Dennis Leinarts

Uncertainty Management for Past State Prognostic Uncertainties in Aerospace Structures

TU Delft

Thesis

Uncertainty Management for Past State Prognostic Uncertainties in Aerospace Structures

Dennis Leinarts

In partial fulfillment of the requirements for the degree of
Master of Science
in Aerospace Engineering

At the Department of Aerospace Structures and Materials in the Intelligent System Prognostics (iSP) Group in the Faculty of Aerospace Engineering at Delft University of Technology (TU Delft)

Studentnumber:	4685709
MSc track:	Structures and Materials
Responsible supervisor:	Dr. Nick Eleftheroglou
Co-supervisors:	Mariana Salinas Camus
	Benjamín Brito Schiele
Date:	2025

Abstract

The aerospace industry increasingly employs composite materials due to their superior strength-to-weight ratios and corrosion resistance. However, the heterogeneous and anisotropic nature of composites introduces complex failure mechanisms such as delamination, matrix cracking and fiber breakage, which make modeling of degradation processes challenging. In addition, variability in material quality due to manufacturing processes, along with operational uncertainties, further challenges the reliable prediction of remaining useful life of life (RUL) of composites. As a result uncertainty management is essential for enabling informed decision-making in prognostics and health management (PHM).

This thesis presents a novel uncertainty management approach specifically targeting past state uncertainties that stem from variability in composite manufacturing processes, such as embedded defects and material quality inconsistencies. By leveraging advanced ultrasonic imaging through Dolphicam technology, internal structural variations in aerospace-grade carbon fiber-reinforced polymer (CFRP) composites were quantified. Subsequently, a similarity-informed methodology was developed, employing both spatial pyramid histogram (SPH) and convolutional neural network (CNN)-based embedding techniques, to group specimens based on internal structural quality. This approach assumes that structurally similar composites exhibit analogous degradation behavior under fatigue loading, enabling targeted and more reliable prognostic modeling.

Experimental validation involved fabricating CFRP laminates with intentionally embedded defects, followed by detailed nondestructive inspections (NDI) and fatigue testing with digital image correlation (DIC)-based strain measurements. The similarity-informed prognostic model demonstrated significant improvements in prediction accuracy and uncertainty quantification compared to conventional methods, confirming the viability and advantages of integrating similarity-based past state uncertainty management in aerospace composite structures to improve reliability in prognostics.

Overall, the results underscore the value of high-fidelity NDI for material quality characterization and highlight the potential of similarity learning techniques to reduce prognostic uncertainty. By enhancing the credibility of RUL predictions, this work contributes to more reliable decision-making capabilities and supports the advancement of predictive maintenance strategies for aerospace composite structures.

Preface

Dear reader,

This thesis marks the final stage of my Master's program in Aerospace Engineering, conducted at Delft University of Technology within the Aerospace Faculty. The work presented here was carried out between September 2024 and July 2025 and focuses on the development of a similarity-informed methodology to manage prognostic uncertainties in composite aerospace structures.

The idea for this research originated from a growing interest in the challenges of data-driven maintenance and the complexities introduced by composite materials in structural health monitoring. During the course of this thesis, I had the opportunity to bridge experimental testing, image-based analysis, and prognostic modeling fields. Given my background in Mechanical Engineering, this mixed aspect suited me perfectly, keeping me both focused and interested.

I am deeply grateful to my supervisor, Nick Eleftheroglou, whose guidance, patience, support and critical insight were invaluable throughout this project. Not only was he a support professionally, but also personally. I would also like to thank my co-supervisors Mariana Salinas Camus and Benjamín Brito Schiele. Benjamín for his for his advice during meetings and personal effort, and Mariana for providing the SL-HSMM modeling framework and helpful discussions on prognostic validation.

On a personal note, I want to thank my family, friends, and especially my girlfriend, for their unwavering encouragement and mental support in stressful times. I hope that readers find this thesis both informative and inspiring, and I invite you to explore its contents.

