced to a panel were discussed. Since one of the goals of this thesis was to investigate a novel augmentation technique called the ”eFlaw”, it was necessary to first have baseline scans of an undamaged sample before introducing the damage. Additionally, the labelling of data was also considered important, so damage techniques that may have left some ambiguity with regard to whether the damage was present or not were also considered important to avoid. The selected damage type was then the flat-bottomed hole, an analogue for delaminations within a laminate. This damage can be easily introduced into the panels by drill press or Computerized Numerical Control (CNC) depending on the requirements for damage size and accuracy of the damage location.

Due to the thickness of the panels an additional dimension can be varied more accurately, and that is the depth of the damage. It was discussed earlier in section 4.3 that the thickness of the panels was designed in such a way to include three different regions, one where the incident wave occurs (the near field), one at the back wall reflection, and one between these two regions. The near field among other areas has always been a difficult region for inspectors due to complex wave interactions that can occur with discontinuities, making damage detection difficult. As a result, it was deemed necessary to include damage at all three of these depths to see how a model would perform at different depths.

4.4.2. Distribution of damage

There are three major forces that drove the selection of the damage distribution. The first is the distribution itself, the driving force for this was the requirement to eventually develop a PoD curve. Annis et al. [22] conducted an in-depth analysis on the PoD curve using the Monte Carlo method to understand the impact of different parameters on the curve. Their findings showed that due to the uncertainty on the position and shape of the PoD curve, a right-skewed distribution would provide the most effective security against this uncertainty (Though in an ideal situation where the position and shape are known, a uniform distribution can prove the most effective [22]). The second is the ideal sample size for an effective PoD, this was also investigated by Annis et al. [22] who found that sample sizes greater than 60 exhibit diminishing returns on the effectiveness of the curve. The third driving force is the range of damage sizes. There are multiple sources of literature that impact this parameter, the first set of literature from Annis et al. [22] and the United States Department of Defence [25] describe the requirement to include damage which guarantees a miss on the detection of damage. This is important since the Monte Carlo simulations found the best PoD curves when the distribution covered probabilities from 0.3 to 0.97, this translates to having a distribution that includes guaranteed misses but also guaranteed hits. Research by Koskinen et al. [42] researched the effect that different flaw data has on the effectivity of a damage classification model. Their findings showed that the training data should predominantly focus on the smaller damage sizes as models generally require more information to learn the smaller features.

Since the performance of the model is not yet known, selecting a damage size that guarantees a miss was difficult to select. Within the NDT community, there has been a general rule of thumb to define a method’s sensitivity by relating the minimum size of damage to the half wavelength of the ultrasonic tool. The origin of this rule of thumb has been difficult to locate, with the oldest source tracing back to the Krautkrämer brothers who had done extensive research on the matter since the 1960s however the publication in which this rule of thumb is mentioned could not be found outside of a discussion on a forum [98]. In a more recent publication by the brothers, Krautkrämer and Krautkrämer [99] describes various topics around the principles of ultrasonic inspection, the relevant part of this publication is the effect of what happens as a defect size approaches and becomes smaller than that of the wavelength. The effect they describe is that as the defect becomes smaller, the scattered wave begins to take on a more spherical shape with the reflection lobe and shadow lobe merging into one. The effect that this has is that more of the energy is distributed and a very sensitive system will be required to detect this defect. As a result of this phenomenon, a conservative damage size of one-quarter of the wavelength was selected. This value was selected in part due to the previous points but also due to the manufacturing limitations and not being able to reliably make smaller damage.

4.4.3. Distributing the damage into the panels

In subsection 4.4.1 and subsection 4.4.2 the types of damage and the distribution the damage would follow was discussed. This subsection will now discuss how, given those parameters, the damage was distributed throughout the panels.

Before distributing the damage a clear set of requirements was required for each distribution. For the training data, the following requirements had to be met:

1. Contain damage ranging from 1.5 mm to 10 mm
2. Contain damage in at least three different depths of the panel
3. Contain variation to further generalize the data.
4. Make the holes more likely to identify and classify correctly.
5. The distribution of damage at a certain depth must be a right-skewed distribution and follow the method described by Annis et al. [22].
6. All distributions must fit within 2.5 lanes within the panels.

For the testing/PoD distribution, the following requirements had to be met:

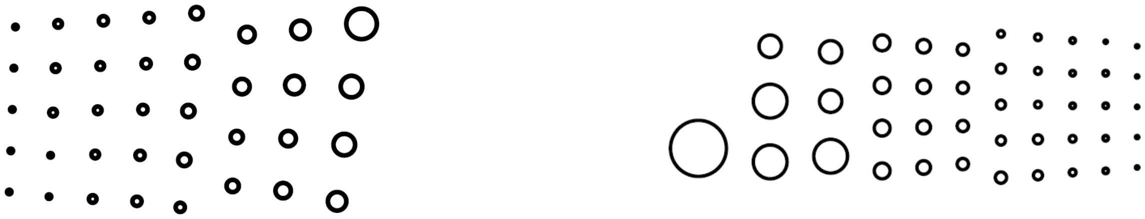
1. Contain damage ranging from 1.5 mm to 25 mm
2. Contain depths and damage sizes outside of the training data.
3. The total number of damages at a certain depth must be greater than 60.
4. The distribution of damage at a certain depth must be a right-skewed distribution and follow the method described by Annis et al. [22].
5. Each damage distribution for a specific depth must be equally distributed between two panels of different layups.
6. All distributions have to fit within 3 lanes on a panel.

It must be noted that a lane is defined as a section of the panel with a dimension of $415mmx80mm$ the 80 mm is a conservative margin applied to ensure that the damage will fall within the roller probe's 100 mm width and not outside its edges.

To distribute the holes while maintaining sufficient separation between each hole a distance of 2 wavelengths (12 mm) was enforced between the edge of each flat bottom hole. For the training data, it was also opted to add a slant to the rows of holes, there were two reasons for this decision: The first was that when labelling the ground truths it would be more difficult to find a single row of holes. This way if the holes are staggered, but still overlap it would be possible to confidently label a damaged region. The second was for more generalized data. Staggering the holes will mean that a B-scan will contain a hole with the full diameter with one hole, and perhaps the edge of an adjacent hole, providing more variation in the data.

The resulting distributions for a single ”unit” of distributed damage are shown in Figure 4.7. In this figure, the applied slant can be more clearly seen. To emphasize the effectivity the example of the 1.5 mm holes will be used. If originally left as one straight row, then the encoder would only catch one slice of the damage or in the best case two partial slices. With this slant applied it's possible to capture the 1.5 mm holes across a 4mm distance.

More details on the distributions can be seen in Figure 4.8, which plots the number of holes per unit. It can be seen that the intended distribution of a right-skewed distribution was achieved. However, one note is that when these damages are broken down into the number of B-scans which can be retrieved from each damage, the resulting distribution changes significantly. This will not affect the



(a) Single Training damage distribution for a single unit consisting of 37 holes

(b) Single Testing/PoD damage distribution for a single unit consisting of 44 holes.

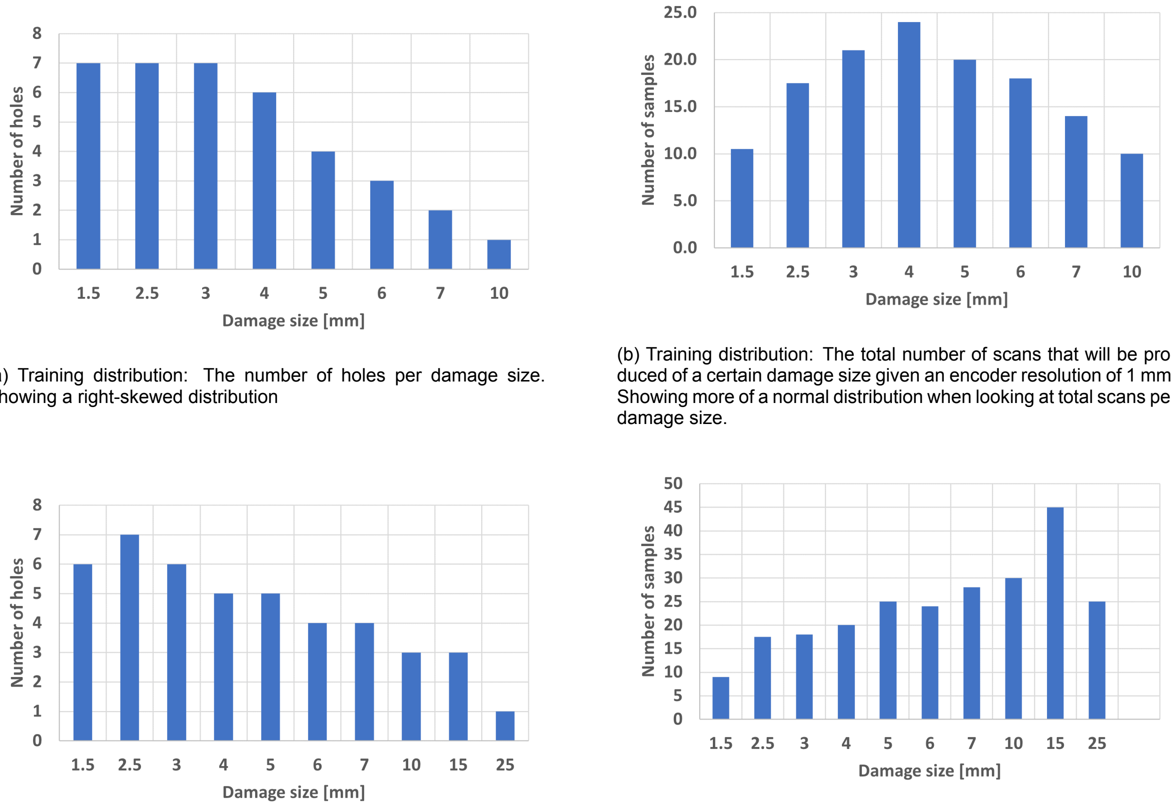
Figure 4.7: Images of the final units used for the damage distribution in the panels for both training and testing.

Table 4.2: Table showing the different depths used in each of the distributions including the number of units (shown in Figure 4.7) per depth.

	List of different depths	**Number of units per depth**	**Number of holes per depth**	**Number of holes per dataset**
	[mm]	**[ ]**	**[ ]**	**[ ]**
Training dist.	3, 4, 5, 6, 7, 8	1	37	222
Testing/PoD dist.	3, 6, 8.5	2	88	264

points discussed earlier, since this variable is technically accounted for in the PoD as it's a logical conclusion that damage that occupies more scans has a higher chance of detection.

The final point to discuss is the different depths that these units are placed in, and the total number of holes per depth. This is best described in Table 4.2. What can be seen from all of the figures in this section is that the original requirements set out have been met and two distinct distributions are ready to be machined into the specimens that were manufactured. For more details on the actual placement of the holes on the panels please refer to Appendix B.



(a) Training distribution: The number of holes per damage size. Showing a right-skewed distribution

(b) Training distribution: The total number of scans that will be produced of a certain damage size given an encoder resolution of 1 mm. Showing more of a normal distribution when looking at total scans per damage size.

(c) Testing/PoD distribution: The number of holes per damage size. Showing a right-skewed distribution

(d) testing/PoD distribution: The total number of scans that will be produced of a certain damage size given an encoder resolution of 1 mm. Showing more of a left-skewed distribution when looking at total scans per damage size.

Figure 4.8: Damage distribution plots showing the different distributions for the training data and testing/PoD respectively on each row of figures.

4.4.4. Manufacturing the damage

It has already been briefly mentioned that the smaller flat bottom holes actually begin to approach the limits of conventional machining, but there are other trade-offs that had to be made due to manufacturing limitations.

The damage was introduced into the panels through the services at the Dienst Elektronische en Mechanische Ontwikkeling (DEMO), who focus on the rapid development and construction of unique experimental setups and prototypes. At DEMO they used a CNC mill to precisely introduce the damage into the panels as described in Appendix B. The only limitation that was made clear was that it would not be possible to make the smaller holes flat bottom holes and that normal drill bits would have to be used for this process. Table 4.3 shows the line at which this switch occurs. This compromise was deemed acceptable for two reasons:

- The first one is that at those sizes the effect of the 135 degree point angle drill bits is not that large. with the largest vertical displacement being 0.41 mm with the 5 mm drill bit.
- The second point refers back to subsection 4.4.2 where the work of Krautkrämer and Krautkrämer [99] was discussed. All of the drill bits are coincidentally also for holes smaller than the wavelength of the roller probe. This means that the lobes of the scattered wave and the shadow will begin converging to form a spherical wave rather than having distinct lobes. In essence, this means the angles induced by the drill bit tip will have even less effect on the interactions with the waves.

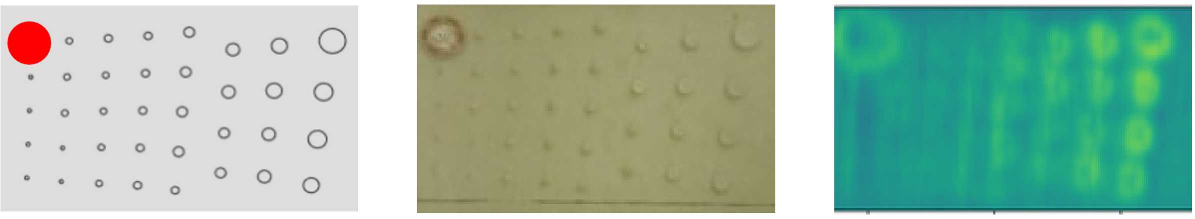
Any effect the drill bits might have on the resulting waves will only reduce the signal strength and as a result, this trade-off has been considered conservative and absolutely acceptable for the given application.

With these compromises considered, the panels were sent to DEMO to be manufactured. Of the 523 holes that had to be machined, only one had a manufacturing defect where the CNC crashed into

Table 4.3: sizes of holes which are made using drill bits, and which ones are made using an endmill (i.e. a flat bottom hole)

	HSS drill bits					End mill				
Size [mm]	1.5	2.5	3	4	5	6	7	10	15	25

the panel resulting in a portion of the panel being damaged and resulting in a shallow hole and a drill bit stuck in the panel. This damage is best illustrated in Figure 4.9. Fortunately, this damage occurred on a panel used for the training data. Since the model will be training on binary input data, this damage is still local enough to confidently label and was deemed suitable to use for training. The only changes were to update the CAD to include this data.



(a) CAD drawing of the plotted holes. The red circle indicates the unintentionaly damaged region.

(b) Image of the damaged panel. The burnt epoxy can be seen in the image caused by the spindle when it crashed into the panel.

(c) Ultrasonic damage features using the E-flaw concept illustrating the effect of the defect in the ultrasonic domain.

Figure 4.9: Different images illustrating the manufacturing defect present in panel #2

4.5. Data preparation

In this section, the dataset used for training will be discussed. This includes the data acquisition, data types, formats, and data management methods employed during the research.

4.5.1. Data acquisition

The data acquisition phase was broken into two distinct parts. The first part consisted of scanning the undamaged panels using the inspection setup from section 4.2 in order to establish the baselines. After this, the panels were sent to DEMO to introduce the damage to the panels, as explained in section 4.4. Once the damage was manufactured into the panels, it was possible to scan the panels again using the same inspection setup. The result from these scans is best summarised in Table 4.4, where the total number of scans per lane is summarised. Different types of scans were made including those where the probes wheel was at the same position every scan and a set of scans where the position of the wheel was randomized. This was only done for the baselines as it was important to record all the necessary data before damaging the panels. It was later determined that it would not be necessary to scan the damaged panels using random starting positions, resulting in the column of zeroes seen in the table.

Table 4.4: Overview of the total number of scans made of each of the panels' lanes (The naming convention is: 'Lane ' + *panel number* + *lane number*).

	Undamaged panel		**Damaged panel**	
	Same position	**Random position**	**Same position**	**Random position**
Lane 21	6	9	14	0
Lane 22	12	9	14	0
Lane 31	12	9	14	0
Lane 32	12	9	14	0
Lane 41	15	5	14	0
Lane 42	12	5	15	0
Total	**69**	**46**	**85**	**0**

To elaborate further on how no random position scans were made of the damaged panels related to the E-flaw concept, which set the goal to isolate individual flat-bottom holes for the model to identify individually, thereby increasing the density of damage one can have on a panel. Since the E-flaw concept requires the scans to be aligned, there was no requirement for scans that began at random positions.

All of these scans were saved on the VEO+, and exported via USB. The total size of all the files is 14 GB, with each file being saved in sonatest’s native filetype, ’.utdata’, which can be used with their proprietary software. Using the software, every file is exported twice as a ’.csv’ file, with the first Comma-Seperated Values (CSV) file being the raw ultrasonic values, and the other CSV containing the normalised values between 0 and 1 of the scan.

4.5.2. Data management

Both the native filetype and CSV’s were still not easy to work with, requiring the data to be further processed. The exported CSV files were in a format where every B-scan was stored as a table. The process involved importing this data and cleaning the file of unnecessary formatting information in order to form a pandas data frame which can be easily saved and imported every time the scan needs to be interacted with.

Before reformatting any of this data, it was important to use a rudimentary method to check the integrity of the files. This check was done by using the file size of each CSV as a checksum to ensure that all of the original files were exported in their complete capacity. This checksum led to a large sum of files being found to be improperly imported due to the export process being interrupted. More detailed checksums could have been performed but were deemed outside of the scope of this thesis as this topic is one that has already been researched quite heavily.

Once it was certain that the entire files had been copied over, the reformatting could begin. The first objective was to reformat the data into a three-dimensional numpy array with the following dimension:

$$41x2122x416$$

Where the first dimension consists of each element used by the probe (in this case 41 due to the line-scanning method), the second dimension shows the entire A-scan of each element (in this case the sampling rate and sampling time resulted in 2122 samples per A-scan), and the third element is the number of B-scans within the scan (in this case fixed to 416 due to the settings and encoder). Once the data is formatted as a numpy array it is then formatted into a pandas multi-indexed dataframe. This places the data into a format which is easily callable and interactable, making future use of the data far easier than the original data (this data and its dimensions are illustrated in Figure 4.11). These data frames were then saved to new CSV files, with a new naming convention being adopted for traceability. This convention is shown in Figure 4.10.

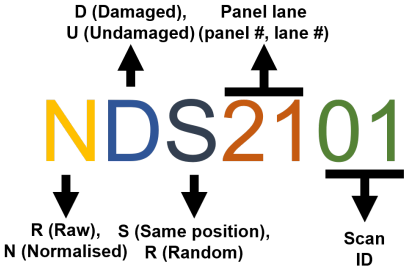


Figure 4.10: Naming convention used for exported scan data. This provided the initial starting point for traceable data.

With these files saved the data processing can take place, first by experimenting with the E-flaw concept, and then by formatting the data to be fed into the model. The result of this preparation is that all of the data is in an easy-to-read format, with unique and traceable naming. Allowing for the quick access of A-, B- and C-scans allowing for complete manipulation and use of the data with little hindrance. This can be neatly illustrated in Figure 4.11

4.6. Augmented eFlaw method

In this section, the eFlaw concept implementation will be discussed. This will cover the initial principle, implementation considerations, quality improvements, and flaw isolation.

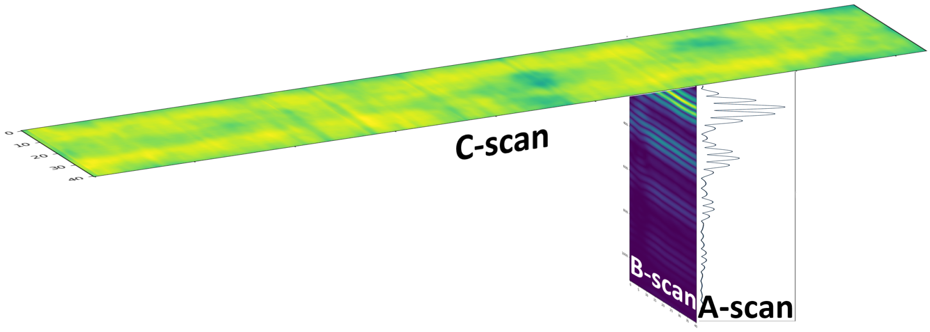


Figure 4.11: Visualisation of the prepared data and the possible representations of the data which can be easily extracted.

4.6.1. eFlaw principle

This concept is described by Virkkunen et al. [33] as an alternative to simulation. It requires that there first be a scan of a defect-free baseline sample. After this sample is scanned damage should be introduced into the same sample, and a new set of scans generated. Subtracting these two scans from one another should in principle result in the damaged features being the only remaining data present in the data. This concept can be expressed with quite a simple equation as shown in Equation 4.1:

$$D - B = E \tag{4.1}$$

Where D denotes the damaged signal, B denotes the undamaged baseline signal, and E denotes the extracted damage features. Since Equation 4.1 is a linear equation, it can be rearranged to solve for D. However, if the extracted features are manipulated in anyway, or the baselines changed for another then the resulting D will not be representative of the original D used to determine E. As a result, from this point on any D generated using this eFlaw concept will be referred to as D' and in the text referred to as augmented damage.

$$B_x + E_y = D' \tag{4.2}$$

In Equation 4.2 both of the terms (B and E) can be manipulated or replaced with different variants. This is denoted by the subscripts x and y, the result is an augmented damage scan (D') which should be unique from the original D used to create the extracted flaw signal used. The uniqueness naturally depends on the extent of the manipulation and variation of Equation 4.2's terms.

4.6.2. Practical considerations when implementing

The principle described in subsection 4.6.1 is quite simple, however, the implementation is less trivial since the largest challenge in the process is the acquisition of a baseline and damaged signal which aligns perfectly. The set-up described in section 4.2 attempts to account for these features and align the signal as best as possible. Despite these efforts there are still a series of factors which result in the baseline and damage mismatching. Some of these factors are:

- Variations in pressure during scanning causing minor phase shifts
- Jerk or jitter during scanning could cause misalignment and phase shifts
- Quality of coupling can cause variations in the signal strength and quality
- Encoder slipping due to the coupling agent getting on the encoder wheel could cause the incorrect locations to be captured, resulting in misalignment.
- Offset of the starting location would cause the entire scans to be misaligned.

All of these factors can be broken down into three possible impacts on the eFlaw methods: Phase misalignment, Amplitude misalignment and Misaligned B-scans

4.6.3. Phase misalignment

The phase misalignment has various different analytical and numerical solutions. The solution settled on was to take each damaged A-scan and compare it to its undamaged twin.

The solution to comparing the two phases of the signals was to use cross-correlation to determine what areas of the signals correlate with one another and shift the one signal to the respective amount.

This cross-correlation was achieved using the SciPy python library, where SciPy could expedite the calculation using the fast Fourier transform to compute the convolutions. Since both of these signals are 1D arrays the cross-correlation of these signals will be in the form [100]:

$$z[k] = (x * y)(k - N + 1) = \sum_{l=0}^{\|x\|-1} x_l y^*_{l-k+N-1} \tag{4.3}$$

for $k = 0, 1, ..., ||x|| + ||y|| - 2$

Where x and y represent each of the signals, $||x||$ is the length of signal x, and N is the maximum length of the two signals.

This method would be done by stepping through each B-scan and comparing every A-scan of those B-scans individually.

4.6.4. Amplitude and B-scan misalignment

The amplitude misalignment was problematic to resolve, since the impact of defects on the amplitude via shadows, or interactions in the near field did not offer any confident corrective measure that could be taken without the risk of changing the damage features too much. As a result, it was decided to not change the amplitude in any way.

Similarly, mismatched or misaligned B-scans would be difficult to identify, particularly if the encoder had slipped slightly. As a result, the opted-for method was to take a large number of measurements and find the best pairs of undamaged and damaged scans which yield the least noisy results.

The solution used to combat these two problems was to use an image comparison metric referred to as the Structural Similarity Index (SSI) [101]. This method is better than other classical methods such as the Mean Square Error (MSE) because this method also assigns an interdependence between pixels spatially. As a result, this method assigns a score to the location of features but also the quality of the features.

The only problem is that this method requires a ground truth, which was easily resolved using the Computer Aided Design (CAD) data to create a mask of what the data should look like Appendix C. Using these masks in conjunction with the SSI it would be possible to create a SSI matrix with all of the damaged/undamaged pairs for the eFlaw concept. The higher the score of the SSI the more closely the extracted flaws are meant to represent the ground truth. The SSI in this case is not being used to assess the individual scores, but rather as a metric for comparison (i.e. the magnitude does not matter as much as the difference between the values).

4.6.5. Flaw isolation

Once a flaw can be extracted, the possibility to augment the damage increases. The vision for this thesis was to introduce a large set of damage near one another, and then augment the damage so that individual damage can be isolated in a scan. The benefits of this are two-fold, the first is that it's more affordable to create the test samples for a large range of damage because the spacing can now be close to one another and can be introduced at once (reducing the cost of manufacturing the damage as well). The second benefit is that now the possibility to augment damage grows further since it's possible to create different combinations of damage via this process of omission. Applying this method of selective omission can yield a total number of combinations using Equation 4.4.

$$c = \sum_{r=1}^{n} \binom{n}{r} = 1 + \sum_{r=1}^{n-1} \frac{n!}{r!(n-r)!}; n \geq 2 \tag{4.4}$$

Where c is the total number of combinations, n is the total number of features (damage) in the scan to which the method is being applied, and r is the varying number of features to have present in the scan. As a short example, a scan with 5 damages present in the frame could then yield a total of 31 unique frames containing different damage. Refering to the PoD damage distribution unit from Figure 4.7 then the 44-hole unit has the potential to have 215 unique hole combinations. When the number of scans is considered then it increases from the minimum of 90 scans without the eFlaw method, to approximately 873 scans using the eFlaw method to create different combinations of B-scans. This increases the generalisation of the data and augments the data by almost a factor of five for the number of damage features present, and almost a factor of 10 for the number of unique B-scans which can exist.

The actual implementation of the omission requires using the extracted feature component E from Equation 4.1 and using the coordinates from the CAD data to crop out the undesired features. Once this is done Equation 4.2 can be used to reintroduce the features and create a damaged B-scan. Variations in the scan between the baseline and damaged scan, as well as the effects the damage has (e.g. shadows), can leave noticeably hard lines in the scan. The solution to this problem was to linearly interpolate 4 pixels of the baseline with the boundary of the cropping for the extracted features.

4.7. Machine Learning model design

During the development process, two separate models were trained to evaluate the impact of transfer learning models versus originally trained models since it was feared there may be insufficient data to converge during training. This section will discuss the two machine-learning models produced and will involve the input data, the model design and the output of the model.

4.7.1. Input data

Both models are CNN's with the one being a transfer learned CNN and the other one being a custom CNN. The driving parameter for the shape of the input data depends on the transfer learning model and on the transition from the convolution layers to the fully connected layers [102]. At this transition, the information is flattened and fed into the fully connected layers. If the dimension of the flattened information were to change it would no longer feed into the fully connected layers properly and the model would cease to work. Since many of the pre-trained models have square input sizes, it remains beneficial to create a dataset with a square shape. This is because, during the transfer learning, it might be decided to unfreeze some of the fully connected layers. This would then require the input shape to be identical to the original training dataset. Other reasons for square inputs exist as well, for instance in the variation of input data, if there are data with different aspect ratios (different scans/structures) then padding can be used without having to retrain the model.

The next question then becomes what the actual dimension of the input shape should be. Another observation from the pre-trained models is that the axis' of their input shapes are powers of 2 and can be represented as 2^n. The reason that these choices are made is twofold. The first is that no matter how large the input size is, it will always be divisible by two, this is relevant for pooling, downsampling or convolution layers since an odd number will result in some information being lost. The second reason focused on performance where efficient memory allocation and memory use can significantly improve the computation time of any computation [103]. These pre-trained models were also trained on large image data sets such as ImageNet [104], this means that the input files are three-channeled data. These three channels would allow for the data set to contain some spatial information about the neighbouring B-scans.

The data preparation process is best illustrated in Figure 4.12. During this process the ultrasonic scan data was cropped so that the signal contained the incident wave, back wall reflection and two echos. The resulting dimension was 41x984. The aspect ratio on this was far too high to consider padding into a square image, as a result, it was then opted to rather downsample the image to 41x164, being sure to at the very least preserve the Nyquist frequency so that some signal information is preserved. This new size still had a considerable aspect ratio so the data was reshaped to an 82x82 square shape so that no padding would be required at all. Adhering to the three-channel limit of the pre-trained models the final dimension of a single training image is 82x82x3.

Ultimately the only factor which may have been better to consider would have been the shape of $82x82$, which in this case should not be so consequential, but if these models are intended for use with other data of different structures or sensors, then a more standardised square dimension would be preferable.

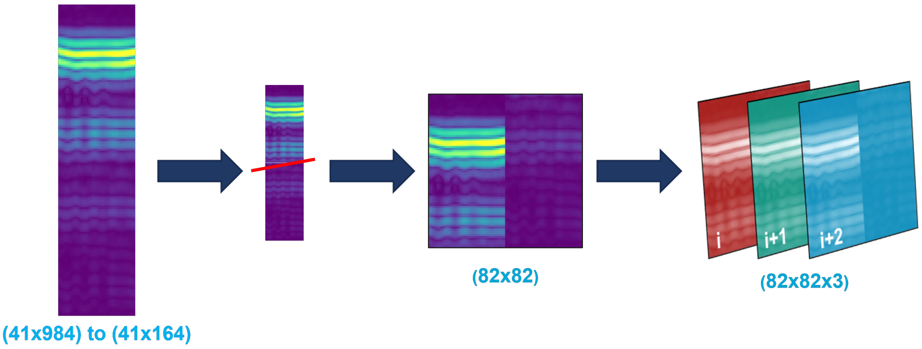


Figure 4.12: Illustration of the steps taken to format the ultrasonic data into data which can be fed into the model.

4.7.2. Model architectures

As mentioned in subsection 4.7.1 two separate models were developed, one custom model, and one pre-trained model which would have transfer learning applied to it. The reason for this was the fear that there was insufficient data to properly train a full CNN from scratch and the curiosity to see other performances between the two. This curiosity stems from the risk of overfitting a model and is directly related to the number of parameters in the model and the amount of data available. There are no clear solutions to determining the size of the fully connected layers, with many sources advising for this to be a hyperparameter which is tuned [105]. However, some sources attempt to provide some methods to estimate the beginning values for this optimisation based on the size of your training data. The premise of this concept is to use the degrees of freedom of the system, combined with a 'generalisation' factor and the size of the available dataset [106]. This concept can be explained in Equation 4.5

$$N_h = \frac{N_s}{\alpha * (N_i + N_o)} \tag{4.5}$$

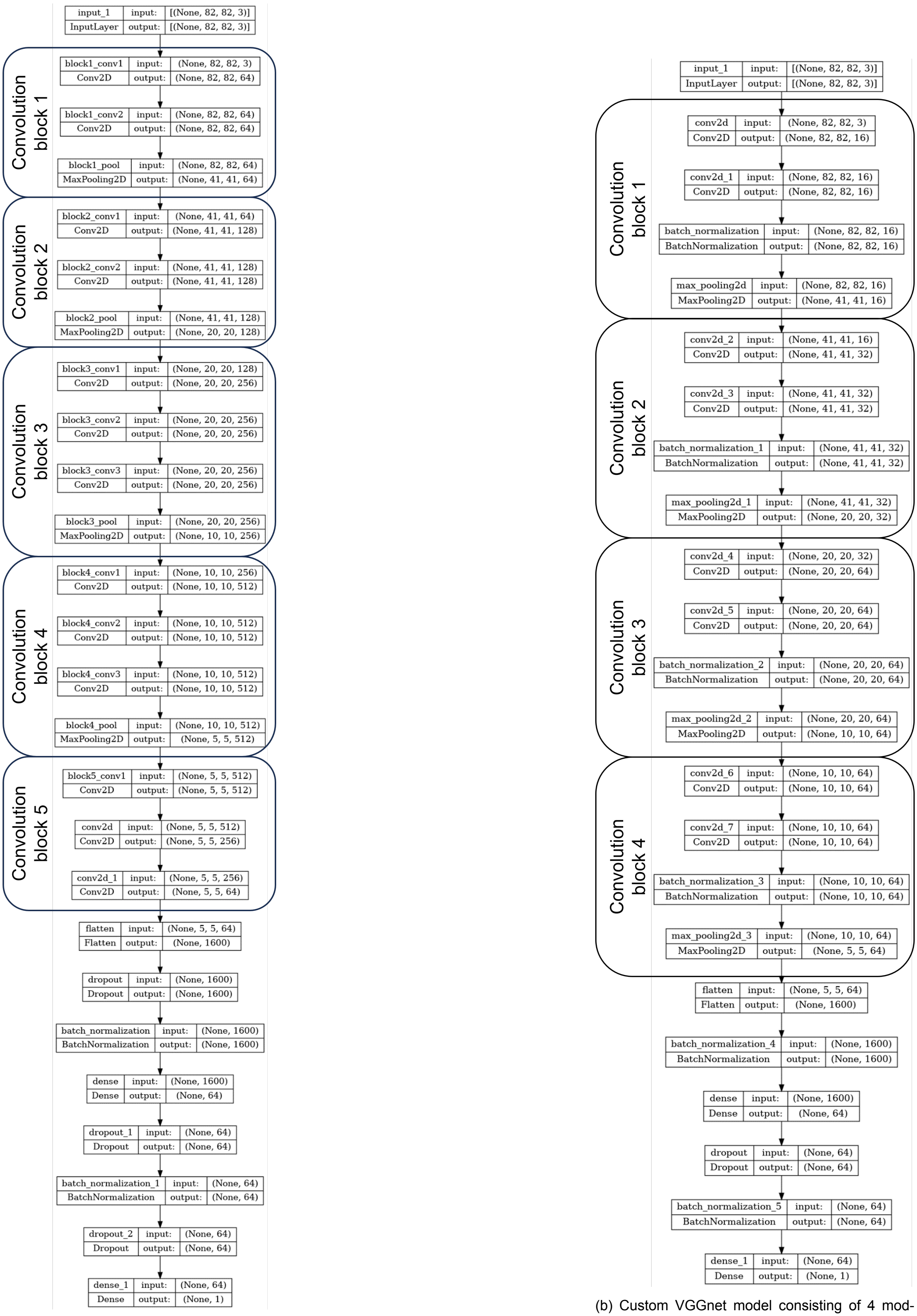
Where N_h is the number of neurons, N_s is the number of samples in the dataset, α is an arbitrary value ranging from 2-10 and represents the 'generalisation' of the model, N_i describes the number of input neurons (the dimension of the final convolution layer after flattening), and N_o number of output neurons (in this case with a binary classification only a single neuron is required)

This guideline will provide an indication of whether there is sufficient data to train the model, but not if the selected number of neurons will be suitable for the desired task of the model. This links back to the previous problem where this becomes a parameter that still needs to be optimised. Thankfully other researchers have developed their own CNNs to classify similar data and the number of neurons and hidden layers used by them can be used as another indication for the starting point.

The transfer learned model can be seen in Figure 4.13a, and is based on the VGG16 architecture [107] trained on the ImageNet dataset [104]. This model has over 130,000,000 parameters and was trained on over 14,000,000 images. Within the aerospace industry datasets of that size are difficult to come by and training such a model would be unimaginable. However, transfer learning allows us to freeze the weights on a large portion of this model and retrain the parameters of interest. In the case of this model, all of the subsequent layers after the fourth convolution block were deleted, and new layers were introduced to tailor the model toward the binary classification and complexity of the data.

The custom CNN will be based on the Visual Geometry Group (VGG) architecture, specifically a heavily modified version of the architecture proposed by Siljama et al. [54]. This comes with the advantage of this architecture showing promising results with ultrasonic data and provides a good point for selecting the number of convolution blocks and hidden layers. The differentiating features with this model versus Siljama et al. [54] are an additional convolution block to further reduce the size of the data, and the additional use of batch normalization in the hidden layers. The model architecture can be seen in Figure 4.13b.

These two models are the final models after an iterative process of optimization. They were trained on the ultrasonic data produced in section 4.5. This training process will be discussed in greater detail in section 4.8.


input_1 InputLayer input: [(None, 82, 82, 3)] output: [(None, 82, 82, 3)]
Convolution block 1
block1_conv1 Conv2D input: (None, 82, 82, 3) output: (None, 82, 82, 64)
block1_conv2 Conv2D input: (None, 82, 82, 64) output: (None, 82, 82, 64)
block1_pool MaxPooling2D input: (None, 82, 82, 64) output: (None, 41, 41, 64)
Convolution block 2
block2_conv1 Conv2D input: (None, 41, 41, 64) output: (None, 41, 41, 128)
block2_conv2 Conv2D input: (None, 41, 41, 128) output: (None, 41, 41, 128)
block2_pool MaxPooling2D input: (None, 41, 41, 128) output: (None, 20, 20, 128)
Convolution block 3
block3_conv1 Conv2D input: (None, 20, 20, 128) output: (None, 20, 20, 256)
block3_conv2 Conv2D input: (None, 20, 20, 256) output: (None, 20, 20, 256)
block3_conv3 Conv2D input: (None, 20, 20, 256) output: (None, 20, 20, 256)
block3_pool MaxPooling2D input: (None, 20, 20, 256) output: (None, 10, 10, 256)
Convolution block 4
block4_conv1 Conv2D input: (None, 10, 10, 256) output: (None, 10, 10, 512)
block4_conv2 Conv2D input: (None, 10, 10, 512) output: (None, 10, 10, 512)
block4_conv3 Conv2D input: (None, 10, 10, 512) output: (None, 10, 10, 512)
block4_pool MaxPooling2D input: (None, 10, 10, 512) output: (None, 5, 5, 512)
Convolution block 5
block5_conv1 Conv2D input: (None, 5, 5, 512) output: (None, 5, 5, 512)
conv2d Conv2D input: (None, 5, 5, 512) output: (None, 5, 5, 256)
conv2d_1 Conv2D input: (None, 5, 5, 256) output: (None, 5, 5, 64)
flatten Flatten input: (None, 5, 5, 64) output: (None, 1600)
dropout Dropout input: (None, 1600) output: (None, 1600)
batch_normalization BatchNormalization input: (None, 1600) output: (None, 1600)
dense Dense input: (None, 1600) output: (None, 64)
dropout_1 Dropout input: (None, 64) output: (None, 64)
batch_normalization_1 BatchNormalization input: (None, 64) output: (None, 64)
dropout_2 Dropout input: (None, 64) output: (None, 64)
dense_1 Dense input: (None, 64) output: (None, 1)
input_1 InputLayer input: [(None, 82, 82, 3)] output: [(None, 82, 82, 3)]
Convolution block 1
conv2d Conv2D input: (None, 82, 82, 3) output: (None, 82, 82, 16)
conv2d_1 Conv2D input: (None, 82, 82, 16) output: (None, 82, 82, 16)
batch_normalization BatchNormalization input: (None, 82, 82, 16) output: (None, 82, 82, 16)
max_pooling2d MaxPooling2D input: (None, 82, 82, 16) output: (None, 41, 41, 16)
Convolution block 2
conv2d_2 Conv2D input: (None, 41, 41, 16) output: (None, 41, 41, 32)
conv2d_3 Conv2D input: (None, 41, 41, 32) output: (None, 41, 41, 32)
batch_normalization_1 BatchNormalization input: (None, 41, 41, 32) output: (None, 41, 41, 32)
max_pooling2d_1 MaxPooling2D input: (None, 41, 41, 32) output: (None, 20, 20, 32)
Convolution block 3
conv2d_4 Conv2D input: (None, 20, 20, 32) output: (None, 20, 20, 64)
conv2d_5 Conv2D input: (None, 20, 20, 64) output: (None, 20, 20, 64)
batch_normalization_2 BatchNormalization input: (None, 20, 20, 64) output: (None, 20, 20, 64)
max_pooling2d_2 MaxPooling2D input: (None, 20, 20, 64) output: (None, 10, 10, 64)
Convolution block 4
conv2d_6 Conv2D input: (None, 10, 10, 64) output: (None, 10, 10, 64)
conv2d_7 Conv2D input: (None, 10, 10, 64) output: (None, 10, 10, 64)
batch_normalization_3 BatchNormalization input: (None, 10, 10, 64) output: (None, 10, 10, 64)
max_pooling2d_3 MaxPooling2D input: (None, 10, 10, 64) output: (None, 5, 5, 64)
flatten Flatten input: (None, 5, 5, 64) output: (None, 1600)
batch_normalization_4 BatchNormalization input: (None, 1600) output: (None, 1600)
dense Dense input: (None, 1600) output: (None, 64)
dropout Dropout input: (None, 64) output: (None, 64)
batch_normalization_5 BatchNormalization input: (None, 64) output: (None, 64)
dense_1 Dense input: (None, 64) output: (None, 1)

(a) Transfer learned VGG16 model trained on ImageNet [107][104], consisting of 5 convolution blocks. All parameters until (and including) the 4th convolution block are locked from training. Total parameters: 11,431,681; Trainable parameters: 1,433,281; Non-trainable parameters: 9,998,400

(b) Custom VGGnet model consisting of 4 modified convolution blocks. All layers are trained. Batch normalization was introduced into the model as opposed to the original VGG architecture. Total parameters: 255,825; Trainable parameters: 252,145; Non-trainable parameters: 3,680

Figure 4.13: The model architectures for both of the models used in the experimental phase.

4.8. Model Training and evaluation

In this section the training process of the model will be discussed. It will cover the hardware used, considerations for the dataset and how this data was augmented.

4.8.1. Hardware and software

All of the data processing, programming and training was done on Linux through a Windows subsystem for Linux. All of the programming was done using Python with the primary packages being Tensorflow [108], Scikit [109], and SciPy [110]. All of the training was accelerated using Graphics Processing Unit (GPU) acceleration. The hardware used for processing and training was done on a personal computer equipped with:

- Central processing unit: AMD Ryzen 7 5800X
- Graphical processing unit: NVIDIA GeForce RTX 3080 Ti
- Memory: 32 GB

4.8.2. Data Augmentation

Data augmentation plays a critical role in the generalization of the data for training and has been shown to positively impact the performance of models and reduce or avoid overfitting as introduced in greater detail in section 2.4. Given the limited amount of data that was collected, augmentation serves an even more important role in the training of these models. Five primary augmentation techniques were adopted, ranging from classical augmentation methods to less conventional techniques tailored for this dataset. These augmentation techniques are mirroring the channels, horizontal mirroring of B-scans, phase shifting of B-scans, wrapping the B-scans and the eFlaw augmentation. These techniques are described in greater detail below:

- **Mirroring the channels [x2]:** This augmentation technique is rather simple, and requires swapping the order of the channels of the input data as illustrated in Figure 4.14. This augmentation mimics the effect of having scanned the same area from the opposite direction and can increase the data by a factor of two.

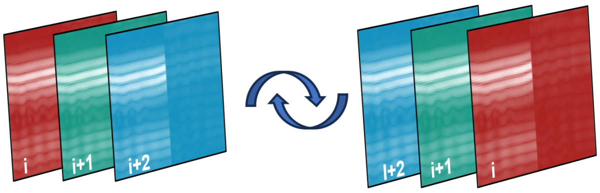


Figure 4.14: Channel mirroring technique

- **Horizontal mirroring of B-scans [x2]:** This method involves mirroring the B-scans along the vertical axis so that it appears as though the damage was on the opposite side of the panel and is illustrated in Figure 4.15.

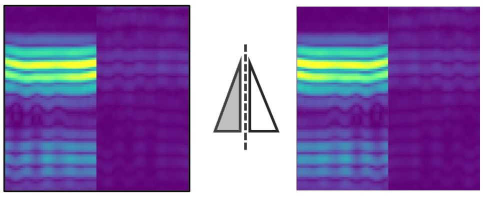


Figure 4.15: Horizontal mirroring B-scan augmentation

- **Phase shifting of B-scans [x5]:** In this augmentation technique, the ultrasonic signal was phase-shifted up or down randomly to 5 different positions. This simulates the effect of varying pressure on the roller probe and is illustrated in Figure 4.16.

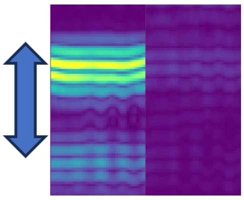


Figure 4.16: Phase shifting B-scan augmentation

- **Wrapping the B-scans [x41]:** This method is a less conventional method, which involves shifting the B-scan to the right and wrapping the pixels that fall outside the right side back to the left side. This is a method to shift the location of the damage more randomly to ensure the features cover the entire width. This is better illustrated in Figure 4.17.

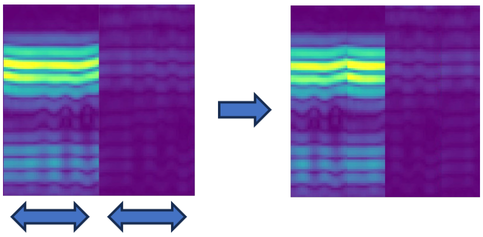


Figure 4.17: Wrapping of B-scans for augmentation

- **eFlaw augmentation:** This method takes advantage of the eFlaw concept to create new unique data which isolate specific features. Figure 4.18 illustrates this best with how the method is able to isolate both of the damaged features and produce two completely new B-scans.

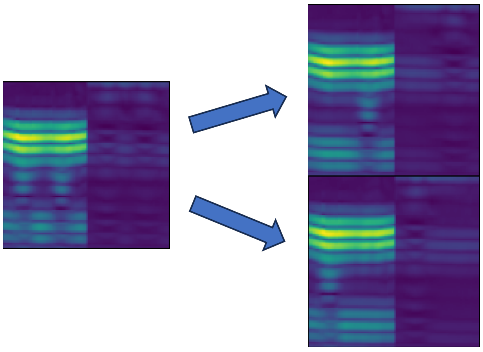


Figure 4.18: Augmentation using the eFlaw method to isolate specific damage features.

The result of these augmentation methods is that it was possible to take approximately 7000 data points and increase it to 79,000 using the eFlaw and phase shifting augmentation methods. Applying the other further methods results in the potential of the model being able to train on 13,000,000 different B-scans. This provides the dataset with an augmentation factor of approximately 1800.

What should be noted here is that the eFlaw concept was not fully utilized. The method was only used to isolate each flaw but was not used to create various combinations of the damaged features. For instance, a B-scan which contained 5 features would only produce 5 unique B-scans as opposed to the calculated 31 using Equation 4.4. If properly implemented the potential size of this dataset could be significantly larger. The reason for only partially implementing the method was that this would be the first application of the method and a cautious approach to the application was taken to avoid potentially detrimentally affecting the dataset.

4.8.3. Overview of the datasets

From all of the scans made during the data acquisition stage, it was possible to create two sizeable datasets one for training and validation, and one for testing. Each dataset has the following number of training images (with square brackets showing the ratio of images which contain damage to images which do not contain damage):

- Training & Validation: 78,500 [46/54]
  - Original B-scans: 12,700
  - Isolated eFlaws: 15,000
  - Phase shifted: 50,800
- Testing: 39,132 [50/50]
  - Original B-scans: 28,781
  - Isolated eFlaws: 10,349

The Training & Validation dataset is split into an 80/20 split across training and validation. What should be noted about the Training & Validation dataset is that this does not include the other augmentation methods, which are only applied during training through a custom generator. The testing dataset will not be augmented beyond the eFlaw concept used to isolate the damage features.

4.8.4. Training process

This topic had hardware limitations which made the training of the model a challenging endeavour. Originally it was intended to train the model using cross-validation as an effective means to assess generalization, however, to complete 50 epochs would take in excess of 8 hours and cost a significant amount of energy. This meant that there was insufficient time to tune the hyper-parameters and conduct cross-validation at the same time, so cross-validation had to be left for future research. All of the models were trained using ADAM as the optimization algorithm, with binary cross entropy being used as the loss function (which is only applicable to this case because the desired output is binary). The hyper-parameters to tune were the learning rate, the size and number of fully connected layers and the number of epochs. The tuning of these parameters was done one variable at a time incrementally to see the effect it had on convergence and performance.

4.8.5. Explainability

Various explainable methods were used to assess the model performance. These methods were implemented using the tf-explain library [111]. Four explainable methods were used: LIME, Integrated gradients, GradCam, and Activation maps. Some of these methods require some fine-tuning with respect to identifying the correct layers to probe. In the case of these models, the final VGG blocks were often far too coarse. This resulted in the heatmaps being rather coarse in resolution and less interpretable.

4.9. F-scan implementation

A novel way to evaluate large PAUT scans was through the development of the Feature-scan (F-scan). The goal of this view was to reduce every B-scan evaluated down into a vector so that every vector could be compiled to produce a C-scan, whose results come from a feature map rather than the raw ultrasonic data. The purpose of this scan is then to improve the identification of damage by marking relevant B-scans as damaged, and by including the explainable features which lead to that classification. With the goal that these explainable features provide the inspector with local information on where to search for damage.

This method was implemented through four steps: identify relevant feature map(s), extract feature maps, average the columns to produce a vector, and finally plot all these vectors together. The final step provides a top view of the sample through the ”eyes” of the model. Using the F-scan view in combination with a C-scan could help inspectors significantly.

This concludes the development of the tool, as now there is a data-driven tool which can identify B-scans of interest and assist the human in their decision-making and workflow.

5 Results

This chapter will show all of the results from the experimental method. Each subsection will be broken down into two parts. The first part will present the results, including descriptions of the results. The subsection is then concluded with a brief evaluation of these results and the interpretation and implications of these results. Three main sections will discuss the results. The first will discuss the eFlaw concept and its feasibility. The second section will discuss the model performance. This will include the training results and model performance on the testing dataset. Finally, the points regarding quantifiability will be discussed. This will involve explainability, and human/machine integration.

5.1. eFlaw demonstration

This section will discuss the findings of the eFlaw concept. It will show the results of the impact that the phase correction had on the quality of the results, the SSI best pair matrices, the damage feature reintroduction capabilities, and the damage isolation.

5.1.1. Phase correction

The phase correction method as described in subsection 4.6.2 was used to minimize the noise during the feature extraction. To illustrate the method two baseline scans of the same location from two different scans were compared. Figure 5.1 shows the A-scans of these two signals.

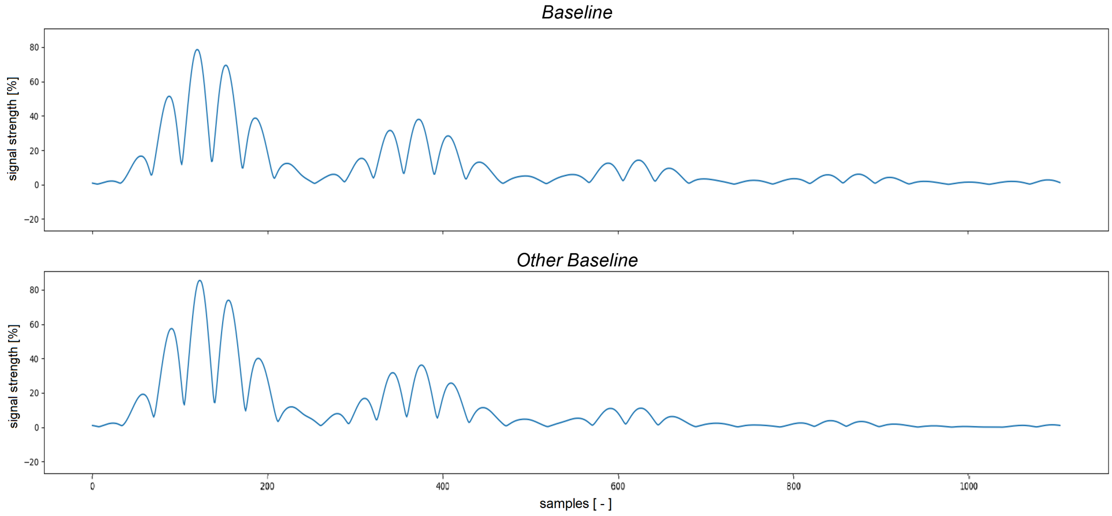


Figure 5.1: A-scans of two separate scans from the same location.

Figure 5.3 shows the resulting signals from the subtraction of these two baselines. The top A-scan of the figure shows a large absolute variation of ± 20% in signal strength. The bottom A-scan shows the impact of the phase-shifted and correlated signals before subtraction. The resulting signal shows a smaller range of only ± 5%. When comparing this variation to the maximum signal amplitude of the baseline signals of 80% the variations of the phase-shifted subtracted signal represent only a variation of ± 6.25% from the amplitude. Though it must be noted that a significant part of this variation is due to a misalignment in amplitudes, a topic which was acknowledged as difficult to correct without jeopardising the quality of the damaged features.