Dennis Leinarts

Delft

June 19, 2025

Contents

Acronyms

List of Figures

List of Tables

1. Introduction

The aerospace industry is in an ongoing transitional phase where the integration of composite materials into aircraft structures is becoming more dominant [4]. In the mid-1990s composites were predominantly used in secondary structures of commercial aircraft and made up 15% - 20% of the total structural weight. While, composites in the latest generation of commercial aircraft such as the Boeing 787 and Airbus A350 account for over 50% of the aircraft's structural weight [5]. Composites, such as Carbon Fiber-Reinforced Polymers (CFRPs), offer significant advantages over traditional materials like aluminum due to their superior strength-to-weight ratios, being corrosion resistant and material tailor-ability [6]. Advances in manufacturing techniques, like laser based Automated Fiber Placement (AFP)[7] allowed for more complex structures with fewer defects to be manufactured. This enables the aerospace industry to create stronger, lightweight structures ranging from fuselage and wing sections to control surfaces and nacelles [8]. Resulting in e.g. improved fuel efficiency, lower operational costs, higher payload, reduced emissions, and longer operational lifespans [4].

Despite the advantages of composite structures, their widespread integration presents new challenges, particularly in maintenance and health management. Composites, being heterogeneous, anisotropic materials, have complex failure mechanisms - such as delamination, matrix cracking and fiber breakage - that are often hard to predict with traditional inspection methods [9].

Therefore, ensuring safety, reliability, and performance of complex structures and systems in the aerospace industry is crucial. Throughout their service life, these assets operate under extreme conditions, making them susceptible to gradual degradation and potential failure [6]. Prognostics and Health Management (PHM) methodologies are critical in maintaining these qualities by providing vital information across the lifespan of the assets. PHM is a comprehensive methodology that includes key elements such as data acquisition, degradation modeling, diagnostics, prognostics, and decision-making [10]. Data acquisition involves collecting real-time data from sensors and monitoring systems embedded in aerospace assets. The quality and quantity of this data are fundamental to the subsequent PHM processes. Degradation modeling characterizes the mechanisms through which assets deteriorate over time, typically using model-based or data-driven models. Diagnostics follows, assessing the current health state of the asset by identifying faults or anomalies that could compromise performance. Prognostics then predicts the future health of the asset, estimating the Remaining Useful Life (RUL) or time of failure [11]. This prognostic step is particularly vital as it provides insights into when maintenance or replacement may be required. Finally, decision-making utilizes diagnostic and prognostic information to support informed decisions, optimizing maintenance schedules, improving availability and reducing operational risks [12].

Among these components, prognostics is the most crucial and challenging due to its role in forecasting the future health states of assets. Especially for composite materials, where modeling the degradation process is challenging due to inherent heterogeneity and anisotropic properties of the material [13]. Moreover, uncertainties inherent to manufacturing and operational conditions make modeling the degradation of composites and predicting their RUL even more complex [14]. To fully comprehend prognostics, it has to be acknowledged that uncertainties are inherent due to its predictive nature. In modeling approaches, for example, assumptions and simplifications are made, thereby introducing uncertainty in the predictions [15].

As prognostics entails predicting the future state of an asset, uncertainties are inherently present. Therefore, it is essential to model the RUL as a random variable to account for the stochastic nature of degradation processes and data variability [11]. A reliable prognostic model is typically categorized by four key attributes: accuracy, low uncertainty, interpretability and explainability. Accuracy refers to how closely the model's mean RUL prediction aligns with the actual observed RUL. This is often evaluated using deterministic error metrics, such as Root Mean Square Error (RMSE) [16], without yet considering the uncertainty around the prediction. 'Low' uncertainty implies that the model outputs narrow confidence intervals, indicating high confidence in the predicted RUL. This is critical because wide uncertainty bounds render the predictions too ambiguous to support practical, reliable decision-making in maintenance scheduling. Interpretability relates to the degree to which a human user can understand the internal mechanics and structure of the model, for instance, understanding how input features influence predictions. Explainability focuses on how well the model's outputs can be justified or traced back to underlying physical or data-driven reasoning, making the prognostic results trustworthy and actionable.