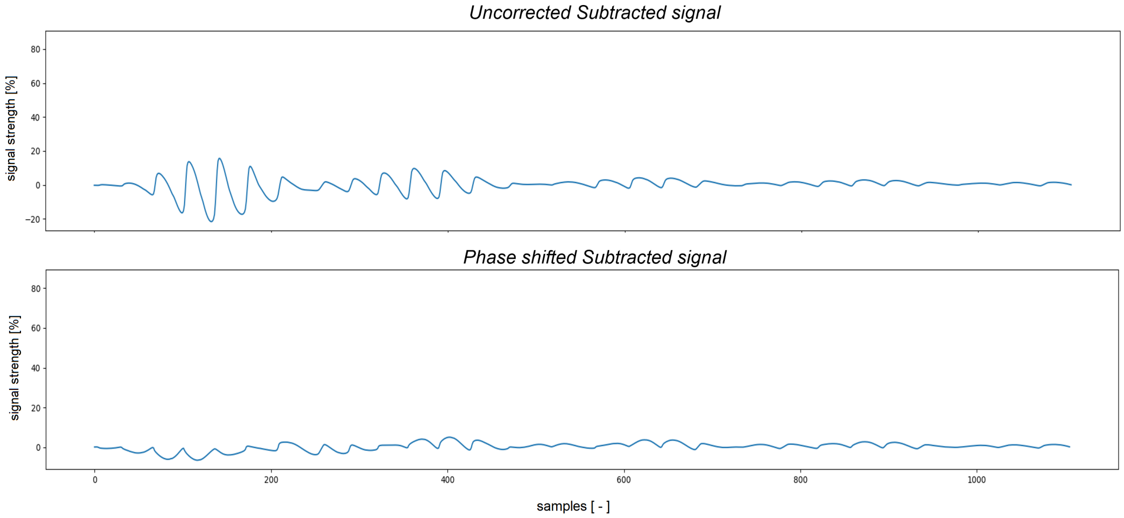


Figure 5.2: A-scans of the resulting subtracted signals from the two baseline signals in Figure 5.1. The top A-scan shows the resulting signal with no corrections applied to the signal and the bottom signal shows the impact that the phase shift correction has on the subtraction.

The impact that this change has on the results can be more clearly illustrated in the C-scan view and is shown in Figure 5.3. It shows the maximum value of the extracted features with the undamaged regions being more consistent and closer to 0 than the uncorrected version. With this particular image, it is possible to make out the individual types of damage.

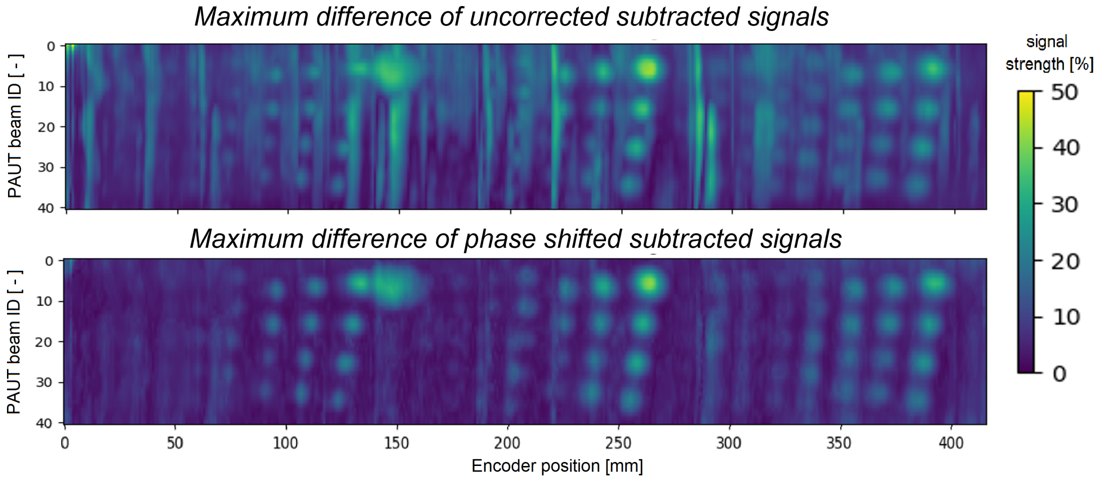


Figure 5.3: C-scan of panel #2 containing the maximum value of the subtracted signal. The top C-scan shows the resulting image without any corrections and the lower image shows the impact that the corrective actions have on the feature extraction quality. [Baseline used was NUS2203, damaged samples was NDS2203].

The phase shift correction provides the most impactful improvement in the quality of feature extraction and shows consistent results in the clarity of the image with the remainder of the lanes scanned being shown in Figure 5.4. However, as mentioned before, the variations in amplitude cause the largest variation in the subtraction of two signals. Figure 5.4 also shows the variation that can come from this method still. Notably, the centre-right image of NUS4115 subtracted with NDS4102 shows a consistent error, this means that the amplitude of these two signals was consistently different by a considerable margin. Similarly, on the left side of all the C-scans on the left, it can be seen that there is an artefact present on the left side, which is caused by the coupling agent not being properly spread at the starting location. This artefact was discussed in earlier sections as well but shows the impact it has on feature extraction. Another of these errors can be seen clearly in the two bottom C-scans on the right side, between the 100 and 150 mark on the x-axis is a clear vertical line. This line was also caused by a jerk during scanning due to the inspection setup having an area where the roller sometimes seized its movement momentarily. The results from this method show strong promise but also makes it evidently clear that not all baseline/damage scan pairs produce clean feature extractions. This leads to the next section which attempts to remedy this problem by identifying the best pairs.

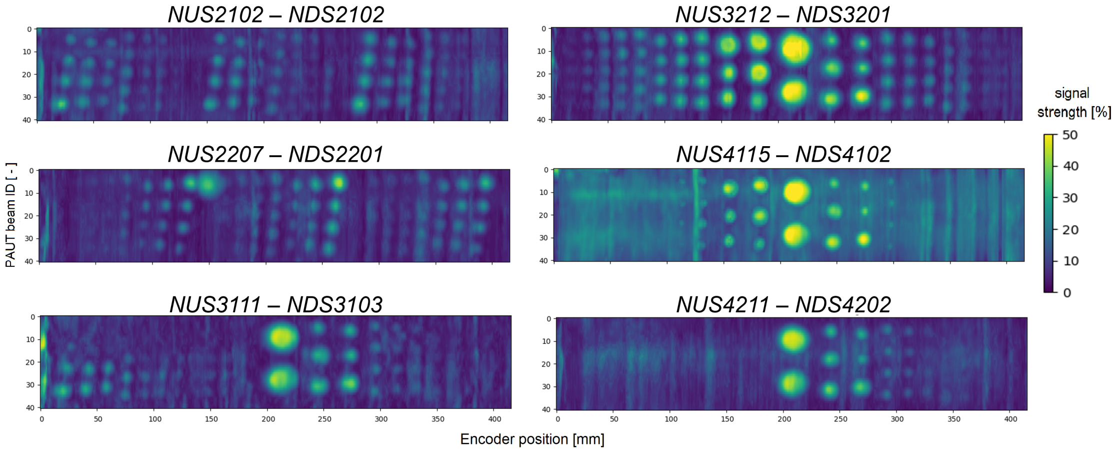


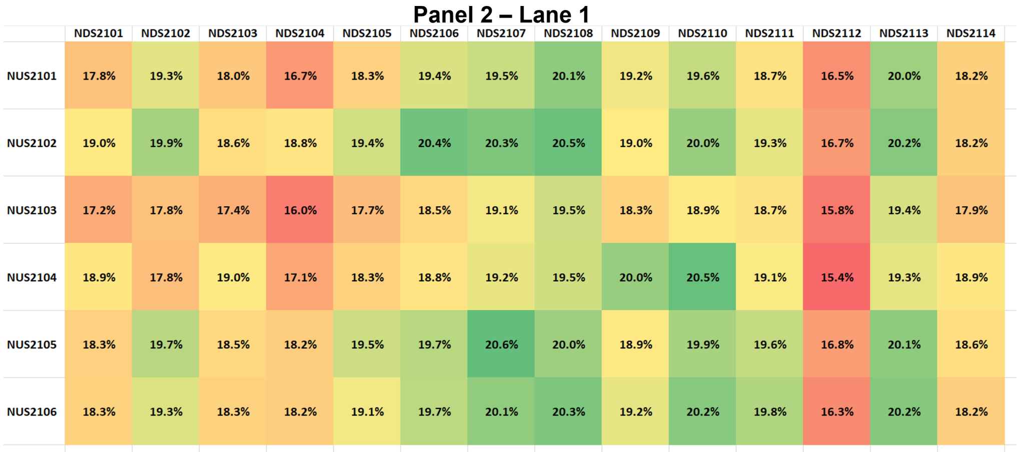
Figure 5.4: C-scan of all the panels. Shows the consistency of the method, but also the presence of artefacts caused by separate problems between the compatibility of the scans. (Recall that the first two digits of the scan number represent the panel number and lane number).

5.1.2. Best pair identification

As stated in the previous section and in subsection 4.6.2 outside of the phase misalignment exists two other errors: amplitude variation and positional variation (from encoder inaccuracies and slippage to poor starting locations). The solution proposed is to use the SSI of each C-scan of the extracted features and create a matrix showing the SSI for every combination possibility.

For each of the lanes, a separate matrix was formed to determine the best pairs. These can be seen in Figures 5.5 - 5.10. What can be seen in these figures is how there are clear rows and columns with distinctly consistent colours, showing that there are some scans which have terrible or excellent compatibility. The percentage values are the SSI score of the extracted feature compared to the baseline. The magnitude has no true value in this case, rather, it is the comparison between the values.

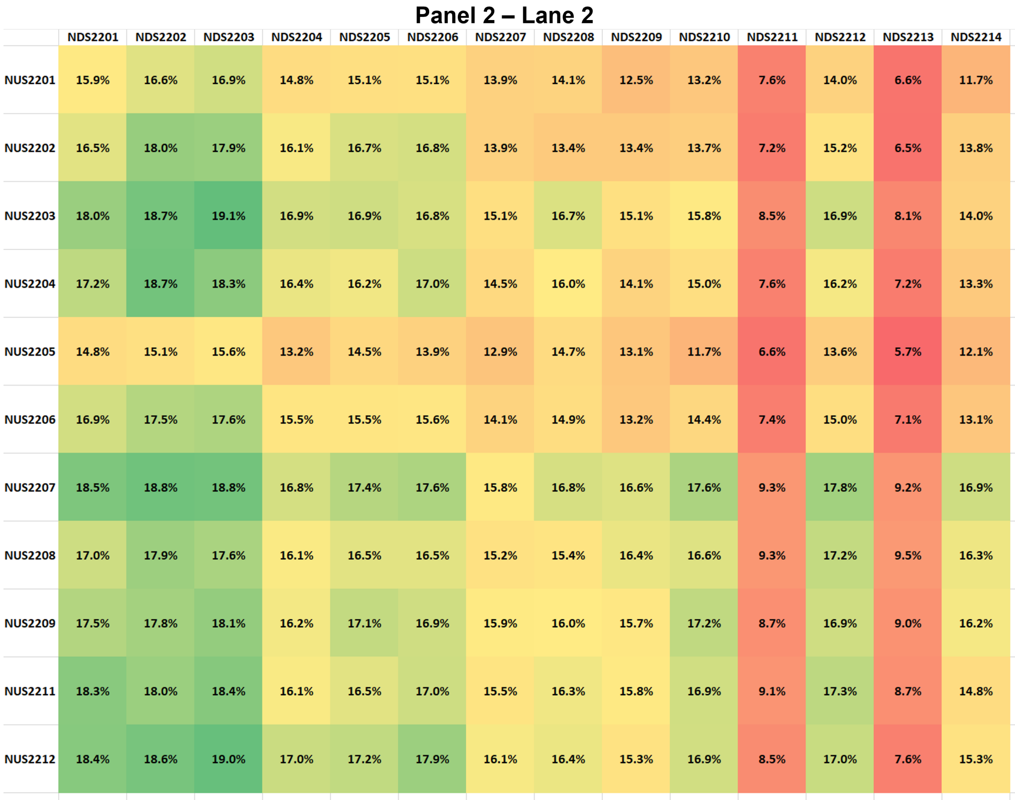
This method is still relatively effective. It has been effective in identifying noisy results, such as those which contain a lot of jerking and amplitude fluctuations, however, extracted features with a constant error such as the centre-right figure of Figure 5.4 still hold a high relative score using this method due to the consistent colour of the image. Another potential issue identified is that the baselines created do not have the same colours as these images, despite both the ground truth and the C-scan being converted to black and white images before being fed into the SSI. A recreation of the ground truth to contain the correct colour/value range could yield more consistent SSI values. Overall this method made it easier to identify good pairs but still required some filtering to find appropriate pairs that were not affected by the aforementioned problems.



Panel 2 – Lane 1

	NDS2101	NDS2102	NDS2103	NDS2104	NDS2105	NDS2106	NDS2107	NDS2108	NDS2109	NDS2110	NDS2111	NDS2112	NDS2113	NDS2114
NUS2101	17.8%	19.3%	18.0%	16.7%	18.3%	19.4%	19.5%	20.1%	19.2%	19.6%	18.7%	16.5%	20.0%	18.2%
NUS2102	19.0%	19.9%	18.6%	18.8%	19.4%	20.4%	20.3%	20.5%	19.0%	20.0%	19.3%	16.7%	20.2%	18.2%
NUS2103	17.2%	17.8%	17.4%	16.0%	17.7%	18.5%	19.1%	19.5%	18.3%	18.9%	18.7%	15.8%	19.4%	17.9%
NUS2104	18.9%	17.8%	19.0%	17.1%	18.3%	18.8%	19.2%	19.5%	20.0%	20.5%	19.1%	15.4%	19.3%	18.9%
NUS2105	18.3%	19.7%	18.5%	18.2%	19.5%	19.7%	20.6%	20.0%	18.9%	19.9%	19.6%	16.8%	20.1%	18.6%
NUS2106	18.3%	19.3%	18.3%	18.2%	19.1%	19.7%	20.1%	20.3%	19.2%	20.2%	19.8%	16.3%	20.2%	18.2%

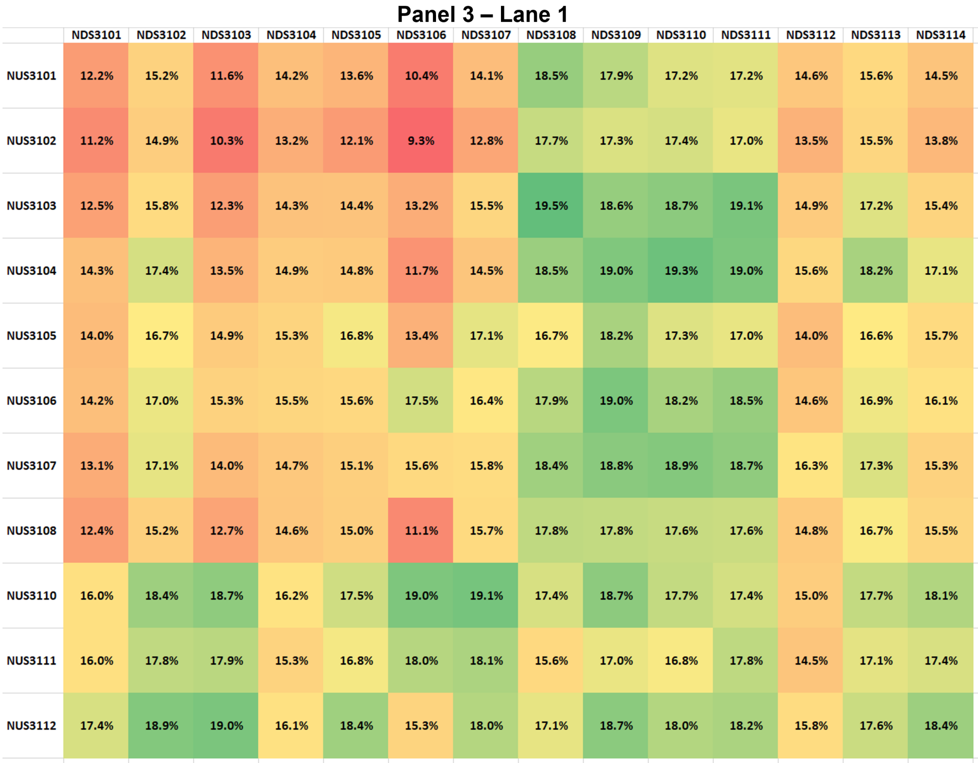
Figure 5.5: SSI of all eFlaw pairs. The vertical and horizontal axis are the baseline and damaged scans respectively.



Panel 2 – Lane 2

	NDS2201	NDS2202	NDS2203	NDS2204	NDS2205	NDS2206	NDS2207	NDS2208	NDS2209	NDS2210	NDS2211	NDS2212	NDS2213	NDS2214
NUS2201	15.9%	16.6%	16.9%	14.8%	15.1%	15.1%	13.9%	14.1%	12.5%	13.2%	7.6%	14.0%	6.6%	11.7%
NUS2202	16.5%	18.0%	17.9%	16.1%	16.7%	16.8%	13.9%	13.4%	13.4%	13.7%	7.2%	15.2%	6.5%	13.8%
NUS2203	18.0%	18.7%	19.1%	16.9%	16.9%	16.8%	15.1%	16.7%	15.1%	15.8%	8.5%	16.9%	8.1%	14.0%
NUS2204	17.2%	18.7%	18.3%	16.4%	16.2%	17.0%	14.5%	16.0%	14.1%	15.0%	7.6%	16.2%	7.2%	13.3%
NUS2205	14.8%	15.1%	15.6%	13.2%	14.5%	13.9%	12.9%	14.7%	13.1%	11.7%	6.6%	13.6%	5.7%	12.1%
NUS2206	16.9%	17.5%	17.6%	15.5%	15.5%	15.6%	14.1%	14.9%	13.2%	14.4%	7.4%	15.0%	7.1%	13.1%
NUS2207	18.5%	18.8%	18.8%	16.8%	17.4%	17.6%	15.8%	16.8%	16.6%	17.6%	9.3%	17.8%	9.2%	16.9%
NUS2208	17.0%	17.9%	17.6%	16.1%	16.5%	16.5%	15.2%	15.4%	16.4%	16.6%	9.3%	17.2%	9.5%	16.3%
NUS2209	17.5%	17.8%	18.1%	16.2%	17.1%	16.9%	15.9%	16.0%	15.7%	17.2%	8.7%	16.9%	9.0%	16.2%
NUS2211	18.3%	18.0%	18.4%	16.1%	16.5%	17.0%	15.5%	16.3%	15.8%	16.9%	9.1%	17.3%	8.7%	14.8%
NUS2212	18.4%	18.6%	19.0%	17.0%	17.2%	17.9%	16.1%	16.4%	15.3%	16.9%	8.5%	17.0%	7.6%	15.3%

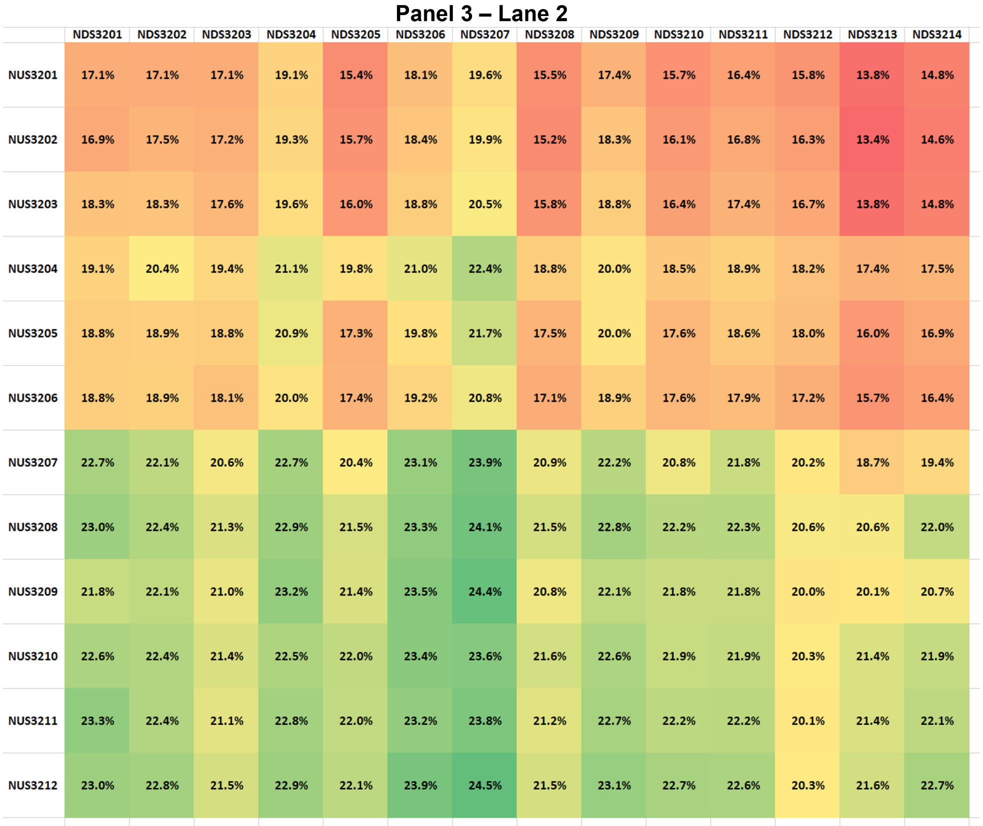
Figure 5.6: SSI of all eFlaw pairs. The vertical and horizontal axis are the baseline and damaged scans respectively.



Panel 3 – Lane 1

	NDS3101	NDS3102	NDS3103	NDS3104	NDS3105	NDS3106	NDS3107	NDS3108	NDS3109	NDS3110	NDS3111	NDS3112	NDS3113	NDS3114
NUS3101	12.2%	15.2%	11.6%	14.2%	13.6%	10.4%	14.1%	18.5%	17.9%	17.2%	17.2%	14.6%	15.6%	14.5%
NUS3102	11.2%	14.9%	10.3%	13.2%	12.1%	9.3%	12.8%	17.7%	17.3%	17.4%	17.0%	13.5%	15.5%	13.8%
NUS3103	12.5%	15.8%	12.3%	14.3%	14.4%	13.2%	15.5%	19.5%	18.6%	18.7%	19.1%	14.9%	17.2%	15.4%
NUS3104	14.3%	17.4%	13.5%	14.9%	14.8%	11.7%	14.5%	18.5%	19.0%	19.3%	19.0%	15.6%	18.2%	17.1%
NUS3105	14.0%	16.7%	14.9%	15.3%	16.8%	13.4%	17.1%	16.7%	18.2%	17.3%	17.0%	14.0%	16.6%	15.7%
NUS3106	14.2%	17.0%	15.3%	15.5%	15.6%	17.5%	16.4%	17.9%	19.0%	18.2%	18.5%	14.6%	16.9%	16.1%
NUS3107	13.1%	17.1%	14.0%	14.7%	15.1%	15.6%	15.8%	18.4%	18.8%	18.9%	18.7%	16.3%	17.3%	15.3%
NUS3108	12.4%	15.2%	12.7%	14.6%	15.0%	11.1%	15.7%	17.8%	17.8%	17.6%	17.6%	14.8%	16.7%	15.5%
NUS3110	16.0%	18.4%	18.7%	16.2%	17.5%	19.0%	19.1%	17.4%	18.7%	17.7%	17.4%	15.0%	17.7%	18.1%
NUS3111	16.0%	17.8%	17.9%	15.3%	16.8%	18.0%	18.1%	15.6%	17.0%	16.8%	17.8%	14.5%	17.1%	17.4%
NUS3112	17.4%	18.9%	19.0%	16.1%	18.4%	15.3%	18.0%	17.1%	18.7%	18.0%	18.2%	15.8%	17.6%	18.4%

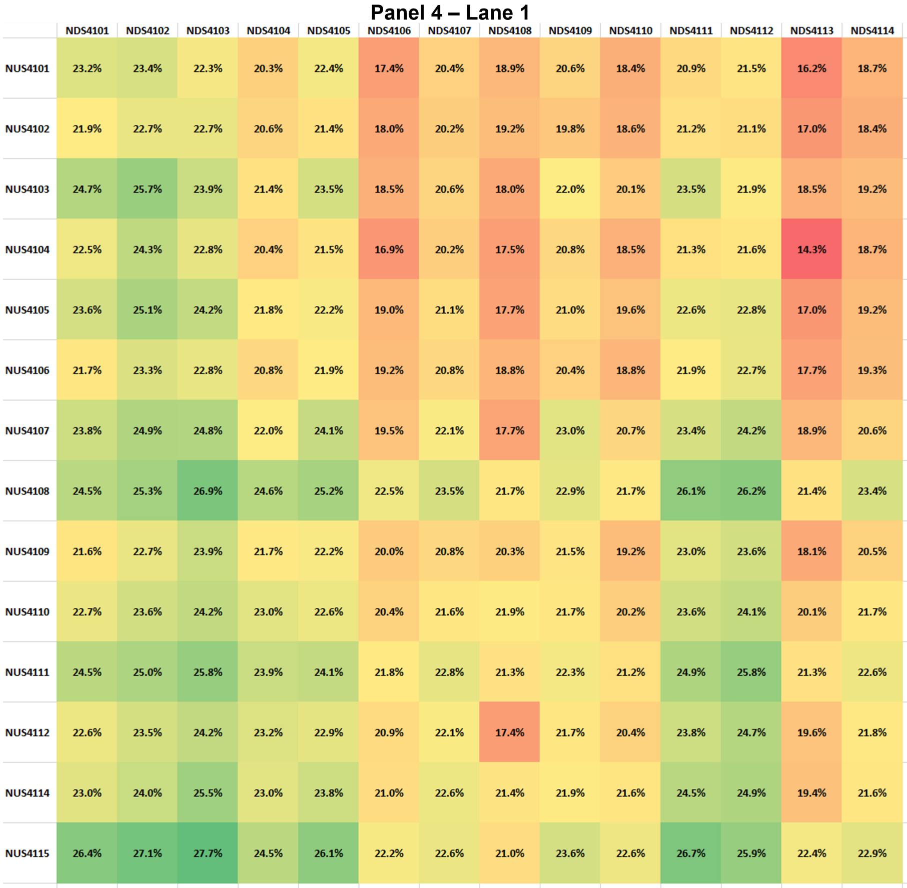
Figure 5.7: SSI of all eFlaw pairs. The vertical and horizontal axis are the baseline and damaged scans respectively.