In the context of prognostic modeling, five primary sources of uncertainty can be distinguished [17, 18]. Past uncertainty refers to incomplete knowledge about the material's initial health condition, including manufacturing-induced variability such as porosity, delaminations, or contamination. Present uncertainty arises from inaccuracies or noise in sensor measurements and diagnostic data captured during the monitoring phase. Future uncertainty is linked to unpredictable operational and environmental conditions that influence degradation over time. Model uncertainty stems from simplifications, assumptions, or limited fidelity in the mathematical degradation model. Lastly, prediction method uncertainty relates to the limitations of the algorithms or learning techniques used to estimate RUL. Each of these sources independently or jointly affect the reliability of prognostic predictions and should be managed appropriately to enable informed decision-making.

This thesis aims to present a novel approach that aids in improving the prognostic reliability for composite materials. The focus will be on addressing uncertainties stemming from manufacturing processes, labeled as "Past State" uncertainties. These uncertainties include the variability in material quality and defects like voids, delaminations and contamination. To the best of the author's knowledge, N. Eleftheroglou [18] is the first to mention this source of uncertainty, but no literature exist that presents a methodology that is able to manage this source of uncertainty. Without adequately addressing these uncertainties, predictions regarding structural integrity and lifespan often remain unreliable or conservative, which can lead to misleading assessments, resulting in inappropriate maintenance strategies and increased operational risks.

Recent advancements in Non-Destructive Testing (NDT) methods can provide high-resolution imaging capable of capturing intricate details about material integrity [19]. These techniques offer a promising solution for quantifying the quality of composite materials post-manufacturing, providing essential data to document material quality and manage past state uncertainties.

However, transforming NDT-derived data into actionable prognostic insights requires a methodology capable of accurately quantifying material quality. By leveraging data from the manufacturing process through NDT and similarity analysis, materials can be quantified on their material quality and grouped based on their internal structure. A fundamental assumption for this methodology is that it is expected that materials with similar structural quality follow a similar degradation process. This data can serve as critical prior knowledge inputs for prognostic models. Rather than training a prognostic model on the entire training set a subset of training samples will be used, aiming to reduce uncertainty bounds and make more reliable RUL predictions.

This thesis seeks to answer the question: *How can uncertainty management for past state uncertainties be realized?* It will explore a methodology that employs a similarity learning technique that specifically leverages data from the Dolphicam [20], known for its ease of use [21], and applications in

the aerospace industry [22]. Similarity metrics like; Structural Similarity Index Measure (SSIM), histogram-based and a Convolutional Neural Network (CNN) are explored to compute the data from NDT into similarity scores, serving as the prior information for prognostics models and as an approach to manage past state uncertainties.

The scope of this research is confined to aerospace-grade CFRP composites subjected to tension-tension fatigue testing conditions. The investigation specifically considers defects intentionally embedded during manufacturing to simulate variability in the material quality. Dolphicam ultrasonic imaging is utilized for defect quantification, and fatigue tests are conducted to collect data from the degradation process of the materials. The main objectives of this research include:

- Identifying and quantifying manufacturing defects in CFRP composites using Dolphicam as a Non-Destructive Inspection (NDI) tool for imaging the internal structure of the material.
- Developing and reflecting upon a similarity learning methodology that is able to observe and categorize CFRP specimen based on Dolphicam data and different types of embedded defects.
- Integrating similarity-based past state uncertainty data into a prognostic model to improve RUL predictions based on the fundamental assumption supporting this methodology.
- Finally evaluating the influence of material quality variability on RUL predictions.