Panel 3 – Lane 2

	NDS3201	NDS3202	NDS3203	NDS3204	NDS3205	NDS3206	NDS3207	NDS3208	NDS3209	NDS3210	NDS3211	NDS3212	NDS3213	NDS3214
NUS3201	17.1%	17.1%	17.1%	19.1%	15.4%	18.1%	19.6%	15.5%	17.4%	15.7%	16.4%	15.8%	13.8%	14.8%
NUS3202	16.9%	17.5%	17.2%	19.3%	15.7%	18.4%	19.9%	15.2%	18.3%	16.1%	16.8%	16.3%	13.4%	14.6%
NUS3203	18.3%	18.3%	17.6%	19.6%	16.0%	18.8%	20.5%	15.8%	18.8%	16.4%	17.4%	16.7%	13.8%	14.8%
NUS3204	19.1%	20.4%	19.4%	21.1%	19.8%	21.0%	22.4%	18.8%	20.0%	18.5%	18.9%	18.2%	17.4%	17.5%
NUS3205	18.8%	18.9%	18.8%	20.9%	17.3%	19.8%	21.7%	17.5%	20.0%	17.6%	18.6%	18.0%	16.0%	16.9%
NUS3206	18.8%	18.9%	18.1%	20.0%	17.4%	19.2%	20.8%	17.1%	18.9%	17.6%	17.9%	17.2%	15.7%	16.4%
NUS3207	22.7%	22.1%	20.6%	22.7%	20.4%	23.1%	23.9%	20.9%	22.2%	20.8%	21.8%	20.2%	18.7%	19.4%
NUS3208	23.0%	22.4%	21.3%	22.9%	21.5%	23.3%	24.1%	21.5%	22.8%	22.2%	22.3%	20.6%	20.6%	22.0%
NUS3209	21.8%	22.1%	21.0%	23.2%	21.4%	23.5%	24.4%	20.8%	22.1%	21.8%	21.8%	20.0%	20.1%	20.7%
NUS3210	22.6%	22.4%	21.4%	22.5%	22.0%	23.4%	23.6%	21.6%	22.6%	21.9%	21.9%	20.3%	21.4%	21.9%
NUS3211	23.3%	22.4%	21.1%	22.8%	22.0%	23.2%	23.8%	21.2%	22.7%	22.2%	22.2%	20.1%	21.4%	22.1%
NUS3212	23.0%	22.8%	21.5%	22.9%	22.1%	23.9%	24.5%	21.5%	23.1%	22.7%	22.6%	20.3%	21.6%	22.7%

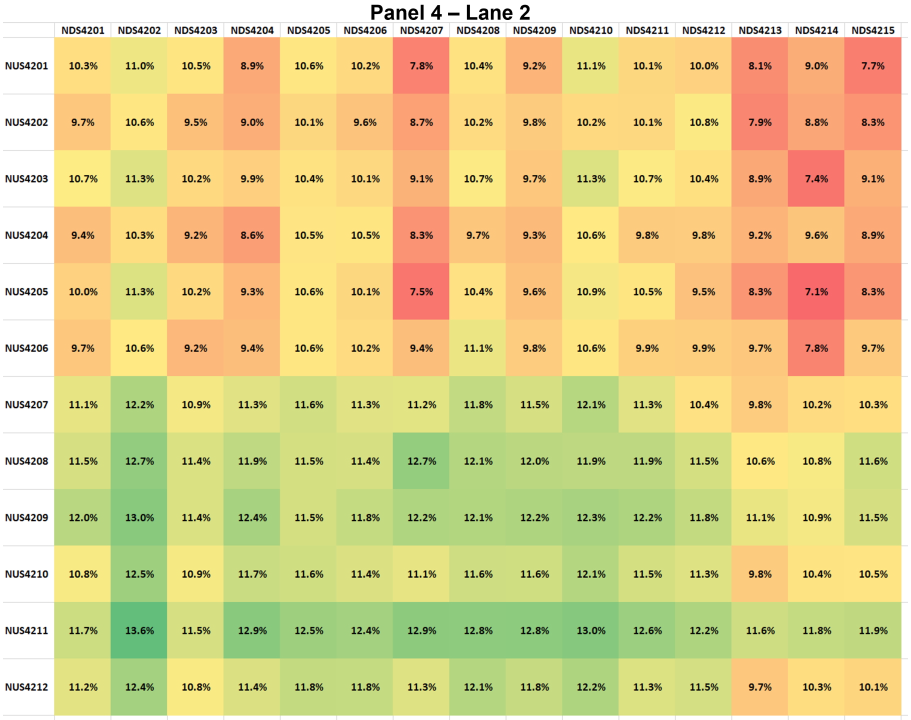
Figure 5.8: SSI of all eFlaw pairs. The vertical and horizontal axis are the baseline and damaged scans respectively.



Panel 4 – Lane 1

	NDS4101	NDS4102	NDS4103	NDS4104	NDS4105	NDS4106	NDS4107	NDS4108	NDS4109	NDS4110	NDS4111	NDS4112	NDS4113	NDS4114
NUS4101	23.2%	23.4%	22.3%	20.3%	22.4%	17.4%	20.4%	18.9%	20.6%	18.4%	20.9%	21.5%	16.2%	18.7%
NUS4102	21.9%	22.7%	22.7%	20.6%	21.4%	18.0%	20.2%	19.2%	19.8%	18.6%	21.2%	21.1%	17.0%	18.4%
NUS4103	24.7%	25.7%	23.9%	21.4%	23.5%	18.5%	20.6%	18.0%	22.0%	20.1%	23.5%	21.9%	18.5%	19.2%
NUS4104	22.5%	24.3%	22.8%	20.4%	21.5%	16.9%	20.2%	17.5%	20.8%	18.5%	21.3%	21.6%	14.3%	18.7%
NUS4105	23.6%	25.1%	24.2%	21.8%	22.2%	19.0%	21.1%	17.7%	21.0%	19.6%	22.6%	22.8%	17.0%	19.2%
NUS4106	21.7%	23.3%	22.8%	20.8%	21.9%	19.2%	20.8%	18.8%	20.4%	18.8%	21.9%	22.7%	17.7%	19.3%
NUS4107	23.8%	24.9%	24.8%	22.0%	24.1%	19.5%	22.1%	17.7%	23.0%	20.7%	23.4%	24.2%	18.9%	20.6%
NUS4108	24.5%	25.3%	26.9%	24.6%	25.2%	22.5%	23.5%	21.7%	22.9%	21.7%	26.1%	26.2%	21.4%	23.4%
NUS4109	21.6%	22.7%	23.9%	21.7%	22.2%	20.0%	20.8%	20.3%	21.5%	19.2%	23.0%	23.6%	18.1%	20.5%
NUS4110	22.7%	23.6%	24.2%	23.0%	22.6%	20.4%	21.6%	21.9%	21.7%	20.2%	23.6%	24.1%	20.1%	21.7%
NUS4111	24.5%	25.0%	25.8%	23.9%	24.1%	21.8%	22.8%	21.3%	22.3%	21.2%	24.9%	25.8%	21.3%	22.6%
NUS4112	22.6%	23.5%	24.2%	23.2%	22.9%	20.9%	22.1%	17.4%	21.7%	20.4%	23.8%	24.7%	19.6%	21.8%
NUS4114	23.0%	24.0%	25.5%	23.0%	23.8%	21.0%	22.6%	21.4%	21.9%	21.6%	24.5%	24.9%	19.4%	21.6%
NUS4115	26.4%	27.1%	27.7%	24.5%	26.1%	22.2%	22.6%	21.0%	23.6%	22.6%	26.7%	25.9%	22.4%	22.9%

Figure 5.9: SSI of all eFlaw pairs. The vertical and horizontal axis are the baseline and damaged scans respectively.



Panel 4 – Lane 2

	NDS4201	NDS4202	NDS4203	NDS4204	NDS4205	NDS4206	NDS4207	NDS4208	NDS4209	NDS4210	NDS4211	NDS4212	NDS4213	NDS4214	NDS4215
NUS4201	10.3%	11.0%	10.5%	8.9%	10.6%	10.2%	7.8%	10.4%	9.2%	11.1%	10.1%	10.0%	8.1%	9.0%	7.7%
NUS4202	9.7%	10.6%	9.5%	9.0%	10.1%	9.6%	8.7%	10.2%	9.8%	10.2%	10.1%	10.8%	7.9%	8.8%	8.3%
NUS4203	10.7%	11.3%	10.2%	9.9%	10.4%	10.1%	9.1%	10.7%	9.7%	11.3%	10.7%	10.4%	8.9%	7.4%	9.1%
NUS4204	9.4%	10.3%	9.2%	8.6%	10.5%	10.5%	8.3%	9.7%	9.3%	10.6%	9.8%	9.8%	9.2%	9.6%	8.9%
NUS4205	10.0%	11.3%	10.2%	9.3%	10.6%	10.1%	7.5%	10.4%	9.6%	10.9%	10.5%	9.5%	8.3%	7.1%	8.3%
NUS4206	9.7%	10.6%	9.2%	9.4%	10.6%	10.2%	9.4%	11.1%	9.8%	10.6%	9.9%	9.9%	9.7%	7.8%	9.7%
NUS4207	11.1%	12.2%	10.9%	11.3%	11.6%	11.3%	11.2%	11.8%	11.5%	12.1%	11.3%	10.4%	9.8%	10.2%	10.3%
NUS4208	11.5%	12.7%	11.4%	11.9%	11.5%	11.4%	12.7%	12.1%	12.0%	11.9%	11.9%	11.5%	10.6%	10.8%	11.6%
NUS4209	12.0%	13.0%	11.4%	12.4%	11.5%	11.8%	12.2%	12.1%	12.2%	12.3%	12.2%	11.8%	11.1%	10.9%	11.5%
NUS4210	10.8%	12.5%	10.9%	11.7%	11.6%	11.4%	11.1%	11.6%	11.6%	12.1%	11.5%	11.3%	9.8%	10.4%	10.5%
NUS4211	11.7%	13.6%	11.5%	12.9%	12.5%	12.4%	12.9%	12.8%	12.8%	13.0%	12.6%	12.2%	11.6%	11.8%	11.9%
NUS4212	11.2%	12.4%	10.8%	11.4%	11.8%	11.8%	11.3%	12.1%	11.8%	12.2%	11.3%	11.5%	9.7%	10.3%	10.1%

Figure 5.10: SSI of all eFlaw pairs. The vertical and horizontal axis are the baseline and damaged scans respectively.

5.1.3. Feature reintroduction

With damage being extracted, and viable undamaged/damaged pairs identified for the creation of clear extracted features, it is possible to reintroduce them into new baselines to see their ability as an augmentation technique.

Adhering to Equation 4.2 it is possible to create new ultrasonic scans with damaged features present in them. The first illustration of this is from the perspective of A-scans and can be seen in Figure 5.11. This reintroduction illustrates the method's ability to introduce distinct features, such as the distinct features found along the x-axis at 200 and 800. But notably it also reduces the amplitude of the back wall reflection at x = 400. However, there is some misalignment in the reintroduction and this can be seen at x = 600 where more jagged curves can be observed.

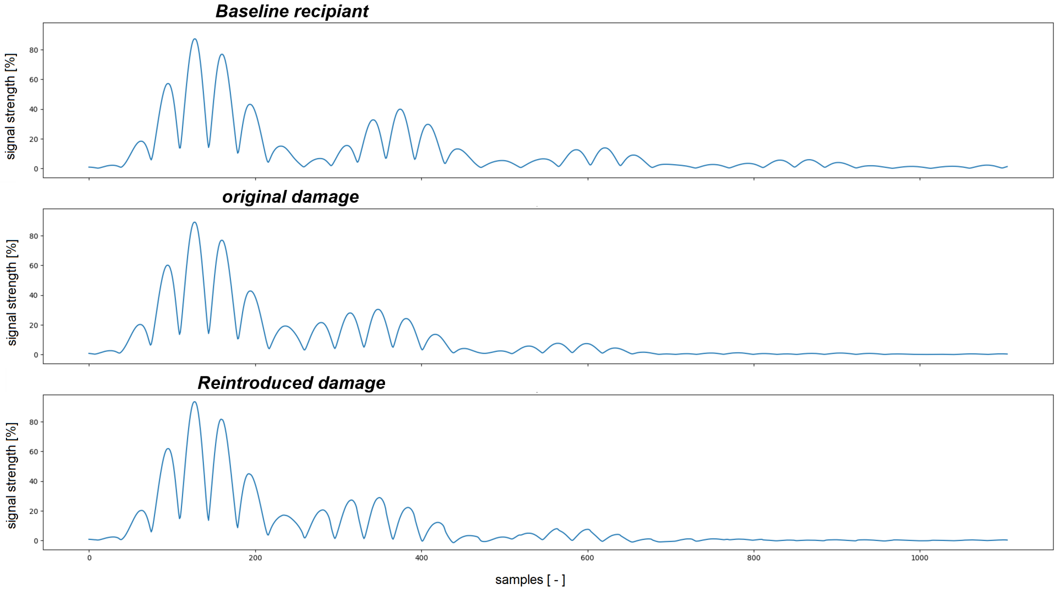


Figure 5.11: Top: Baseline A-scan absent of damaged features; Middle: A-scan containing damaged features; Bottom: A-scan of extracted damage features introduced into the baseline A-scan.

When expanding the view to B-scans it is possible to view the feature in two dimensions. Figure 5.12 shows the comparison between two B-scans. One of the original damage and the other with the extracted feature was introduced into a separate scan. There are some differences such as the intensity of the incident wave and even in the shape of the feature in certain areas (such as the back wall reflection).

Augmenting the damage by translating the flaw to different locations proved to be challenging. In Figure 5.12 it can be seen how there is a slight slope to the back wall reflection. This, as previously discussed, was caused by the vacuum infusion process which left the two faces not fully parallel. As a result, the extracted flaws could not be shifted left or right (within the B-scan reference frame), because the component of the feature will not align with the baseline's back-wall reflection. This variation of thickness, however, is far more consistent along the length of the panel which means that the flaws could be shifted along the length.

Figure 5.13 shows the eFlaw concept's ability to translate the extracted damage features along the length of the panel. What can be seen is how the systemic error from the roller probe stays in place while the distinct damage features have moved around.

This ability provides the potential for damage to be introduced anywhere along the length of the panel either for training purposes or for further data augmentation.

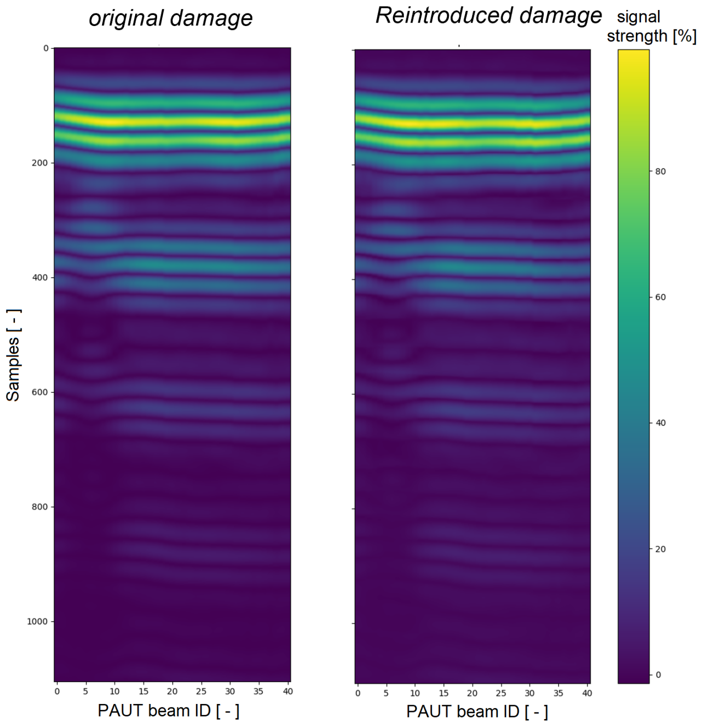


Figure 5.12: Right: Original B-scan with damage; Left: Baseline B-scan with eFlaw introduced into it.

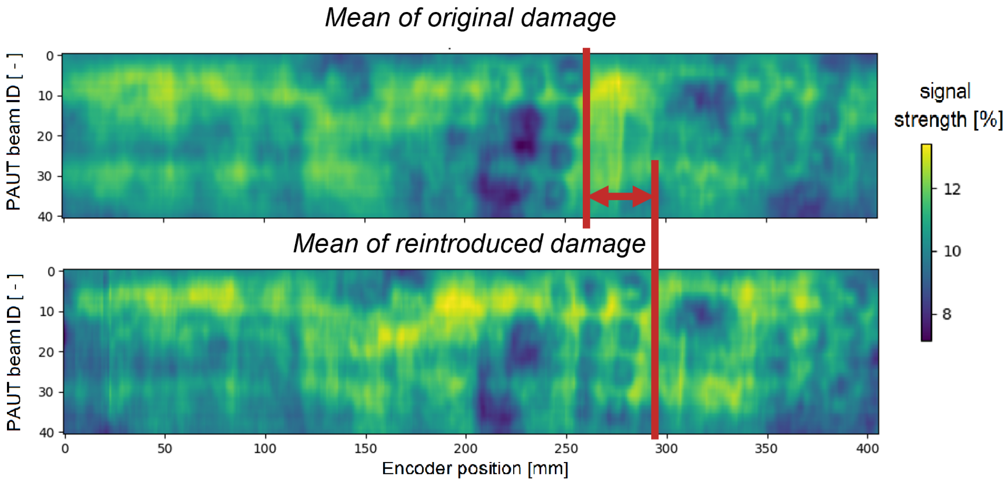
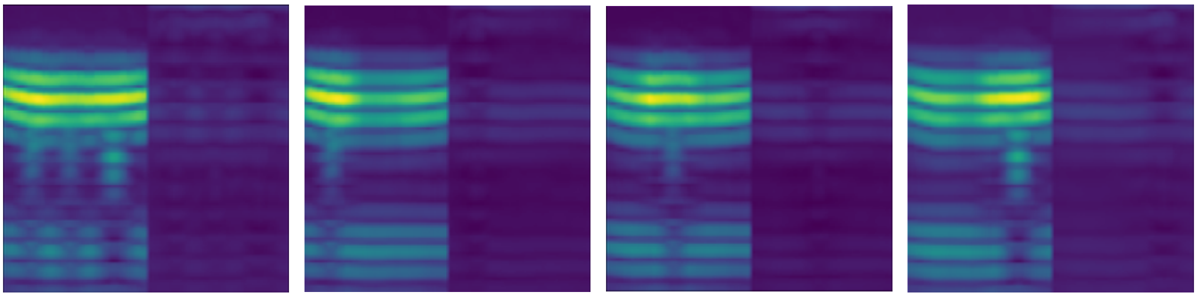


Figure 5.13: Figure showing the eFlaws ability to be translated along the length of the panels

5.1.4. Feature isolation

The translation is not the only type of augmentation that can be applied to the extracted flaws. Another method discussed was the ability to isolate specific features before reintroduction, opening the possibility for more variations of the same B-scan. Figure 5.14 shows the results of this process. Using the method described it was possible to isolate the three damages present in Figure 5.14a. An important feature is how bright the region is above the damage in the incident wave. It was not fully implemented during this thesis but according to Equation 4.4 there would be 6 unique B-scans that could result from the original scan, totalling 7 B-scans.



(a) original scan with all damage (b) 10mm FBH isolated (c) 10mm FBH isolated (d) 15mm FBH isolated

Figure 5.14: Scan from panel 4 lane 1, the eFlaw concept here was applied using NUS4115 and the damage was extracted from NDS4101 and reintroduced into NUS4115 to minimize noise.

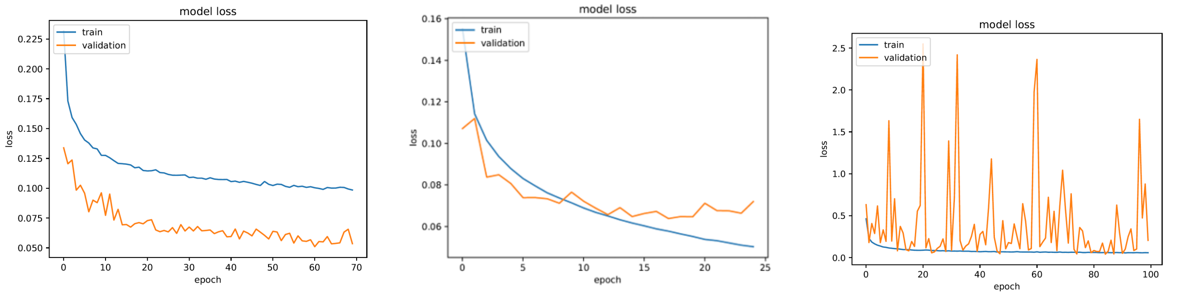
Despite not being fully implemented, it still demonstrated the successful ability to create an additional 25,000 images for the data set. This method of augmentation has proven effective for composite materials as well, with a few caveats surrounding the uniformity of thickness of the sample and the structure being a plate-like structure. The potential for this method, as discussed by its founders [33], open the potential for more easily generalised data and has now been proven for composite materials. The implementation of this concept will now be seen in how it impacts the performance of training a model.

5.2. Model performance

This section will discuss the results of the final models. Including a discussion on the training of the models and the impact which the eFlaw concept had on the performance. The resulting models and their performance on the test data set and finally the explainability of the models being assessed will be discussed.

5.2.1. Model training

Training the models proved challenging with the initial training dataset. Despite the extensive amount of augmentation being applied, while excluding the eFlaw augmentation, the models were not able to converge properly. Instead, both models either exhibited signs of overfitting or underfitting and in addition signs of either the validation data or training data being too easy to predict/train. Figure 5.15 shows some of these issues which were encountered during training. The end results were models which performed poorly or had memorized the training data.



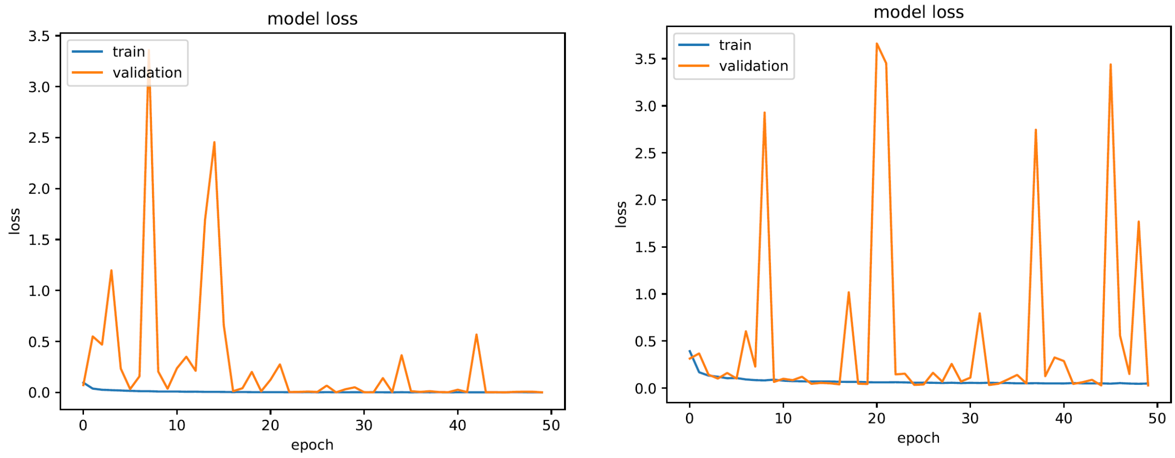
(a) Custom VGG model's learning curve showing that the validation data is easier to predict and resulting in underfitting.

(b) Transfer learning model showed signs of overfitting and potentially unrepresentative validation data.

(c) Transfer learning model with unrepresentative validation data.

Figure 5.15: Three plots showing the challenges faced during training.

After these results, it was apparent that the dataset, as it was, was incapable of training a CNN effectively for use in damage detection. This was observed to be a result of data which was not sufficiently generalized for training. The eFlaw concept is used to create the additional 25,000 images described earlier (nowhere near the full utilization of the concept). What followed were learning curves which converged, and yielded immediate improved results. Figure 5.16 shows the two learning curves for the custom VGG model and the transfer learned model. What can be immediately seen is that the custom model converges more cleanly than that of the transfer learning model, which still appears to exhibit some trouble with the validation data and raises questions on the validity of the transfer learning model.



(a) Custom VGG models training curve after 50 epochs

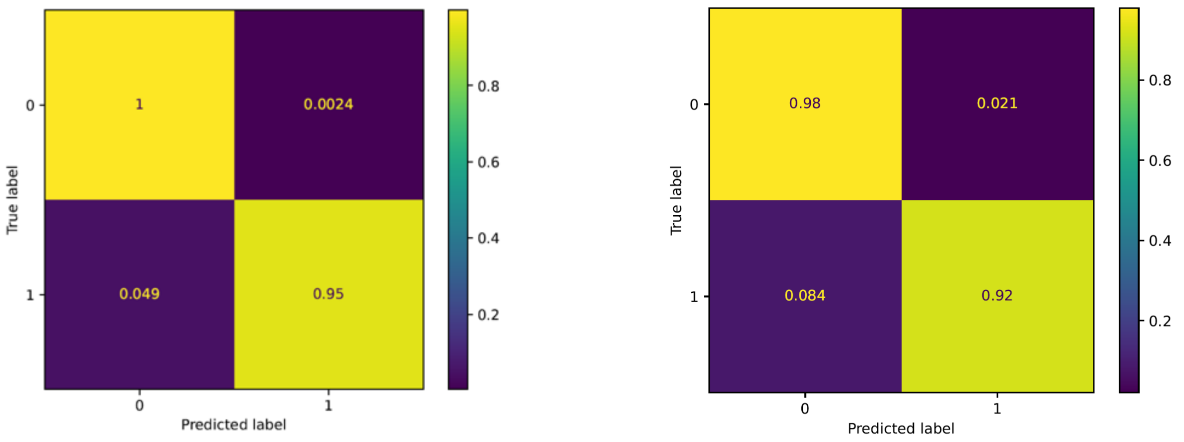
(b) Transfer learning model training curve after 50 epochs

Figure 5.16: Final learning curves for both models.

It would have been ideal to run these models for more epochs, however, the increase in data size increased the computational time required per epoch where it was no longer feasible to train models for longer periods as the energy consumption had begun to become too costly. This point of the model's training was deemed acceptable to proceed with the project, and with the use of the performance metrics and the explainability, it will come to show whether these models needed more training or not.

5.2.2. Model performance

The first step to evaluate the performance of the model was to generate the confusion matrices for the models using the testing data. These matrices can be seen in Figure 5.17, where each cell corresponds with the same cell denoted in Figure 2.11. What can be seen is that the true negative and true positive rate for each matrix is quite high, with the custom VGG model having the best results. These values are a percentage horizontally comparing the true and false positives/negatives and should add up to 100%.



(a) Confusion matrix for the custom VGG model.

(b) Confusion matrix for the transfer learned model

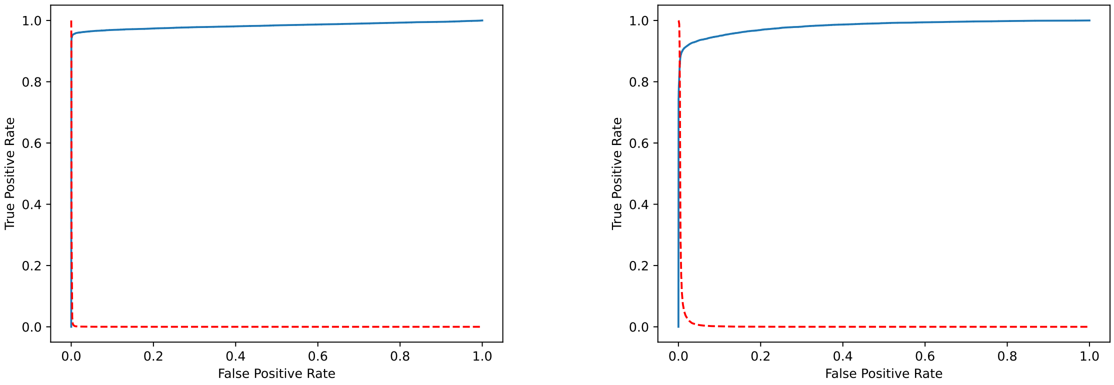
Figure 5.17: Confusion matrices of the final trained models. 1 denotes damage and 0 denotes undamaged

To give a clearer understanding of these values the various metrics derived from these matrices are tabulated below. Table 5.1 shows more clearly how the custom VGG model outperforms the transfer learned model. The recall most notably being the lowest value for both models, which is brought down by both matrices having relatively high False negative values. The integrity of the True negative and False negative is not fully certain, since the labelling of the infusion defect in panel 4 lane 2 may not have been labelled correctly, resulting in potentially mislabelled scans as damaged when in actuality they are undamaged.

Table 5.1: Table showing the various performance metrics derived from the confusion matrix.

	Accuracy	**Recall**	**Specificity**	**Precision**	**F1-score**
Custom VGG model	97.4%	95.3%	99.8%	99.8%	97.5%
Transfer learned model	94.8%	92.1%	97.8%	97.9%	94.9%

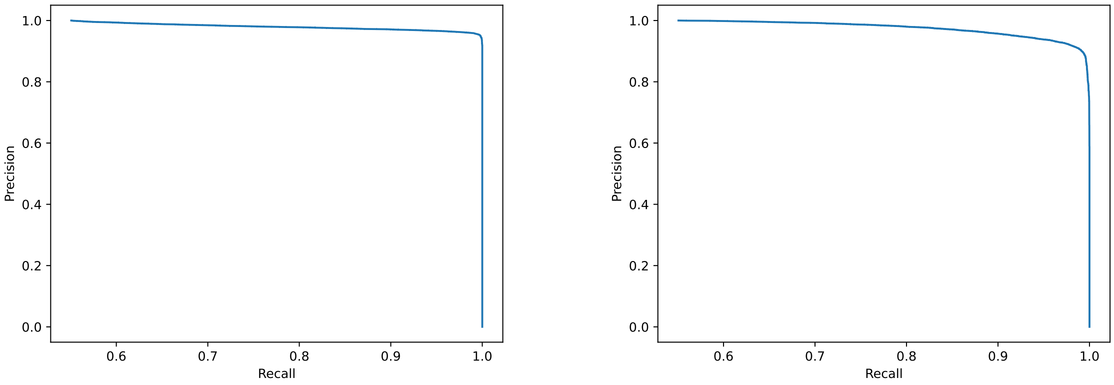
Another area of interest is the sensitivity of the model to thresholds and whether the two distributions of damaged and undamaged predictions overlap significantly. To visualise this, the ROC curve is used and can be seen in Figure 5.18. In these curves, two lines can be seen, a red dashed line and a blue solid line. The blue solid line shows the ROC curve for each of the models, and the red dashed line shows the respective threshold value for each False positive rate. What can be seen from these curves is the fast and sharper response of the custom VGG model compared to the transfer learned model which has a more gradual slope and indicates more overlap between the damage and undamaged distributions. What this means for the results is that the custom VGG model will always perform better than the transfer learned model with the correct thresholds set (in this case a threshold of 0.03). The same observations can be seen for the Precision/Recall plots which illustrate a similar relationship and can be seen in Figure 5.19 where once again the VGG model is almost completely square and the other model exhibits less sharpness and lower area under the curve.



(a) ROC curve of the custom VGG model

(b) ROC curve of the transfer learned model

Figure 5.18: ROC curves for both models. The blue line is the resulting ROC curve, and the red dashed line denotes the different threshold values for each false positive rate.



(a) Precision/Recall curve of the custom VGG model

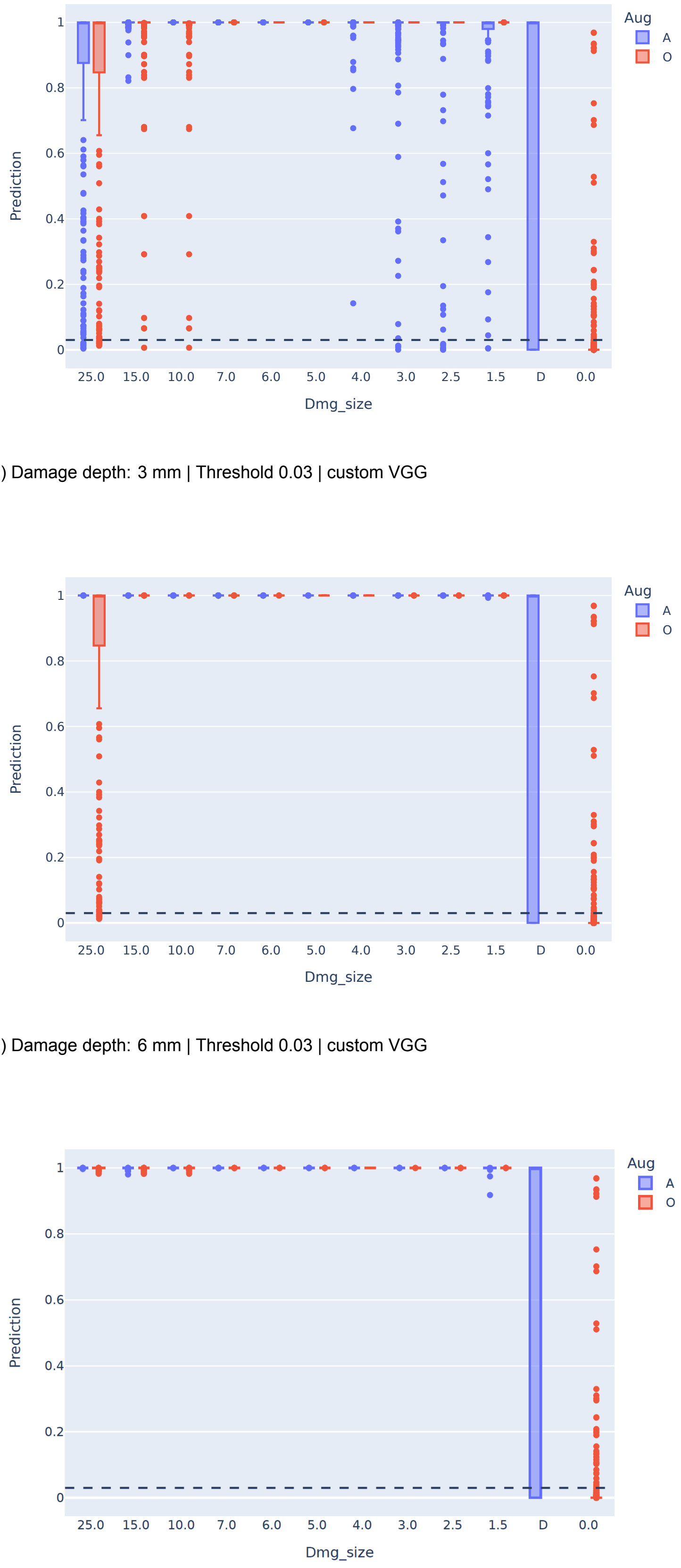
(b) Precision/Recall curve of the transfer learning model

Figure 5.19: Precision/Recall curves for both models.

5.2.3. Size and depth prediction performance

From the previous results, it was clear that the custom VGG model outperformed the transfer learning model. To finalise which model would be used for the final tool it was also decided to plot the prediction performance of the model for different depths and damage sizes as box plots. These box plots provided a way to evaluate the prediction capabilities of the augmented eFlaw data to the original unaltered data. Effectively being able to peer into the damaged class and identify how it performed on different damage sizes and depths.

Figure 5.20 and Figure 5.21 show these box plots for each of the initial damage depths. Since the model was trained on damage ranging from 10 mm to 2.5 mm (as 1.5 mm was excluded due to a lack of confidence in the data labelling) the two key areas to pay attention to are the 25mm damage and the 1.5 mm damage. For the custom VGG model (Figure 5.20), it can be seen that damage at a depth of 3 mm provides some challenge to the model with a very small number of outliers falling below the threshold for damage sizes on either end of the tail (both small and large). For the smaller damage, there are some misses on the augmented damage, while the original damage has a 100% detection rate. The 8.5 mm depth damage has the best performance with not a single miss in the data with damage. The 6 mm depth damage oddly enough only had misses from the original damage at the 25mm size (this is odd since it detected 1.5 mm damage without a problem). As for the undamaged inputs, there were also consistently good predictions, with a portion of the outliers falling above the threshold. What this means is that over 99.68% of each damage-size and undamaged sample is being classified correctly. The likely culprit for the reduced accuracy is the manufacturing defect in panel 4 lane 2, which had to be labelled by hand where the B-scans should be considered damaged. For the most part, the median and upper limits of the box plot are at 100%, but the first quartile also goes to 0. It is known for a fact that some of this data is mislabeled, and should be revised. The transfer learning model (Figure 5.21), shows poorer performance. With a wider interquartile range across all depths for all of the smaller damage. Where the custom VGG model had the median at 100% for all cases, the transfer learning model has it varying across the entire range, but still above the threshold. The worst of these was the box plot of the 1.5 mm damage at 8.5 mm depth whose median was below the threshold meaning a significant number of those damages were being labelled as undamaged. Similarly, for the infusion defect, it performed worse than the other model with a lower median, and it also had a larger distribution of undamaged samples above the threshold, with some even at the 100% certainty mark (unlike the other model which only reached 98%). Overall from these plots, it is clear that the custom VGG model has more consistent and polar predictions as opposed to the transfer learning model which exhibits more overlap between its damaged/ undamaged predictions. Additionally, the transfer learning model is consistently poor with smaller damages with the box plots often crossing the threshold. As a result, the custom VGG model became the model of interest for explainability topics.

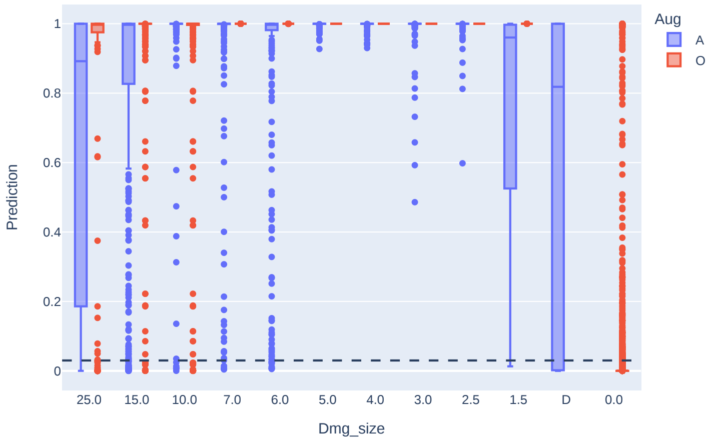


(a) Damage depth: 3 mm | Threshold 0.03 | custom VGG

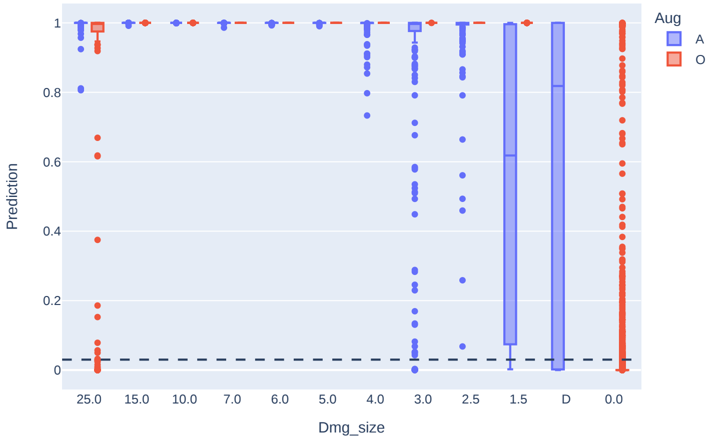
(b) Damage depth: 6 mm | Threshold 0.03 | custom VGG

(c) Damage depth: 8.5 mm | Threshold 0.03 | custom VGG

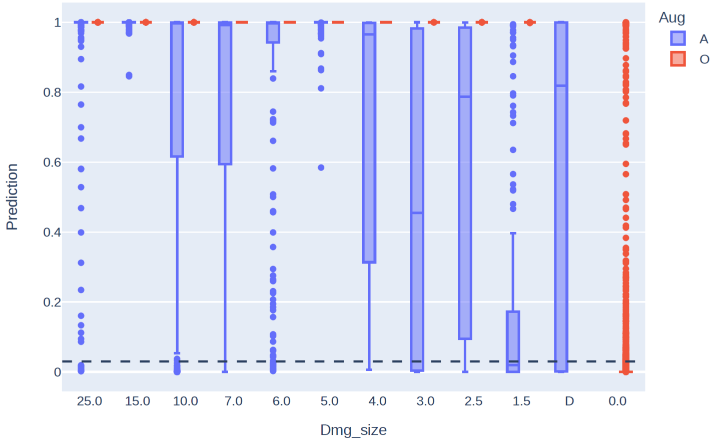
Figure 5.20: Box and whisker plots of the different damage size prediction for the custom VGG model. Each dot represents a prediction that falls outside ±1.5 times the interquartile range (outside of 99.36% of the distribution). Blue: For the isolated eFlaw data, Red: Original unaltered data. The ’D’ represents the predictions on the region of the infusion error made in panel 4 lane 2, and 0.0 represents no damage (hence no augmented data)



(a) Damage depth: 3 mm | Threshold 0.03 | Transfer learning model



(b) Damage depth: 6 mm | Threshold 0.03 | Transfer learning model



(c) Damage depth: 8.5 mm | Threshold 0.03 | Transfer learning model

Figure 5.21: Box and whisker plots of the different damage size prediction for the transfer learning model. Each dot represents a prediction that falls outside ± 1.5 times the interquartile range (outside of 99.36% of the distribution). Blue: For the isolated eFlaw data, Red: Original unaltered data. The 'D' represents the predictions on the region of the infusion error made in panel 4 lane 2, and 0.0 represents no damage (hence no augmented data)

5.2.4. Explainability

From the previous sections, it was observed that the custom VGG model performed significantly better than the transfer learned model. From this conclusion, it was decided to only investigate the custom VGG model moving forward. This section aims to evaluate the different explainable methods, and which ones offer the best insight into how the decisions are being made.

The first explainable method applied was LIME and provided mixed results. Figure 5.22 illustrates the method in use. What can be seen is that the method found the upper portions of the images as the most relevant areas for damage detection. meaning that the incident wave and first echo provide the model with sufficient information to make a decision. Across different depths and damage sizes it was found to have a similar shape. This result only provided a coarse perspective of the relevant areas for the model, but no precise information on locations could be derived.

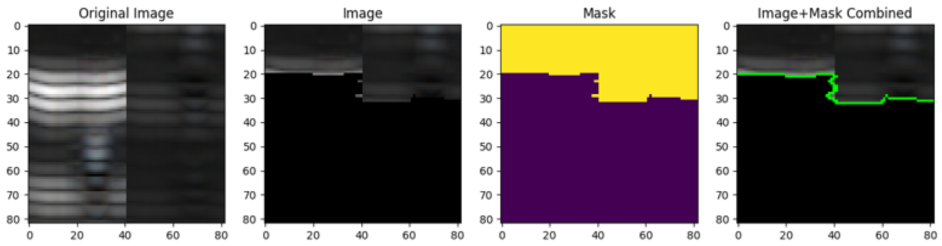
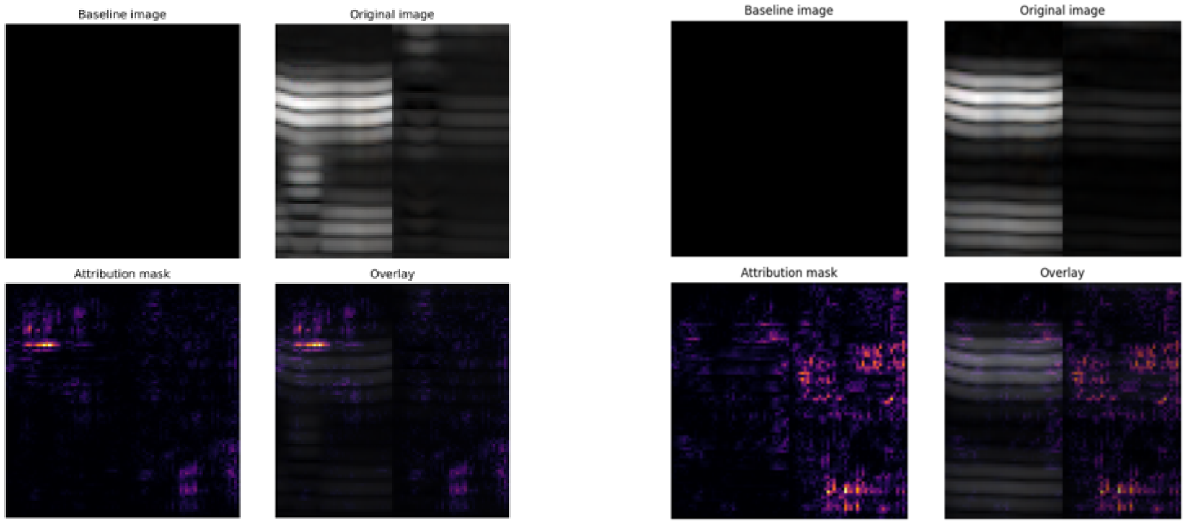


Figure 5.22: LIME implementation on an augmented eFlaw image with 25mm damage size.

The next method was the integrated gradients method, which yielded almost incomprehensible results. Figure 5.23 shows the inconsistency of these results. In the best case the method was able to highlight the rough coulumns where the damage was present (in many cases it was unable to even do this) and can be seen in the image on the right. In the worst case, there is remotely no information to be derived (such as the image on the right).



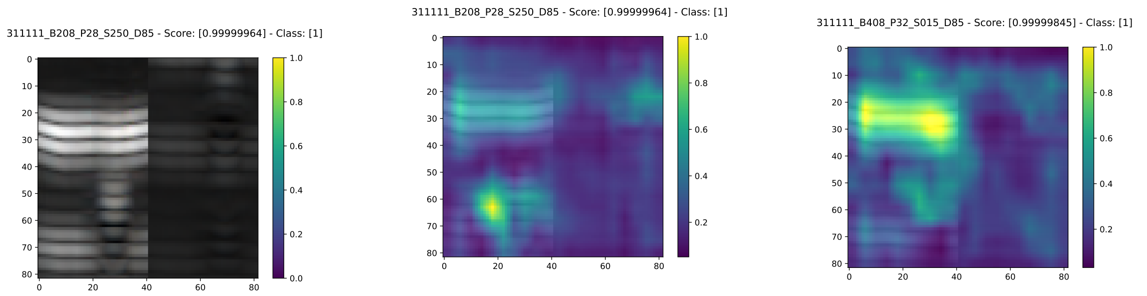
(a) Integrated gradients method applied to an isolated 25mm damage | Damage depth: 8.5 mm | Prediction: 0.9999939 - Truth: 1 | file: 311111_B213_P9_S250_D85

(b) Integrated gradients method applied to isolated 1.5 mm damage | Damage depth: 8.5 mm | Prediction: 0.9999901 - Truth: 1 | file: 311111_B394_P4_S015_D85

Figure 5.23: Integrated gradients application using a black backdrop for the integration and illustrating the inconsistency of the method.

To contrast the performance of this method was the GradCam method, which showed promising results. The GradCam method was applied to the second last VGG block because the final block had too coarse a resolution to effectively work for smaller damage. The results are shown in Figure 5.24 where a localised explanation can be observed. For the 25mm damage, it was seen that it attempted to locate not only the column but the depth of the damage as shown in Figure 5.24b. The features it highlights are the surrounding features of the damage.

For the remaining damage, the top part of the images was mostly used for inference. This is shown in Figure 5.24c where the incident wave is the area highlighted. There is a lot of activity which roughly correlates to the locations of the damage and provides some confidence that the model is truly identifying damage and not some arbitrary features or memorization.



(a) Image of the 25mm damage

(b) GradCam result overlayed on the 25mm damage image. True damage positon: x=28 (56mm)

(c) GradCam result for damage of 1.5mm in size. True damage position: x=32 (64mm)

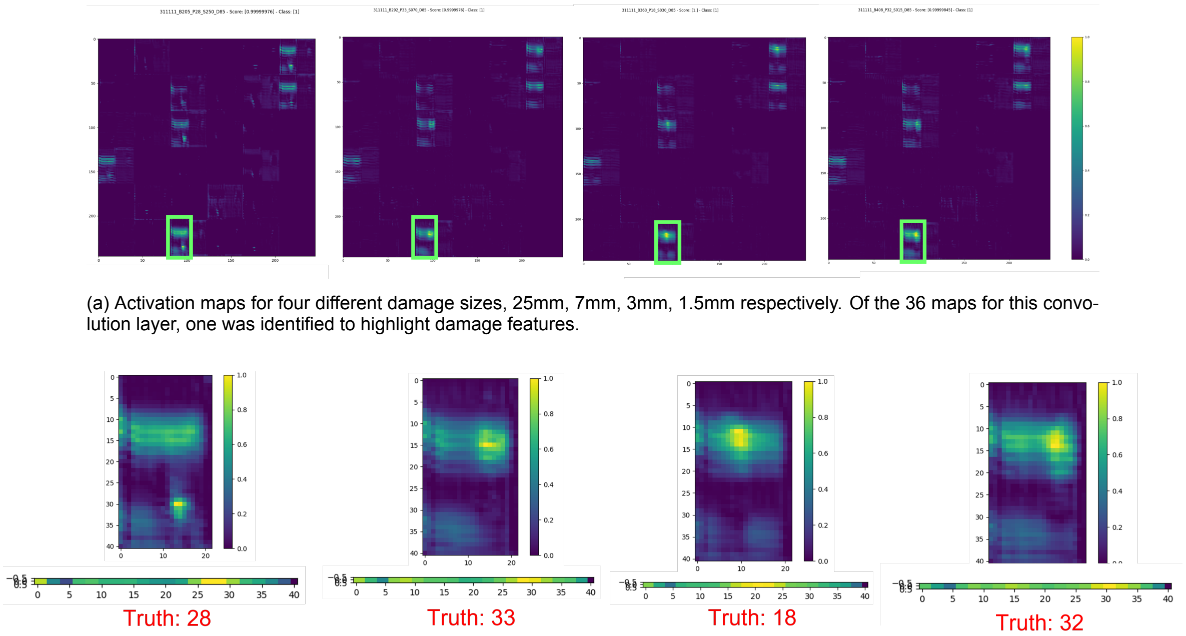
Figure 5.24: GradCam implementation on two types of damage to illustrate the different methods it highlights the damage.

The final and most effective method is the inspection of the activation maps of the model. This allows each VGG block to be peered into and each trained convolution map to be viewed. The inspection of these activation maps finds that the final convolution layer of the second VGG block has a feature map which highlights specific features consistently. This can be seen in Figure 5.25a where one-half of the feature map is highlighted in green. This feature map appears to highlight features similar to how the Gradcam method's results appeared but with more localised precision. To further process this data the feature map in question was extracted for each of the damage cases used and is shown in Figure 5.25b. What can be seen from these extracted maps is how this feature map is activated in the columns where the damage is present. Because this feature map is extracted from the second VGG block it means its spatial dimensions are halved, making the resolution slightly coarser and reducing the accuracy which can be derived from the localisation and limiting this method to the initial VGG blocks where the resolution is still high enough.

As a further method of processing the extracted map, it was desired to reduce the maps down into a 1D vector where it would be easier to evaluate the observed location of the damage compared to the ground truth. Using the resulting 1D vectors, the centre of the maximum pixels was calculated and assumed to be the central position of the damage. From here Table 5.2 was generated to evaluate the general accuracy of the method. From this brief evaluation, a maximum deviation of 7mm was observed, which means that an NDT inspector would have to scan an area of a few centimetres at most to locate the damage if this highlighted region was used as the starting point. An observation from all of these explainable methods was that the echo data did not appear to play a significant role in the detection of damage. Since the custom model proved to be more effective, the input shape could have been changed, and more resolution preserved of the image.

Table 5.2: Observed damage locations using the vectors maximum location. These values are multiplied by the PAUT element width to calculate the true location in millimetres.

	Damage size			
	25mm	**7mm**	**3mm**	**1.5mm**
Observed location [px]	27.5	29.5	20.5	30.5
Ground truth [px]	28	33	18	32
Deviation [px]	-0.5	3.5	2.5	-1.5
Deviation [mm]	-1	7	5	-3



(a) Activation maps for four different damage sizes, 25mm, 7mm, 3mm, 1.5mm respectively. Of the 36 maps for this convolution layer, one was identified to highlight damage features.

(b) Extracted feature map which highlights the respective damaged features. Taking the average of each map column produces a vector which provides insight into the damage location (below each map). The true location is listed in red.

Figure 5.25: Activation map method employed. It allowed for a common feature map to be identified which identified the damaged features.

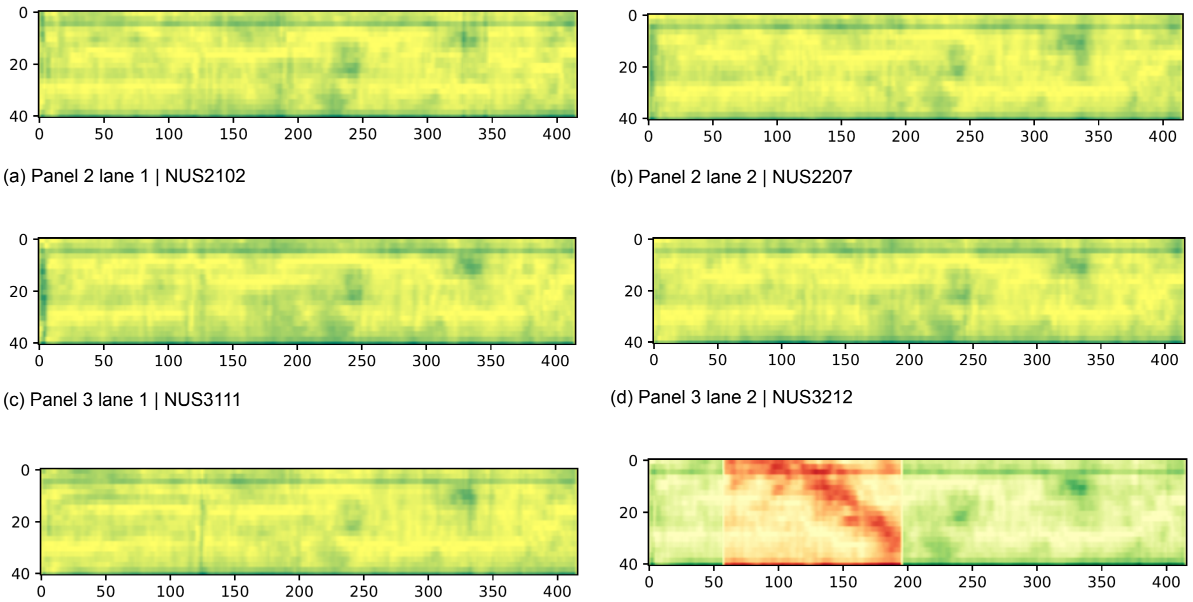
5.2.5. Human interface (through the F-scan)

The section on explainability revealed that the model was identifying some relevant features which allowed damage to be localised. This resulted in the creation of the Feature-scan, or F-scan. As opposed to using gates to view a C-scan, a technician can refer to the activation maps and select the feature map of choice to display the C-scan view. In other words, it becomes a dynamic feature-driven gate which highlights them along the length of the panel. The implementation of this principle can be seen in Figure 5.26 and Figure 5.27 where the method will highlight a B-scan as green for no damage detected and red for damage detected. The vectorised feature map from the previous second is used to create that slice of the C-scan.

Figure 5.26 shows fully undamaged panels for each of the lanes with the only exception being the final lane (Figure 5.26f) which had the manufacturing defect present in it. The F-scan is also able to highlighting the defect which resembles the defect shown in Figure 4.6 perfectly. This is a positive sign as it shows the model is generalised enough to detect these features which do not resemble a FBH.

Figure 5.27 shows the model attempting to locate damage within the damaged panels. Due to the high density of damage, most of these panels are almost completely red except for the starting points (where a few false positive occurs in Figure 5.27a). The feature map produces some clear visualisations allowing for some localisation. Figure 5.27c has a feature right before x=200 showing no damage, when consulting the ground truths based on the CAD data, this green area aligns with the undamaged regions of the panels.

One observation is that despite there being 12mm of space between each damage these regions are still being highlighted as damage. A likely culprit is that the damage is likely not sufficiently far away enough from each other and there are some interactions with the PAUT beam which results in certain features which the model has learnt to identify as the presence of damage. The overall performance of this method is positive as it shows consistency and detail and most importantly communicates it in an effective format as close to the standard format used by inspectors.



(a) Panel 2 lane 1 | NUS2102

(b) Panel 2 lane 2 | NUS2207

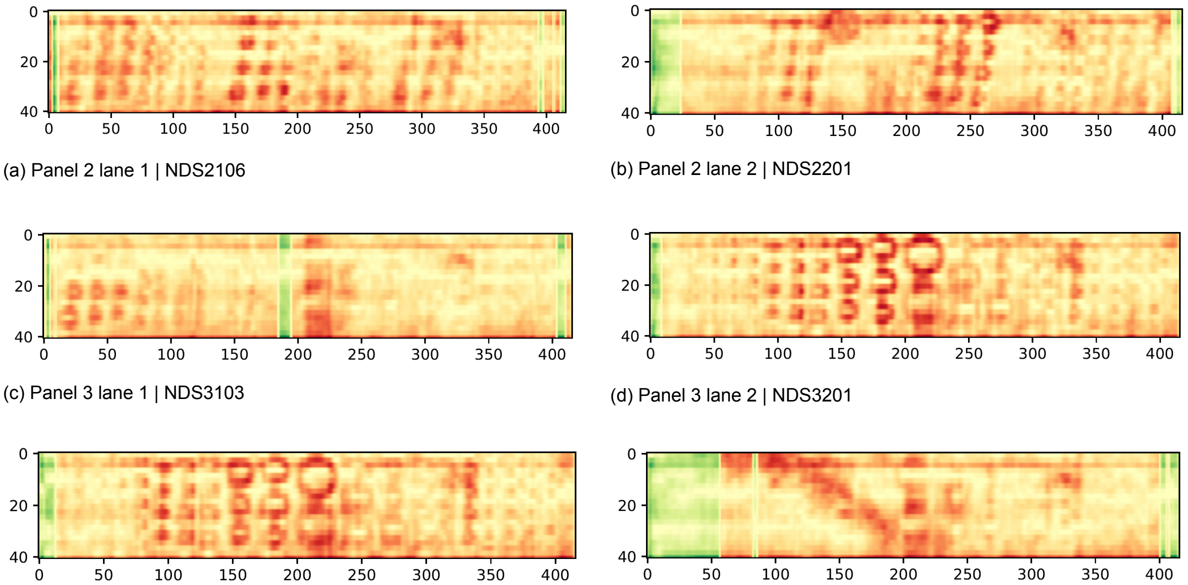
(c) Panel 3 lane 1 | NUS3111

(d) Panel 3 lane 2 | NUS3212

(e) Panel 4 lane 1 | NUS4115

(f) Panel 4 lane 2 | NUS4211 | Note the F-scan identifying the infusion defect illustrated in Figure 4.6.

Figure 5.26: F-scans of all the panels and lanes. These scans were all of the undamaged panels before the introduction of the FBHs.



(a) Panel 2 lane 1 | NDS2106

(b) Panel 2 lane 2 | NDS2201

(c) Panel 3 lane 1 | NDS3103

(d) Panel 3 lane 2 | NDS3201

(e) Panel 4 lane 1 | NDS4101

(f) Panel 4 lane 2 | NDS4202

Figure 5.27: F-scans of all the panels and lanes. These scans were all of the damaged panels with the FBHs present.

5.3. Reflection on certifying AI for aerospace

This section will discuss the EASA AI guidelines, this thesis, and how they relate to one another. This will include best practices, identified problems and a review of the research questions.

5.3.1. Guidelines and the project

Upon reflection on this thesis, a lot was learnt about both the maturity of the guidelines, current literature and the development of such a tool. With a focus on the three main domains associated with AI (learning assurance, AI explainability and AI safety risk mitigation), it was clear that there were differences in the level of their maturity. Learning assurance by far is one of the more mature domains within research. The reason for this is that all AI topics require data sets, and they require that the model is able to converge onto a solution from this data. A lot of these principles are generally model agnostic and allow for a lot of interdisciplinary development on the topic. What this means is that the barrier of entry has been lowered as there is both a lot of literature but also tools and software which allow for the quick and easy application of various methods. The guidelines as a result are also able to provide more details for the expectations and deliverables for learning assurance. They cover the end-to-end development process in what they refer to as the W-shaped process [2]. For this project, the iterative learning assurance process was used to great success.

A supplemental document within the guidelines is another EASA document titled ”Concepts of Design Assurance for Neural Networks” [112]. This document served as a building block for both the guidelines and roadmap, and provided their own proposal for developing these tools, similarly emphasising the three domains already mentioned. Data quality was a large part of both of these documents, where a set of requirements were outlined. How this project performed on those requirements is summarised below:

- *Accuracy:* Achieved through CNC machined damage, rigid inspection setup, CAD data to accurately relocate the data. For all of the damage ranging from 25 mm to 2.5 mm, it is certain the PAUT probe captured the correct slices as they were identifiable through the B-scan data. For the 1.5 mm damage, there was some uncertainty which is why it was omitted from the training data (a decision which did not appear to negatively impact the performance in detecting damage that small). Additionally, the infusion defect was also poorly labelled and has impacted the results of the confusion matrices, however, the Figure 5.20 was able to show the classification performance of each damage individually.

- *Resolution:* The data was kept to the Nyquist frequency of the captured data. As much data was conserved, and after the results, it was clear that the echo data was not as relevant as initially suspected.

- Assurance level: This was not exactly considered within the scope of this thesis, but minor attempts were made to ensure that data was not corrupted or changed throughout the process. Mainly by a trivial checksum using the filesize, and by never moving the files around once secured on the personal computer. Other literature proposes more in-depth methods to ensure data integrity.

- *Traceability:* This was also a major focus for this project. Every scan and every augmented data can be traced back to the original data used to produce it. This was attributed to the naming convention applied to the data which encoded various attributes including original files, damage size and even location. This proved to be a very useful consideration when debugging, or identifying bad data. An instance of bad data was the infusion defect. Traceability allowed the detailed plots of Figure 5.20 and 5.21 to show the true performance of the actual damage features.

- *Timeliness:* This requirement is relevant for model drift as input data might change over time, resulting in the model no longer being able to detect the correct features. This was not considered during the project, but is definitely a consideration for generalisation as the different environmental conditions may affect the model performance (but this is speculation and requires investigation).

- *Completeness:* section 4.4 discusses the considerations made here to include a proper distribution of damage within the data to properly perform.

5.3.2. Challenges with qualification

One topic on traceability and transparency which was missing from this thesis was a quantitative analysis of the frequency of error present in the data, as well as a list of discrepancies that would have been valuable to document in this thesis.

Validating the model has been acknowledged by the guidelines as one of the more difficult tasks given the complexity of NN based models. The method used in this thesis was coverage-based white-box testing, which looked at some agnostic methods which considered the input and outputs for performance, and some explainable methods to peer into the ”black box”. However, the ideal situation for the guidelines is to have a formal verification method which has formally derived bounds for the model’s performance. In this case, there was little to contribute, as the only formal method that was intended to be used was the PoD curve. However, the model’s performance exceeded expectations, to the point where it was no longer possible to generate a viable PoD curve. This exposed two major issues with this process, the first is that conventional qualification techniques have come into question on their applicability. The second is that the selected distribution was insufficient to properly test the limits of this model. A problem not easily fixed since the damage size was already reaching the limits of conventional machining.

Finally, these guidelines focus on the operational monitoring of these models, which is arguably the area with the highest potential for learning and furthering the understanding of the model’s capabilities. This, however, is not addressed within the scope of this thesis and should be considered when planning to introduce a data-driven tool into the field. Considering the guidelines recognise the lower risk in maintenance and training tools than flying tools, it furthers the argument for an ultrasonic data-driven tool to be the first case study to further develop these guidelines.

5.4. Research questions

The research question aimed to be answered in this thesis was as follows:

What would the development process for a qualifiable maintenance machine learning model, which is capable of classifying damage in ultrasonic scans of aerospace-grade composite panels, look like?

And its three subquestions:

1. What are the impacts of the eFlaw concept on the process?
2. What are the types of features the model identifies?
3. How can the output of the model be interpreted and used in a practical ultrasonic non-destructive testing application, and what are the implications for safety and reliability in the aerospace industry?

5.4.1. Main research question

In order to answer the main research question, it must be divided into several subtopics. Beginning with the development process, section 3.3 as well as all the content in the experimental setup and methodologies provides a strong foundation for the development of such a tool. The thesis acts as a starting point for maintenance-related tools and contributes more than is provided in the appendix of the guidelines. This process, however, is lacking objective methods to guarantee the performance of the model. To investigate this point further, it may be advisable to conduct a conventional probability of detection study using a human in the loop and evaluate the results from there.

As for the capability of the tool to classify damage, the outcome of this thesis was that a data-driven tool in the form of a convolution neural network was developed with the capability of detecting damage as small as 1.5 mm ($\frac{1}{4}\lambda$). With reference to the qualifiability, in section 4.1 it was discussed that the goal would be to design a tool which could provide level 1-B automation which was defined as a tool which provides human cognitive assistance in decision and action selection. The outcome of this was that a tool was developed which could assist an inspector in identifying areas of interest. This was achieved with reference to the list made by Gunning et al. [83] discussed in section 2.7.

The list has been restated below, with a discussion on how the developed tool relates to these points.

1. **Explanation only assists user performance if the task is difficult enough to require explanation:** For the purpose of this tool the explanations will only become relevant for damages below 2.5 mm where the visibility on the B-scans becomes difficult.
2. **Users preferred and trusted a system that offered an explanation over a system which only provided decisions:** This point was taken into consideration since a C-scan with the relevant B-scans highlighted would not provide much insight into what was being selected and why. Developing the F-scan was necessary to meet this point.
3. **If the interpretation of the explanation requires a high cognitive load it may hinder user performance:** To avoid this being a problem it was chosen to keep the explanation in the same input space of the model. This allows the explanation to be easily compared to with the input B-scans. The verdict of this approach and using activation maps with higher resolutions offer the least cognitive load to interpret the results. Additionally, the activation maps were further simplified to produce the F-scan, which offers another perspective which is comparable with the C-scan of the data.
4. **An incorrect explanation is considered extremely valuable for edge-case investigations:** This tool already provided some explanations which did not align with the general methods of interpreting the results. The explainable methods showed that the incident wave had the largest amount of attention from the model, and an investigation into whether the incident wave is all that is required to identify damage could be interesting for the computational and data requirements for such a tool, but may also pave the way for other methods to interpret ultrasonic data.

The guidelines discuss the thresholds between level 1A and level 1B, where level 1A tools are only those which augment the information presented to the end user and level 1B tools support the decision-making process by interpreting the results as well [58]. Since the output of this tool also classifies data as damaged or undamaged it aids towards the decision-making of the inspector. The threshold for this tool to become a level 2A tool is based on its ability to automatically make a decision, this would likely require the tool to characterise the damage, its size, location and whether it is a structural concern, before submitting a report for evaluation. This threshold to 2A is far more complex and would require a great deal of thought into the type of model and how explainability would be integrated with such a system. The current tool however already provides a significant amount of utility in its ability to assist an inspector, and could likely be used as a development point into level 2A. As a result, this tool meets the requirements to be theoretically classified as a level 1B tool, but must still be tested in practice with a qualitative study to be fully validated.

Regarding model performance, it is believed that this model has demonstrated its ability to work reliably for GFRP panels of a thickness between 10 mm and 11 mm. However, more research must be made on the effect that large damage has on the model's performance, as well as smaller damage (though this will become challenging as the size begins to reach the limits of the damage manufacturing process). Similarly, the capture rate of the encoder will likely have to decrease to reliably detect smaller damage.

5.4.2. Sub-questions

It's now possible to move to the three sub-questions, which each have an interest in various aspects of the development process. Each one of these questions will be discussed below:

1. **What are the impacts of the eFlaw concept on the process?** The eFlaw concept proved to be extremely successful as an augmentation technique. It offers a method which allows for the more affordable creation of test samples since various types of damage can be placed in closer proximity. It also allows for the isolation of specific features, which allows for the creation of numerous combinations of features. Translation of these features is slightly more problematic as this concept only works for uniform thickness plates as of now, where a method will have to be derived to further isolate damage features of each component (incidence wave, feature reflection, back-wall reflection, etc.). In the case of this model, the use of eFlaws was the inflexion point where the model was suddenly able to converge towards a model with high performance. As it is the first application to composite materials, this concept can be considered a success with a lot of potential for future research.

2. **What are the types of features the model identifies?** From the results it was clear that the model was able to identify defects of different shapes and sizes, even being able to identify voids caused by a manufacturing error. Using explainable methods it was determined that the model mostly only required the incident wave to determine whether there was damage present or not. This means that the model has been able to identify features within the near field of the beam and make an accurate informed decision. Only for significantly large damages of 25mm did the explainable methods begin highlighting features in the depth. Finally, it was also observed that the shadow reflections on the right side of each input image had little to no activation in any of the activation maps. As a result, it is suggested that the incident and back wall reflection region is sufficient. It also means that the model's input dimensions can be adjusted and a slightly higher resolution will be possible.

3. **How can the output of the model be interpreted and used in a practical ultrasonic NDT application, and what are the implications for safety and reliability in the aerospace industry?** Using the final model it was possible to create a tool which would be able to process the ultrasonic data and produce a new form of the C-scan referred to as an F-scan. This scan, in the same view as the C-scan, highlights a B-scan as either green for undamaged, or red for damaged. The process also used a specific feature map from the model's convolution layers to provide some spatial context on what areas along the width of the B-scan it believes there to be damage. A tool like this allows inspectors to quickly identify regions of interest, and focus on identifying the damage within that region.

   This tool however is very sensitive and can detect damage which exceeds the capabilities of human inspectors, this trait is a double-edged sword. Starting with the strength, this means the tool is so sensitive that it increases the detection envelope for the low-frequency wheel probe. This is rather significant as this tool could slowly begin reducing the trade-off inspectors have to make between the frequency of their probe and the thickness of the material. Meaning that the inspection of thick composites could become far more accessible, opening the door for thicker damage-tolerant primary composite structures. The downside of this trait however is that it might be too sensitive for the acceptable damage sizes which can be present in the structure, this would mean that the inspector would have irrelevant damage flagged to their device and waste their time since it poses no risk at the moment. To remedy this, additional thresholds could be applied to potentially reduce the sensitivity to align more with acceptable defects. In a level-3 system, this increased sensitivity could allow for extremely sensitive detection systems which are capable of evaluating the damage using low-frequency probes.

   As for safety and reliability, it is still an open topic to develop some discrete metrics to assure the reliability of this model. If the model labels a region as undamaged, it could prove catastrophic if the inspector misses the damage. Low reliability would also reduce the trust that the inspector has in the tool, impacting the effectiveness of the tool and the strength of the human interface with the tool. However, continuous safety analysis of a tool such as the one in this thesis could provide valuable data for retraining, but also on how to develop evaluation methods to assess the performance of the tool during its operational phase.

The results from this thesis have contributed to the industrialisation of AI for NDT in aerospace. A model was developed which offers exceptional performance. The results and process here have the potential to provide the guidelines with a more detailed case study for the appendix. There are many other points and topics within the guidelines that have to be considered before this tool can be remotely considered for operational use, however, this thesis takes the first steps to what is believed to be the best point of entry for AI tools used directly on aircraft structures.

6 Conclusion

This thesis investigated the development of a level 1-B qualifiable CNN for in-situ ultrasonic damage classification in aerospace composite structures. The scope of this topic was vast and required a lot of interdisciplinary knowledge. It involved a clear understanding of the drafted guidelines by EASA and the development of a list of clear deliverables that should be expected of any research topic working towards the same goal.

A large contribution of this thesis has been the experimental setup and methodology, where it first introduces the concept of operation. The concept of operation discussed the vision and goals of the tool that would be developed and offers insight to the reader into what they can expect from this tool. This was followed by the inspection setup, which included the equipment and the jigs and fixtures used. It would go on to describe the four panels that were set out to be manufactured, with the end result being only three usable panels for this study. Regardless of these limitations, a clear distribution of damage was selected with an emphasis to have the distribution be right-skewed.

This thesis demonstrated the first successful application of the eFlaw concept in composite materials. The principle proved extremely successful in its ability to augment data, translate data, and isolate data. This method ultimately became the augmentation technique that allowed the model to converge. This is attributed to its ability to increase the data set size by 24% and also provide significantly more generalised data. It should be restated that this implementation of the method was also underutilized as it only isolated individual damage and did not create various combinations of the damage, which would have only further increased the data set size by a significant margin.

Using the eFlaw concept and various augmentation techniques, it was possible to train two separate CNNs. The first model was a custom VGG inspired model and the second was a transfer learning model of a pre-trained VGG16 model. Limitations due to hardware, time and cost restricted the depth of training and verification methods that could be utilised. Regardless of these limitations, two models were successfully trained with both models achieving F1-scores of 97.5% and 94.9% respectively. From the various metrics it was observed that the custom VGG was consistently more effective than the transfer learning model. An evaluation of classification results for different damage sizes at different depths revealed that the transfer learning model struggled across most damage sizes, with particular trouble evaluating the smaller damage. The custom VGG model, on the other hand, excelled across the board and was also able to detect manufacturing defects effectively. Overall, the custom VGG model proved extremely effective and broke the rule of thumb that damage below half a wavelength could not be detected reliably.

This strong performance, however, prevented a conventional reliability method from being applied. The results from the model were unable to be used for the creation of a PoD curve. As a result, it brings into question the many papers that make use of this method and whether it is a good performance metric to use when comparing different models.

Explainable methods were then applied to the custom VGG model with mixed results. LIME and Integrated gradients proved ineffective in inferring any local features from the model. It should be noted that LIME did pay particular attention to the incident wave of the image, which would later be the dominant feature which the model focused on in order to infer its decisions. GradCam and studying the activation maps proved to be the most successful, with Gradcam being able to provide some insight into the location of damage while focusing mostly on the incident wave as well. The activation maps provided a clearer result, with a feature map being identified which was highlighting the damaged feature areas consistently.

This feature map was then used to develop a tool which would provide a C-scan from the perspective of this feature map, requiring no gates or tuning. This new type of scan was referred to as an F-scan and was able to identify damaged regions with the potential to assist inspectors in their tasks. It opens the possibility for inspectors to be guided to the relevant areas where after their own evaluation they can conclude what must be done with this damaged region.

The potential of this very sensitive model is that it increases the envelope of detectable damage for very low-frequency ultrasonic probes. This finding was new, as it paves the way for the ability to inspect very thick composites with low attenuation but still with the ability to detect near-millimeter-sized defects. The results of this could soon reduce the tradeoff that inspectors must make between probe frequency, attenuation and damage detectability.

The guidelines have clearly stated that training and maintenance tools built on top of already certified hardware, need not undergo initial risk assessments. Given the results of this paper and this leniency from the guidelines, it is clear that if operational performance needs to be evaluated then AI for maintenance is a viable option to research systems that operate so closely with the aircraft, and gain a better understanding for potential methods which can be used to evaluate the performance and more importantly the reliability of these models.

This thesis set out to cover a large scope by investigating the end-to-end development of a data-driven tool for use in aircraft maintenance. Its goal was never to produce a high-performing model with exceptional accuracy, though impressive, that point has already been demonstrated by many other researchers. The real contribution that this thesis aimed to offer was a transparent development process for data-driven maintenance tools with the goal of directing research towards assisting the development of the guidelines. This thesis also contributes by providing a large carefully labelled dataset of ultrasonic data, contributing to the small amount of data currently available to researchers who want to study AI and ultrasonic inspection.

Given the results of this thesis, it has become clear that there is a very strong potential for AI in the maintenance of aerospace structures and the new potential for better inspection capabilities for thick composites. This thesis has also provided an overview for future researchers to quickly develop these tools for research, so that they may focus their efforts on less developed areas within the field. With this technology, it is without a doubt that the already high safety standards of the industry will only be elevated with the advent of such tools and that it will be possible to meet the increased growth of the industry while reducing the strain and demand from human inspectors.

7 Future work

This thesis covered a broad scope in the development of data-driven maintenance aviation tools. Within the time frame of a Master's thesis, it was naturally not possible to investigate every topic in depth, additionally, there were also a series of observations that could be made only at the end of this project which can pave the way for future research topics. This section intends to discuss these topics in the hopes of guiding researchers into exploring these topics. To discuss these points this section discuss future research topics surrounding the eFlaw concept, the model and training design, model evaluation, and finally the explainable methods.

The eFlaw concept proved to be a successful and valuable method for the training and performance of the resulting model, while also paving the way for cost-effective generalized training sets. However, with this being the first introduction of the concept into composite materials there were a few observations and lessons learned from the process. The first notable observation was the impact that varying thicknesses within the specimen had on limiting the manipulation of the extracted features, particularly in the back wall reflection not aligning correctly due to the phase mismatch. Research into perhaps isolating these features to then transform them could open more possibilities in the data augmentation of such data. The second observation of this method was the proximity of the different damage features from one another, and to what degree they interfere with each other's signals. This is relevant particularly when cropping these features. A distance of two wavelengths does not appear to have been sufficient to prevent interference. The third point of development on the eFlaw concept was to investigate the implementation of different damage analogues. A potential research avenue is to use simulated baseline ultrasonic scans for damage analogues which must be introduced during the manufacturing of the samples. This also paves the way for the use of generative adversarial networks to augment the simulated data to represent real-world ultrasonic data more accurately. The final point of research is related to the ability of this method to create a large number of different combinations depending on the number of features present within the data. During the course of this thesis, the concept was not fully utilised and damage features were only isolated, but no combinations were made. An interesting point of investigation on this topic would be to investigate the impact that having all of the possible combinations present would have on the model performance compared to the limited application of the concept in this thesis.

The next area of focus for future research is regarding the training data and model design. This thesis attempted to investigate the relation that damage size and depth have on detectability. This focus was on the damaged B-scans on how generalised the model was at classifying these damages, however, little focus was made on undamaged samples. To research the generalisation of undamaged scans it would be interesting to include samples with significantly varying thickness as the edge cases.

Another interesting observation was that the model sometimes struggled with larger damage features, and the question then becomes does this performance degrade the large the damage feature becomes? Conversely, the model was able to detect damage where the size was a quarter wavelength, understanding the limitations of this model could prove helpful in being able to understand the model's uncertainty and plotting PoD curves of the model's performance. However, it is also acknowledged that the damage sizes will slowly approach the limitations on the sizes that mechanical manufacturing methods can achieve, and might prove too challenging to practically investigate.

During the explainability phase of the thesis, it was found that the model paid particular attention to the incident wave of the signal while rarely (if ever) referring to the echo signals that follow the back-wall reflection. This introduces the potential point of research for a more compact dataset but also a more computationally efficient model, an important point of research now that many industries are discussing the sustainability of data-driven systems. This focus on the incident wave also suggests that there is information within the near field which can already communicate whether the damage is present or not and could lead to different methods for human inspectors to interpret signals.

The transfer learning model was also another point that could be improved. There are two possible directions regarding this topic. The first direction is to continue transfer learning on models trained on very large datasets of visual images, and experimenting with more simple changes to the models as opposed to what was implemented in this thesis. The second direction is to investigate the creation of a large generalised ultrasonic/signal-based dataset and to train a model on this data to provide a foundation of a heavily generalised model for researchers to transfer learning from. The potential of this application would be an even further reduction in experimental costs and the reduced demand for data if the parameters are further reduced. The latter direction could also pave the way for a unified dataset for researchers to compare their results with, similar to the other machine vision competitions currently held by other industries.

The performance of the evaluation of these models was also found to be a key point of research. The conventional PoD curves will prove challenging to implement unless a full study is done to include a human inspector where all of the human-related factors are accounted for. Additionally, it would have been interesting to perform some form of uncertainty quantification using Bayesian neural networks.

Finally, this chapter will end by emphasizing how this thesis, despite its best efforts, has only scraped the tip of the iceberg when it comes to evaluating the EASA guidelines. A further investigation on this subject, particularly with reference to the guidelines may reveal potential insights missed by this thesis, and identify further points of research to aid in the development of the guidelines.

Bibliography

[1] European Aviation Safety Agency. *Artificial Intelligence Roadmap*. 2020.

[2] Guillaume Soudain. *EASA Concept Paper: guidance for Level 1 and 2 machine learning applications*. Feb. 2023.

[3] Stephan Schmidt. *Qualifying A.I. for Ultrasonic inspection: Literature Study*. Delft University of Technology, Jan. 2023.

[4] Sergio Cantero-Chinchilla, Paul D. Wilcox, and Anthony J. Croxford. "Deep learning in automated ultrasonic NDE – Developments, axioms and opportunities". In: *NDT E International* 131 (Oct. 2022), p. 102703. ISSN: 0963-8695. DOI: `10.1016/J.NDTEINT.2022.102703`.

[5] Hossein Towsyfyan, Ander Biguri, Richard Boardman, and Thomas Blumensath. "Successes and challenges in non-destructive testing of aircraft composite structures". In: *Chinese Journal of Aeronautics* 33 (3 Mar. 2020), pp. 771–791. ISSN: 1000-9361. DOI: `10.1016/J.CJA.2019.09.017`.

[6] William Roeseler, Branko Sarh, and Max Kismarton. "composite structures: the first 100 years". In: *16th international conference on composite materials* (2016).

[7] Iikka Virkkunen, Ulf Ronneteg, Göran Emilsson, Thomas Grybäck, and Kaisa Miettinen. "Feasibility study of using eflaws on qualification of nuclear spent fuel disposal canister inspection". In: *12th International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurized Components* (2018). URL: `http://www.ndt.net/?id=22532`.

[8] Lester W. Schmerr. *Fundamentals of Ultrasonic Phased Arrays*. Vol. 215. Springer International Publishing, 2015. ISBN: 978-3-319-07271-5. DOI: `10.1007/978-3-319-07272-2`. URL: `http://link.springer.com/10.1007/978-3-319-07272-2`.

[9] Sandy Cochran. "Ultrasonics, Part 12. Fundamentals of ultrasonic phased arrays". In: *Insight: Non-Destructive Testing and Condition Monitoring* 48 (4 Apr. 2006), pp. 212–217. ISSN: 13542575. DOI: `10.1784/insi.2006.48.4.212`.

[10] Chi Hyung Seo and Jesse T. Yen. "Sidelobe suppression in ultrasound imaging using dual apodization with cross-correlation". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 55 (10 Oct. 2008), pp. 2198–2210. ISSN: 08853010. DOI: `10.1109/TUFFC.919`.

[11] Alison C J Glover. "Non-Destructive Testing Techniques for Aerospace Applications". In: *DSTG International Conference on Health and Usage Monitoring* (2011). URL: `https://humsconference.com.au/Papers2011/Glover_A_Non-Destructive_Testing_Techniques_for.pdf`.

[12] Hossein Taheri and Ahmed Arabi Hassen. "Nondestructive Ultrasonic Inspection of Composite Materials: A Comparative Advantage of Phased Array Ultrasonic". In: *Applied Sciences* 9 (8 Apr. 2019), p. 1628. ISSN: 2076-3417. DOI: `10.3390/app9081628`.

[13] Mertol Gökelma, Damien Latacz, and Bernd Friedrich. "A Review on Prerequisites of a Set-Up for Particle Detection by Ultrasonic Waves in Aluminium Melts". In: *Open Journal of Metal* 06 (01 2016), pp. 13–24. ISSN: 2164-2761. DOI: `10.4236/ojmetal.2016.61002`.

[14] Lester W. Schmerr. *Fundamentals of Ultrasonic Nondestructive Evaluation: A modeling approach*. Second. Springer International Publishing, 2016. ISBN: 978-3-319-30461-8. DOI: `10.1007/978-3-319-30463-2`.

[15] Olympus. *RollerFORM: Phased Array Wheel Probe*. URL: `https://www.olympus-ims.com/en/rollerform/`.

[16] Edward A Ginzel, Ryan Thomson, and Robert K Ginzel. *A Qualification Process for Phased-Array UT using DNV RP-F118 Guidelines*. Materials research Institute, 2011. URL: `https://www.ndt.net/article/ndtnet/2011/21_Ginzel.pdf`.

[17] European Committee for Standardization. *About CEN - CEN-CENELEC*. URL: `https://www.cencenelec.eu/about-cen/`.

[18] European Committee of standards. *EN 4179: Aerospace series - Qualification and approval of personnel for non-destructive testing*. 2009.

[19] European committee of standards. *EN 473: Non-destructive testing-Qualification and certification of NDT personnel-General principles*. 2008.

[20] João da Cruz Payão Filho, Vinicius Pereira Maia, Elisa Kimus Dias Passos, Rodrigo Stohler Gonzaga, and Diego Russo Juliano. "Probability of detection of discontinuities by ultrasonic phased array inspection of 9% Ni steel joints welded with alloy 625 as the filler metal". In: *Ultrasonics* 119 (Feb. 2022), p. 106582. ISSN: 0041-624X. DOI: `10.1016/J.ULTRAS.2021.106582`.

[21] Joint Research Centre and Institute for Energy and Transport, L Gandossi, and C Annis. *Probability of detection curves : statistical best-practices*. Publications Office, 2011. DOI: `doi/10.2790/21826`.

[22] Charles Annis, Luca Gandossi, and Oliver Martin. "Optimal sample size for probability of detection curves". In: *Nuclear Engineering and Design* 262 (2013), pp. 98–105. ISSN: 00295493. DOI: `10.1016/J.NUCENGDES.2013.03.059`.

[23] American Society for Testing and Materials. *ASTM - E3023-21: Standard Practice for Probability of Detection Analysis for â Versus a Data*. 2021. DOI: `10.1520/E3023-21`.

[24] Adam C Cobb, Jay Fisher, and Jennifer E Michaels. *Model-Assisted Probability of Detection for Ultrasonic Structural Health Monitoring*. 2010. URL: `www.ndt.net/index.php?id=8333`.

[25] United States Department of Defence. *MIL-HDBK-1823A: NONDESTRUCTIVE EVALUATION SYSTEM RELIABILITY ASSESSMENT*. 2009. URL: `http://www.everyspec.com`.

[26] American Society for Testing and Materials. "ASTM - E2862: Standard Practice for Probability of Detection Analysis for Hit/Miss Data". In: (2018). DOI: `10.1520/E2862-18`. URL: `www.astm.org,`.

[27] Shuncong Zhong and Walter Nsengiyumva. *Nondestructive Testing and Evaluation of Fiber-Reinforced Composite Structures*. Springer Nature Singapore, 2022. ISBN: 978-981-19-0847-7. DOI: `10.1007/978-981-19-0848-4`.

[28] S. I. B. Syed Abdullah. "Drop Test Impact Analysis—Experimental and Numerical Evaluations". In: *Composites Science and Technology* (2021), pp. 35–46. DOI: `10.1007/978-981-16-1323-4_3`. URL: `https://link.springer.com/chapter/10.1007/978-981-16-1323-4_3`.

[29] Song Zhou, Yan Li, Kunkun Fu, and Xiaodi Wu. "Progressive fatigue damage modelling of fibre-reinforced composite based on fatigue master curves". In: *Thin-Walled Structures* 158 (Jan. 2021), p. 107173. ISSN: 0263-8231. DOI: `10.1016/J.TWS.2020.107173`.

[30] P. Blain, J.-F. Vandenrijt, F. Languy, M. Kirkove, L.-D. Théroux, J. Lewandowski, and M. Georges. *Artificial defects in CFRP composite structure for thermography and shearography nondestructive inspection*. Ed. by Anand K. Asundi. June 2017. DOI: `10.1117/12.2271701`.

[31] Paul E. Mix. *Introduction to Nondestructive Testing*. John Wiley Sons, Inc., May 2004. ISBN: 9780471719144. DOI: `10.1002/0471719145`.

[32] Herberth Birck Fröhlich. *Evaluation of adhesion failures in composite laminated plates using deep learning-based object detection in shearography images*. 2021. URL: `https://repositorio.ufsc.br/handle/123456789/226970`.

[33] Iikka Virkkunen, Kaisa Miettinen, and Tapani Packalén. "Virtual flaws for NDE training and qualification". In: *11th European Conference on Non-Destructive Testing* (2014).

[34] Richard J. Pyle, Rhodri L. T. Bevan, Robert R. Hughes, Rosen K. Rachev, Amine Ait Si Ali, and Paul D. Wilcox. "Deep Learning for Ultrasonic Crack Characterization in NDE". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 68 (5 May 2021), pp. 1854–1865. ISSN: 0885-3010. DOI: `10.1109/TUFFC.2020.3045847`.

[35] Christopher Bishop. *Pattern Recognition and Machine Learning*. 1st ed. Springer, Aug. 2006. ISBN: 978-0-387-31073-2.

[36] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer US, 2021. ISBN: 978-1-0716-1417-4. DOI: `10.1007/978-1-0716-1418-1`.

[37] Zihao Guo, Jianqiao Zhou, and Di Zhao. "Thyroid Nodule Ultrasonic Imaging Segmentation Based on a Deep Learning Model and Data Augmentation". In: *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020* (June 2020), pp. 549–554. DOI: `10.1109/ITNEC48623.2020.9085093`.

[38] Duje Medak, Luka Posilovic, Marko Subasic, Marko Budimir, and Sven Loncaric. "Automated Defect Detection from Ultrasonic Images Using Deep Learning". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 68 (10 Oct. 2021), pp. 3126–3134. ISSN: 15258955. DOI: `10.1109/TUFFC.2021.3081750`.

[39] Wangzhe Du, Hongyao Shen, Jianzhong Fu, Ge Zhang, and Quan He. "Approaches for improvement of the X-ray image defect detection of automobile casting aluminum parts based on deep learning". In: *NDT and E International* 107 (Oct. 2019). ISSN: 09638695. DOI: `10.1016/J.NDTEINT.2019.102144`.

[40] Nauman Munir, Hak Joon Kim, Jinhyun Park, Sung Jin Song, and Sung Sik Kang. "Convolutional neural network for ultrasonic weldment flaw classification in noisy conditions". In: *Ultrasonics* 94 (Apr. 2019), pp. 74–81. ISSN: 0041-624X. DOI: `10.1016/J.ULTRAS.2018.12.001`.

[41] Thibault Latête, Baptiste Gauthier, and Pierre Belanger. "Towards using convolutional neural network to locate, identify and size defects in phased array ultrasonic testing". In: *Ultrasonics* 115 (Aug. 2021). ISSN: 0041624X. DOI: `10.1016/J.ULTRAS.2021.106436`.

[42] Tuomas Koskinen, Iikka Virkkunen, Oskar Siljama, and Oskari Jessen-Juhler. "The Effect of Different Flaw Data to Machine Learning Powered Ultrasonic Inspection". In: *Journal of Nondestructive Evaluation* 40 (1 Mar. 2021), p. 24. ISSN: 0195-9298. DOI: `10.1007/s10921-021-00757-x`.

[43] Sebastian Meister, Nantwin Möller, Jan Stüve, and Roger M. Groves. "Synthetic image data augmentation for fibre layup inspection processes: Techniques to enhance the data set". In: *Journal of Intelligent Manufacturing* 32 (6 Aug. 2021), pp. 1767–1789. ISSN: 15728145. DOI: `10.1007/s10845-021-01738-7`.

[44] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. "Generative Adversarial Networks: An Overview". In: *IEEE Signal Processing Magazine* 35 (1 Jan. 2018), pp. 53–65. ISSN: 10535888. DOI: `10.1109/MSP.2017.2765202`.

[45] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. *Improving neural networks by preventing co-adaptation of feature detectors*. July 2012. URL: `https://arxiv.org/abs/1207.0580v1`.

[46] Sangchul Hahn and Heeyoul Choi. "Understanding dropout as an optimization trick". In: *Neurocomputing* 398 (July 2020), pp. 64–70. ISSN: 0925-2312. DOI: `10.1016/J.NEUCOM.2020.02.067`.

[47] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *32nd International Conference on Machine Learning, ICML 2015* 1 (Feb. 2015), pp. 448–456. URL: `https://arxiv.org/abs/1502.03167v3`.

[48] Matthew D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. Dec. 2012. URL: `https://arxiv.org/abs/1212.5701v1`.

[49] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. "Activation functions in deep learning: A comprehensive survey and benchmark". In: *Neurocomputing* 503 (Sept. 2022), pp. 92–108. ISSN: 0925-2312. DOI: `10.1016/J.NEUCOM.2022.06.111`.

[50] Rémi Bardenet, Mátyás Brendel, Balázs Kégl, and Michèle Sebag. "Collaborative hyperparameter tuning". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 2. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 199–207. URL: `https://proceedings.mlr.press/v28/bardenet13.html`.

[51] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. *Deep Learning Scaling is Predictable, Empirically*. Dec. 2017. DOI: `10.48550/arxiv.1712.00409`. URL: `https://arxiv.org/abs/1712.00409v1`.

[52] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" In: (Nov. 2015). DOI: `10.48550/arxiv.1511.06348`. URL: `https://arxiv.org/abs/1511.06348v2`.

[53] Rosa L. Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H. Ngo. "Predicting sample size required for classification performance". In: *BMC Medical Informatics and Decision Making* 12 (1 Feb. 2012), pp. 1–10. ISSN: 14726947. DOI: `10.1186/1472-6947-12-8/TABLES/1`. URL: `https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-8`.

[54] Oskar Siljama, Tuomas Koskinen, Oskari Jessen-Juhler, and Iikka Virkkunen. "Automated Flaw Detection in Multi-channel Phased Array Ultrasonic Data Using Machine Learning". In: *Journal of Nondestructive Evaluation* 40 (3 Sept. 2021). ISSN: 15734862. DOI: `10.1007/s10921-021-00796-4`.

[55] Kushal Virupakshappa and Erdal Oruklu. "Ultrasonic flaw detection using Support Vector Machine classification". In: *2015 IEEE International Ultrasonics Symposium (IUS)*. IEEE, Oct. 2015, pp. 1–4. ISBN: 978-1-4799-8182-3. DOI: `10.1109/ULTSYM.2015.0128`.

[56] Kushal Virupakshappa and Erdal Oruklu. "Ultrasonic flaw detection using Hidden Markov Model with wavelet features". In: *IEEE International Ultrasonics Symposium (IUS)*. IEEE, Sept. 2016, pp. 1–4. ISBN: 978-1-4673-9897-8. DOI: `10.1109/ULTSYM.2016.7728491`.

[57] Kushal Virupakshappa, Michael Marino, and Erdal Oruklu. "A Multi-Resolution Convolutional Neural Network Architecture for Ultrasonic Flaw Detection". In: *IEEE International Ultrasonics Symposium, IUS* 2018-October (Dec. 2018). ISSN: 19485727. DOI: `10.1109/ULTSYM.2018.8579888`.

[58] Mukhammed Garifulla, Juncheol Shin, Chanho Kim, Won Hwa Kim, Hye Jung Kim, Jaeil Kim, and Seokin Hong. "A case study of quantizing convolutional neural networks for fast disease diagnosis on portable medical devices". In: *Sensors* 22 (1 Jan. 2022). ISSN: 14248220. DOI: `10.3390/S22010219`.

[59] Asif Khan, Izaz Raouf, Yeong Rim Noh, Daun Lee, Jung Woo Sohn, and Heung Soo Kim. "Autonomous assessment of delamination in laminated composites using deep learning and data augmentation". In: *Composite Structures* 290 (June 2022), p. 115502. ISSN: 0263-8223. DOI: `10.1016/J.COMPSTRUCT.2022.115502`.

[60] Iikka Virkkunen, Tuomas Koskinen, Oskari Jessen-Juhler, and Jari Rinta-aho. "Augmented ultrasonic data for machine learning". In: *Journal of Nondestructive Evaluation* 40.1 (2021). DOI: `10.1007/s10921-020-00739-5`.

[61] Joseph Melville, K. Supreet Alguri, Chris Deemer, and Joel B. Harley. "Structural damage detection using deep learning of ultrasonic guided waves". In: *AIP Conference Proceedings* 1949 (1 Apr. 2018), p. 230004. ISSN: 0094-243X. DOI: `10.1063/1.5031651`. URL: `https://aip.scitation.org/doi/abs/10.1063/1.5031651`.

[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[63] Umberto Michelucci. *Advanced applied deep learning: Convolutional neural networks and object detection*. Apress Media LLC, Sept. 2019, pp. 1–285. ISBN: 9781484249765. DOI: `10.1007/978-1-4842-4976-5`.

[64] Apeksha Shewalkar, Deepika Nyavanandi, and Simone A. Ludwig. “Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU”. In: *Journal of Artificial Intelligence and Soft Computing Research* 9 (4 Oct. 2019), pp. 235–245. ISSN: 2083-2567. DOI: 10.2478/jaiscr-2019-0006.

[65] S Meister. “Automated Defect Analysis using Optical Sensing and Explainable Artificial Intelligence for Fibre Layup Processes in Composite Manufacturing”. 2022. DOI: 10.4233/uuid:34442378-e3a2-4c99-865f-57be3f13b96f. URL: https://doi.org/10.4233/uuid:34442378-.

[66] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Networks”. In: *Science Robotics* 3 (January June 2014), pp. 2672–2680. ISSN: 10495258. DOI: 10.48550/arxiv.1406.2661. URL: https://arxiv.org/abs/1406.2661v1.

[67] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep Unsupervised Learning using Nonequilibrium Thermodynamics”. In: *32nd International Conference on Machine Learning, ICML 2015* 3 (Mar. 2015), pp. 2246–2255. DOI: 10.48550/arxiv.1503.03585. URL: https://arxiv.org/abs/1503.03585v8.

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

[69] Open AI. *DALL·E 2*. 2022. URL: https://openai.com/dall-e-2/.

[70] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86 (11 1998), pp. 2278–2323. ISSN: 00189219. DOI: 10.1109/5.726791.

[71] Qingxue Zhang, Dian Zhou, and Xuan Zeng. “HeartID: A Multiresolution Convolutional Neural Network for ECG-Based Biometric Human Identification in Smart Health Applications”. In: *IEEE Access* 5 (2017), pp. 11805–11816. ISSN: 21693536. DOI: 10.1109/ACCESS.2017.2707460.

[72] Viswanatha V, Chandana R K, and Ramachandra A. “Real Time Object Detection System with YOLO and CNN Models: A Review”. In: *journal of xi an university of architecture technology* (July 2022). DOI: 10.48550/arxiv.2208.00773. URL: https://arxiv.org/abs/2208.00773v1.

[73] Jiazhi Liang. “Image classification based on RESNET”. In: *Journal of Physics: Conference Series* 1634 (Sept. 2020), p. 012110. DOI: 10.1088/1742-6596/1634/1/012110.

[74] Mingxing Tan, Ruoming Pang, and Quoc V. Le. “EfficientDet: Scalable and Efficient Object Detection”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Nov. 2019), pp. 10778–10787. ISSN: 10636919. DOI: 10.48550/arxiv.1911.09070. URL: https://arxiv.org/abs/1911.09070v7.

[75] Mingxing Gong. “A Novel Performance Measure for Machine Learning Classification”. In: *International Journal of Managing Information Technology* 13 (1 Feb. 2021), pp. 11–19. ISSN: 09755926. DOI: 10.5121/ijmit.2021.13101.

[76] Wissam Siblini, Jordan Fréry, Liyun He-Guelton, Frédéric Oblé, and Yi Qing Wang. “Master Your Metrics with Calibration”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12080 LNCS (2020), pp. 457–469. ISSN: 16113349. DOI: 10.1007/978-3-030-44584-3_36/FIGURES/7. URL: https://link.springer.com/chapter/10.1007/978-3-030-44584-3_36.

[77] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC Genomics* 21 (1 Jan. 2020), pp. 1–13. ISSN: 14712164. DOI: 10.1186/S12864-019-6413-7/TABLES/5. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7.

[78] M Wall, F A Wedgwood, and S Burch. *Modelling of NDT Reliability (POD) and applying corrections for human factors*. May 1998. URL: https://www.ndt.net/article/ecndt98/reliabil/325/325.htm.

[79] Christopher D. Brown and Herbert T. Davis. “Receiver operating characteristics curves and related decision measures: A tutorial”. In: *Chemometrics and Intelligent Laboratory Systems* 80 (1 Jan. 2006), pp. 24–38. ISSN: 0169-7439. DOI: 10.1016/J.CHEMOLAB.2005.05.004.

[80] David L. Streiner and John Cairney. “What’s under the ROC? An Introduction to Receiver Operating Characteristics Curves”. In: *The Canadian Journal of Psychiatry* 52 (2 Feb. 2007), pp. 121–128. ISSN: 14970015. DOI: 10.1177/070674370705200210. URL: https://journals.sagepub.com/doi/10.1177/070674370705200210.

[81] Defense Advanced Research Projects Agency. *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency, Aug. 2016. URL: https://www.darpa.mil/program/explainable-artificial-intelligence.

[82] Andrew D Selbst and Julia Powles. “Meaningful information and the right to explanation”. In: *International Data Privacy Law* 7 (4 Nov. 2017), pp. 233–242. ISSN: 2044-3994. DOI: 10.1093/idpl/ipx022.

[83] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. “DARPA ’s explainable AI (XAI) program: A retrospective”. In: *Applied AI Letters* 2 (4 Dec. 2021). ISSN: 2689-5595. DOI: 10.1002/ail2.61.

[84] Yann LeCun, Corinna Cortes, and Christopher Burges. *MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges*. URL: http://yann.lecun.com/exdb/mnist/.

[85] Alexander Tong, David van Dijk, Jay S. Stanley III, Matthew Amodio, Kristina Yim, Rebecca Muhle, James Noonan, Guy Wolf, and Smita Krishnaswamy. “Interpretable Neuron Structuring with Graph Spectral Regularization”. In: *Advances in Intelligent Data Analysis*. Vol. XVII. 2020, pp. 509–521. DOI: 10.1007/978-3-030-44584-3_40.

[86] Jeroen van Doorenmalen and Vlado Menkovski. “Evaluation of CNN Performance in Semantically Relevant Latent Spaces”. In: *Advances in Intelligent Data Analysis*. Vol. XVII. 2020, pp. 145–157. DOI: 10.1007/978-3-030-44584-3_12.

[87] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol. 13-17-August-2016. Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778.

[88] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *34th International Conference on Machine Learning, ICML 2017* 7 (Mar. 2017), pp. 5109–5118. URL: https://arxiv.org/abs/1703.01365v2.

[89] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus Robert Müller, and Wojciech Samek. “Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9887 LNCS (Apr. 2016), pp. 63–71. ISSN: 16113349. DOI: 10.1007/978-3-319-44781-0_8. URL: https://arxiv.org/abs/1604.00825v1.

[90] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *International Journal of Computer Vision* 128 (2 Oct. 2016), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: http://arxiv.org/abs/1610.02391%20http://dx.doi.org/10.1007/s11263-019-01228-7.

[91] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *CoRR* abs/1311.2901 (2013). arXiv: 1311.2901. URL: http://arxiv.org/abs/1311.2901.

[92] Sonatest. *Large Low Frequency WheelProbe Technical documentation*. URL: https://pdf.directindustry.com/pdf/sonatest-ltd/large-low-frequency-wheelprobe/21671-478711.html.

[93] Sonatest. *veo+: Applications and solutions*. URL: https://sonatest.com/resources/application-notes/veo.

[94] Witte. *Alufix Classic*. URL: https://www.witte-barskamp.com/modular-fixturing-systems/alufix-classic-alufix-eco/.

[95] Hexcell. *Hexforce 7581: Product data*. URL: https://www.hexcel.com/user_area/content_media/raw/DSF_7581.pdf.

[96] Hexion. *Epikure/Epikote 04908: Data sheet*. URL: https://www.swiss-composite.ch/pdf/t-Hexion-Harz-EPR04908.pdf.

[97] Bekir Yenilmez, Talha Akyol, Baris Caglar, and E. Murat Sozer. "Minimizing Thickness Variation in the Vacuum Infusion (VI) Process". In: *Advanced Composites Letters* 20 (6 Nov. 2011), p. 096369351102000. ISSN: 2633-366X. DOI: 10.1177/096369351102000603.

[98] *Half wavelength limit*. 2017. URL: https://www.ndt.net/forum/thread.php?msgID=45501.

[99] Josef Krautkrämer and Herbert Krautkrämer. *Ultrasonic Testing of Materials*. Springer Berlin Heidelberg, 1990. ISBN: 978-3-662-10682-2. DOI: 10.1007/978-3-662-10680-8.

[100] The SciPy community. *scipy.signal.correlate — SciPy v1.11.1 Manual*. 2023. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.correlate.html.

[101] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. "Image quality assessment: From error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13 (4 Apr. 2004), pp. 600–612. ISSN: 10577149. DOI: 10.1109/TIP.2003.819861.

[102] Mahdi Hashemi. "Enlarging smaller images before inputting into convolutional neural network: Zero-padding vs. interpolation". In: *Journal of Big Data* 6.1 (2019). DOI: 10.1186/s40537-019-0263-7.

[103] Richard Jones, Antony Hosking, and Eliot Moss. *The garbage collection handbook : art of automatic memory management.* CRC Press, 2012, 481 blz. ISBN: 9781420082791.

[104] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[105] Martin Hagan, Howard Demuth, Mark Beale, and Orlando Jesus. *Neural Network Design*. 2nd ed. Martin Hagan, 2014.

[106] Frontline Solvers. *Training an Artificial Neural Network - Intro | solver*. URL: https://www.solver.com/training-artificial-neural-network-intro.

[107] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (Sept. 2014). URL: https://arxiv.org/abs/1409.1556v6.

[108] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[109] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-Learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12.null (Nov. 2011), pp. 2825–2830. ISSN: 1532-4435.

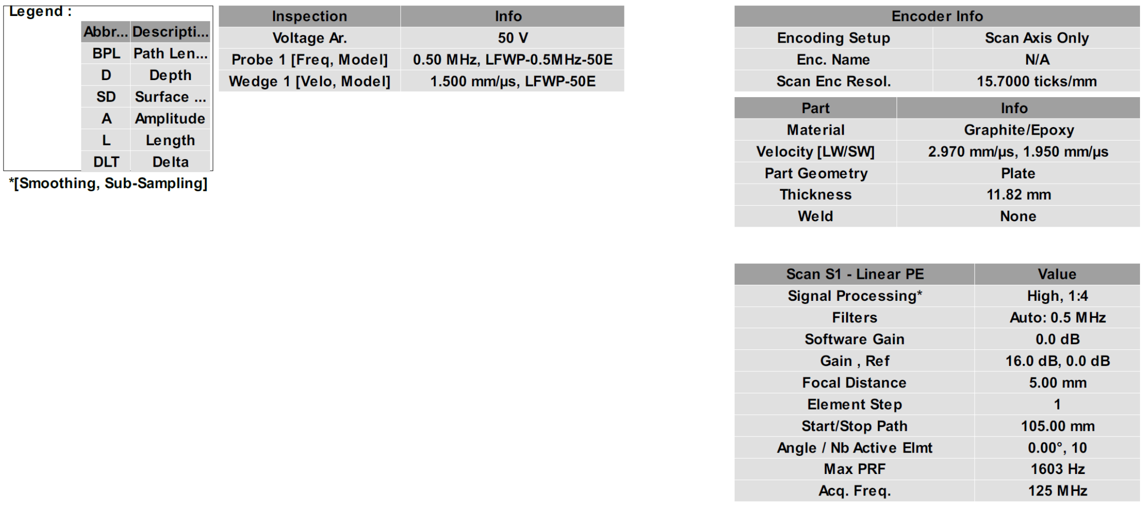
[110] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[111] Raphael Meudec. *tf-explain*. Feb. 2021. DOI: 10.5281/zenodo.5711704. URL: https://github.com/sicara/tf-explain.

[112] Jean Marc Cluzeau, Xavier Henriquel, Georges Rebender, Guillaume Soudain, Luuk van Dijk, Alexey Gronskiy, and David Haber. *Concepts of Design Assurance for Neural Networks (CoDANN)*. EASA, 2020.

A

Low-frequency wheel probe settings



Legend :

Abbr...	Descripti...
BPL	Path Len...
D	Depth
SD	Surface ...
A	Amplitude
L	Length
DLT	Delta

*[Smoothing, Sub-Sampling]

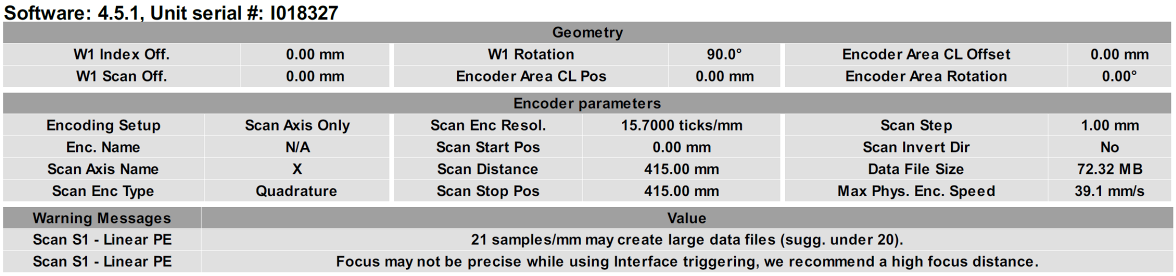
Inspection	Info
Voltage Ar.	50 V
Probe 1 [Freq, Model]	0.50 MHz, LFWP-0.5MHz-50E
Wedge 1 [Velo, Model]	1.500 mm/µs, LFWP-50E

Encoder Info	
Encoding Setup	Scan Axis Only
Enc. Name	N/A
Scan Enc Resol.	15.7000 ticks/mm

Part	Info
Material	Graphite/Epoxy
Velocity [LW/SW]	2.970 mm/µs, 1.950 mm/µs
Part Geometry	Plate
Thickness	11.82 mm
Weld	None

Scan S1 - Linear PE	Value
Signal Processing*	High, 1:4
Filters	Auto: 0.5 MHz
Software Gain	0.0 dB
Gain , Ref	16.0 dB, 0.0 dB
Focal Distance	5.00 mm
Element Step	1
Start/Stop Path	105.00 mm
Angle / Nb Active Elmt	0.00°, 10
Max PRF	1603 Hz
Acq. Freq.	125 MHz

Figure A.1: Overview of the wheel probe and material details



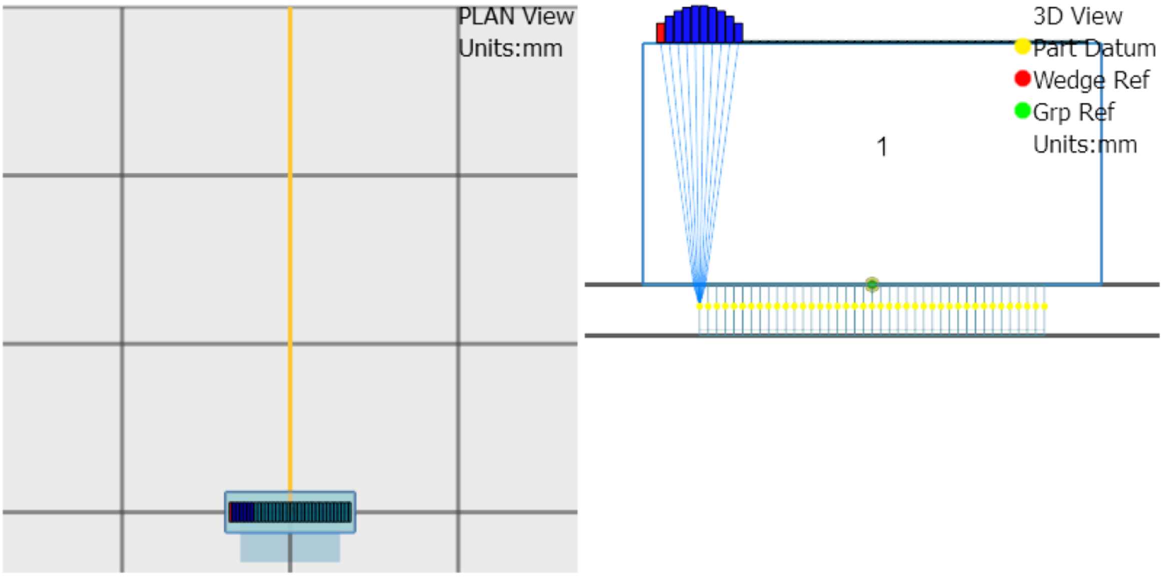
Software: 4.5.1, Unit serial #: I018327

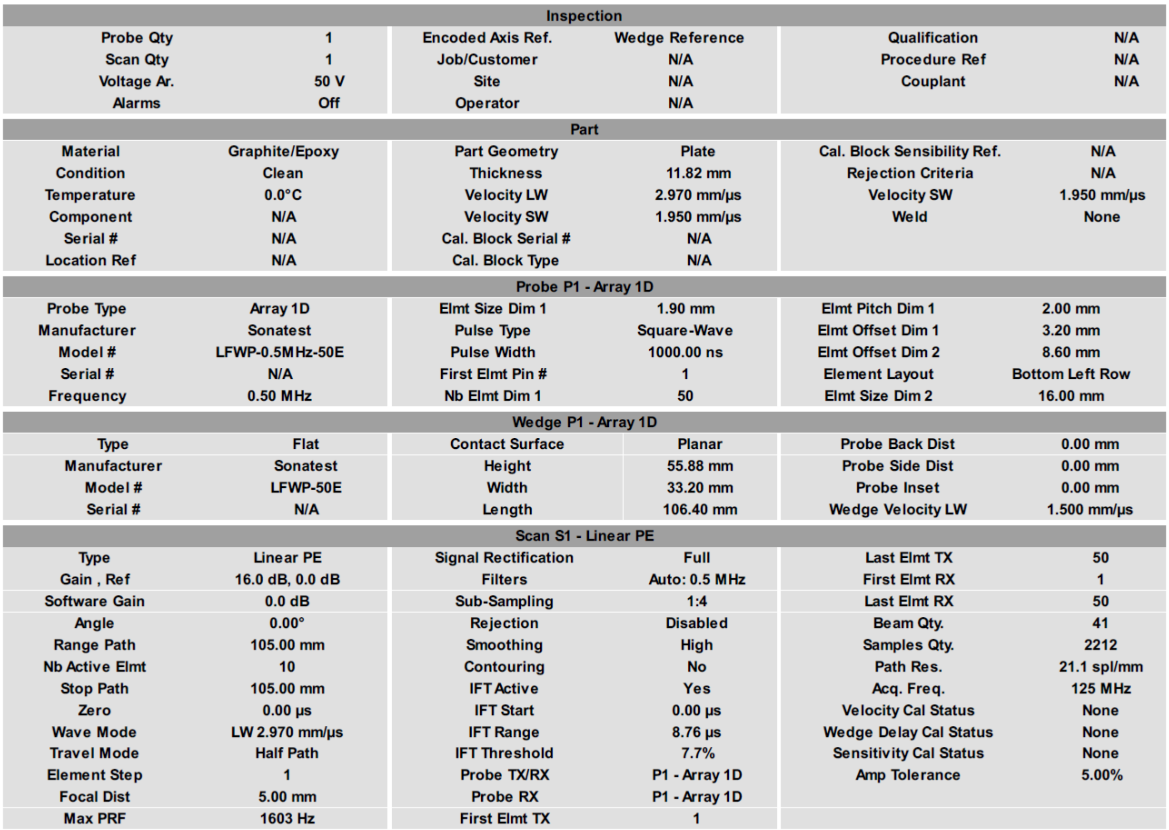
Geometry					
W1 Index Off.	0.00 mm	W1 Rotation	90.0°	Encoder Area CL Offset	0.00 mm
W1 Scan Off.	0.00 mm	Encoder Area CL Pos	0.00 mm	Encoder Area Rotation	0.00°

Encoder parameters					
Encoding Setup	Scan Axis Only	Scan Enc Resol.	15.7000 ticks/mm	Scan Step	1.00 mm
Enc. Name	N/A	Scan Start Pos	0.00 mm	Scan Invert Dir	No
Scan Axis Name	X	Scan Distance	415.00 mm	Data File Size	72.32 MB
Scan Enc Type	Quadrature	Scan Stop Pos	415.00 mm	Max Phys. Enc. Speed	39.1 mm/s

Warning Messages	Value
Scan S1 - Linear PE	21 samples/mm may create large data files (sugg. under 20).
Scan S1 - Linear PE	Focus may not be precise while using Interface triggering, we recommend a high focus distance.

Figure A.2: Details on encoder parameters, warnings, and geometry of the scan set up which in this case was not set.





Inspection

Probe Qty	1	Encoded Axis Ref.	Wedge Reference	Qualification	N/A
Scan Qty	1	Job/Customer	N/A	Procedure Ref	N/A
Voltage Ar.	50 V	Site	N/A	Couplant	N/A
Alarms	Off	Operator	N/A		

Part

Material	Graphite/Epoxy	Part Geometry	Plate	Cal. Block Sensibility Ref.	N/A
Condition	Clean	Thickness	11.82 mm	Rejection Criteria	N/A
Temperature	0.0°C	Velocity LW	2.970 mm/µs	Velocity SW	1.950 mm/µs
Component	N/A	Velocity SW	1.950 mm/µs	Weld	None
Serial #	N/A	Cal. Block Serial #	N/A		
Location Ref	N/A	Cal. Block Type	N/A		

Probe P1 - Array 1D

Probe Type	Array 1D	Elmt Size Dim 1	1.90 mm	Elmt Pitch Dim 1	2.00 mm
Manufacturer	Sonatest	Pulse Type	Square-Wave	Elmt Offset Dim 1	3.20 mm
Model #	LFWP-0.5MHz-50E	Pulse Width	1000.00 ns	Elmt Offset Dim 2	8.60 mm
Serial #	N/A	First Elmt Pin #	1	Element Layout	Bottom Left Row
Frequency	0.50 MHz	Nb Elmt Dim 1	50	Elmt Size Dim 2	16.00 mm

Wedge P1 - Array 1D

Type	Flat	Contact Surface	Planar	Probe Back Dist	0.00 mm
Manufacturer	Sonatest	Height	55.88 mm	Probe Side Dist	0.00 mm
Model #	LFWP-50E	Width	33.20 mm	Probe Inset	0.00 mm
Serial #	N/A	Length	106.40 mm	Wedge Velocity LW	1.500 mm/µs

Scan S1 - Linear PE

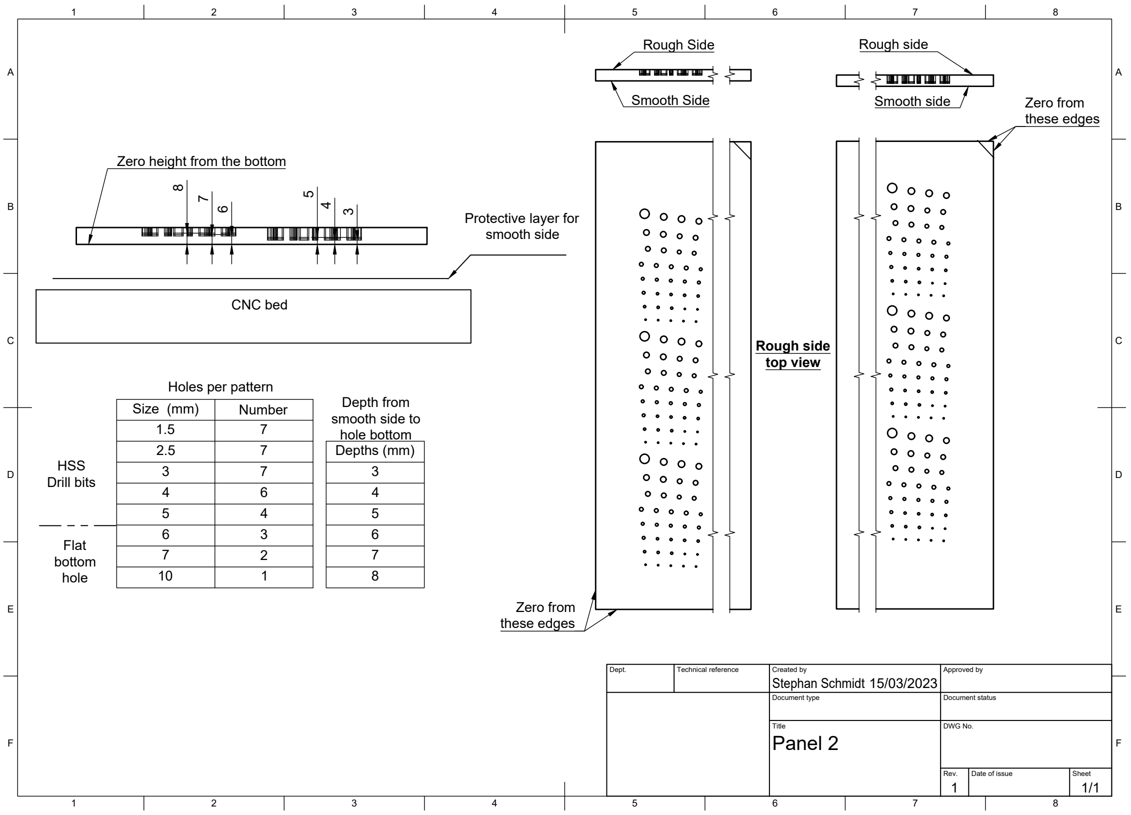
Type	Linear PE	Signal Rectification	Full	Last Elmt TX	50
Gain , Ref	16.0 dB, 0.0 dB	Filters	Auto: 0.5 MHz	First Elmt RX	1
Software Gain	0.0 dB	Sub-Sampling	1:4	Last Elmt RX	50
Angle	0.00°	Rejection	Disabled	Beam Qty.	41
Range Path	105.00 mm	Smoothing	High	Samples Qty.	2212
Nb Active Elmt	10	Contouring	No	Path Res.	21.1 spl/mm
Stop Path	105.00 mm	IFT Active	Yes	Acq. Freq.	125 MHz
Zero	0.00 µs	IFT Start	0.00 µs	Velocity Cal Status	None
Wave Mode	LW 2.970 mm/µs	IFT Range	8.76 µs	Wedge Delay Cal Status	None
Travel Mode	Half Path	IFT Threshold	7.7%	Sensitivity Cal Status	None
Element Step	1	Probe TX/RX	P1 - Array 1D	Amp Tolerance	5.00%
Focal Dist	5.00 mm	Probe RX	P1 - Array 1D		
Max PRF	1603 Hz	First Elmt TX	1		

Figure A.3: Full details on the settings, including the samples, the array, the wedge (roller), and the scanning parameters

B

Technical drawings of the inspection panels

B.1. Panel 2 hole overview



Holes per pattern

Size (mm)	Number
1.5	7
2.5	7
3	7
4	6
5	4
6	3
7	2
10	1

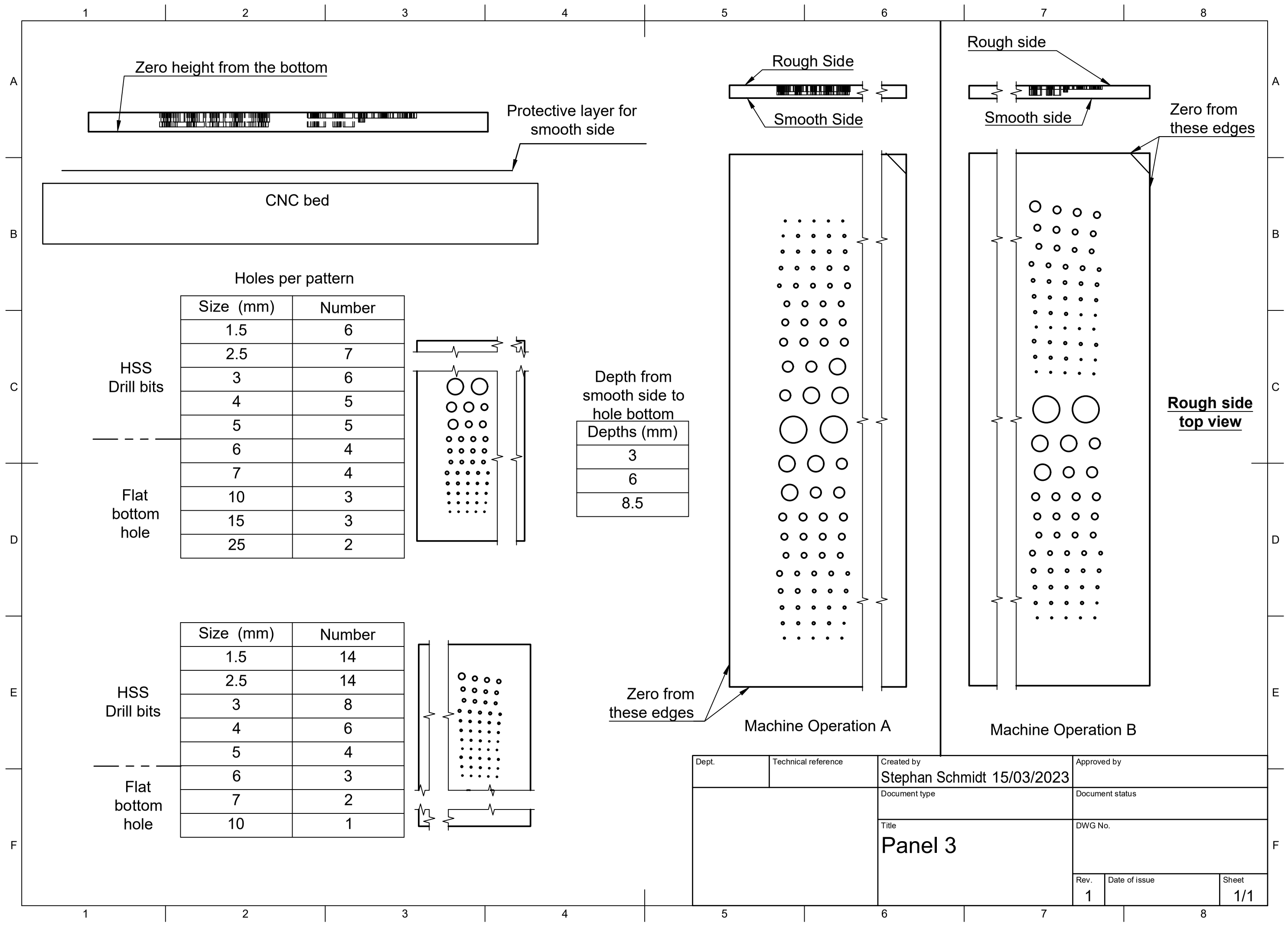
Depth from smooth side to hole bottom

Depths (mm)
3
4
5
6
7
8

Dept.	Technical reference	Created by Stephan Schmidt 15/03/2023	Approved by
		Document type	Document status
		Title Panel 2	DWG No.
			Rev. 1 / Date of issue / Sheet 1/1

Figure B.1: Technical drawing showing the layout of the damage in panel 2. including the total number of each damage size.

B.2. Panel 3 hole overview



Holes per pattern

Size (mm)	Number
1.5	6
2.5	7
3	6
4	5
5	5
6	4
7	4
10	3
15	3
25	2

Size (mm)	Number
1.5	14
2.5	14
3	8
4	6
5	4
6	3
7	2
10	1

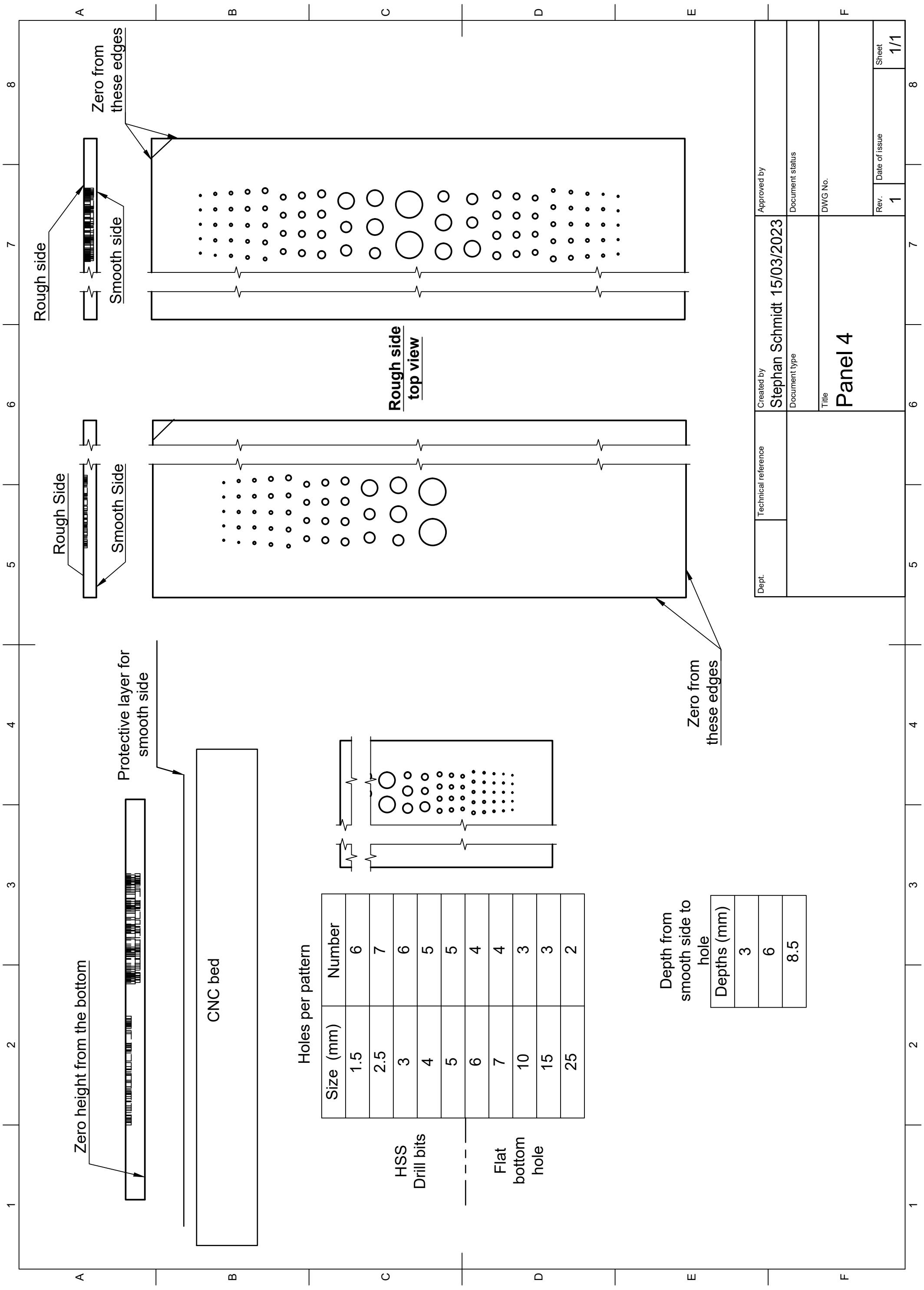
Depth from smooth side to hole bottom

Depths (mm)
3
6
8.5

Dept.	Technical reference	Created by Stephan Schmidt 15/03/2023	Approved by
		Document type	Document status
		Title Panel 3	DWG No.
			Rev. 1 / Date of issue / Sheet 1/1

Figure B.2: Technical drawing showing the layout of the damage in panel 3. including the total number of each damage size.

B.3. Panel 4 hole overview



Holes per pattern

Size (mm)	Number
1.5	6
2.5	7
3	6
4	5
5	5
6	4
7	4
10	3
15	3
25	2

Depth from smooth side to hole

Depths (mm)
3
6
8.5

Figure B.3: Technical drawing showing the layout of the damage in panel 4. including the total number of each damage size.

C

Structural similarity index masks

The strange aspect ratio of the masks is due to the resolution of the scanner and will result in this image with a 1:1 pixel aspect ratio.

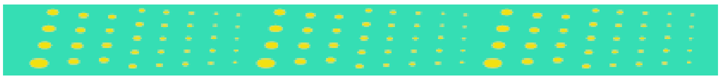


Figure C.1: Mask created using the CAD data for panel 2 lane 1.

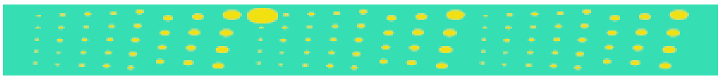


Figure C.2: Mask created using the CAD data for panel 2 lane 2.

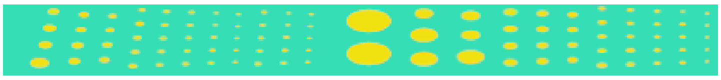


Figure C.3: Mask created using the CAD data for panel 3 lane 1.

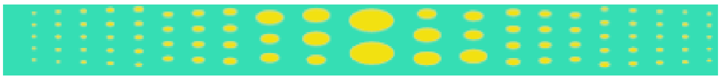


Figure C.4: Mask created using the CAD data for panel 3 lane 2.

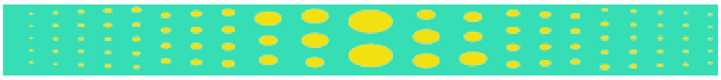


Figure C.5: Mask created using the CAD data for panel 4 lane 1.

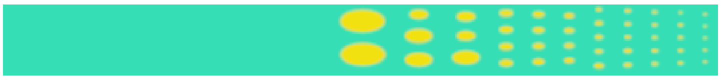


Figure C.6: Mask created using the CAD data for panel 4 lane 2.