This thesis is structured as follows, Chapter 2 details the Literature Review, providing an overview of existing literature on prognostics, composite material defects, similarity learning methodologies, and their application in prognostics. In this chapter a methodology is proposed and the research questions that arised from the review and the methodology are presented. In, Chapter 3, the case study created for the Thesis will be described. It includes the materials used, the process of introducing defects, NDT scanning protocols, fatigue testing setup, and data collection procedures. The data from the case study is used as an input for similarity methods to tailor them specifically for the scan results. Next, Chapter 4 will provide detailed development of the identified similarity learning techniques from the literature review, also providing the similarity scores. Chapter 5 presents the results of the RUL predictions and a prognostic performance comparison of the methodologies. Chapter 6 details and discusses the findings, placing them within the context of the research objectives. Finally, in Chapter 7 the conclusion summarizes the findings, provides the research contributions, highlights its implications, and proposes recommendations for future research.

2. Literature Review

The increasing adoption of composite materials in aerospace structures presents new challenges in the field of PHM, driven by both the intrinsic material complexities and the uncertainties inherent in prognostic modeling. Beyond their heterogeneous and anisotropic nature, composites exhibit damage behaviors that are difficult to capture, especially under variable operational and manufacturing conditions. As outlined in the introduction, the reliability of RUL predictions is strongly dependent on how effectively these uncertainties are quantified and managed. Mitigating uncertainty is therefore critical for improving prognostic reliability, supporting informed maintenance decisions, and safeguarding structural integrity. This chapter presents a comprehensive review of current PHM methodologies for composites, with particular emphasis on how they address, or overlook, key uncertainty sources. It further highlights the emerging role of past state information obtained through NDI as a means to reduce prognostic uncertainty and enhance model reliability.

This review is structured as follows: The first section delves deeper in the concept of prognostics, discussing how inherent uncertainties are typically addressed. Section 2.2 outlines traditional uncertainty management methods and introduces a novel approach. Next, Section 2.3 introduces the category of *Past State Uncertainties* and proposes a methodology for their quantification. Section 2.4 presents various uncertainty management methods, detailing how data from past state uncertainties can enhance prognostic models. The review concludes with a proposed methodology in Section 2.5 that addresses and integrates past state uncertainties in a framework that improves the reliability of RUL predictions.

2.1. Prognostics

Generally, prognostics involves predicting the RUL of assets by leveraging real-time condition monitoring data along with a degradation model. A distinction can be made between supervised and unsupervised techniques. Supervised techniques, like Machine Learning (ML), leverage data from historical and monitoring data with known labels, where the labels represent the ground truth values, which in this context are the actual RUL or failure times of assets. These models learn to map observed features directly into RUL estimates, without explicitly modeling the degradation process. Whereas unsupervised techniques do not rely on labels. Instead, unsupervised methods distinguish patterns or groupings within the dataset without requiring prior knowledge of class labels. These techniques rely solely on the intrinsic structure and relationships present in the data to identify clusters or trends, making them especially valuable when labeled failure data is unavailable or limited [23]. Both supervised and unsupervised techniques can be used to assess the time until a particular event or state is reached within the service life of the asset. In this thesis' context, "state" refers to the various failure modes of a composite structure, where each state can exhibit different types of damages, such as fiber breakage, matrix cracking, delaminations, etc., associated with a probability of being in that specific state. While the "failure state" is binary — indicating that the structure has either failed or not. The event marking the occurrence of failure is typically referred to as End of Life (EoL). Prognostic models forecast the asset's state evolution

over time, predicting when the EoL might occur within a defined future period, as well as future state values or other quantities related to the asset's health [24].

The RUL is defined as the time from the current prediction point, t_p, to the end of the asset's useful life, or failure time, denoted as the EoL. This can be expressed as:

$$\text{RUL} = \text{EoL} - t_p \quad | \quad \text{EoL} >= t_p \tag{2.1}$$

Since the remaining time after t_p is inherently stochastic, the RUL should be modeled as a random variable rather than a deterministic one [11]. Figure 2.1 illustrates the concept of RUL prediction, where $Y = \{y_1, \ldots, y_n\}$ represents the collected data samples up to the prediction point t_p. Probabilistic estimation of the degradation level at t_p can be performed through diagnostic methods, which help quantify the asset's current degradation level and associated uncertainties.

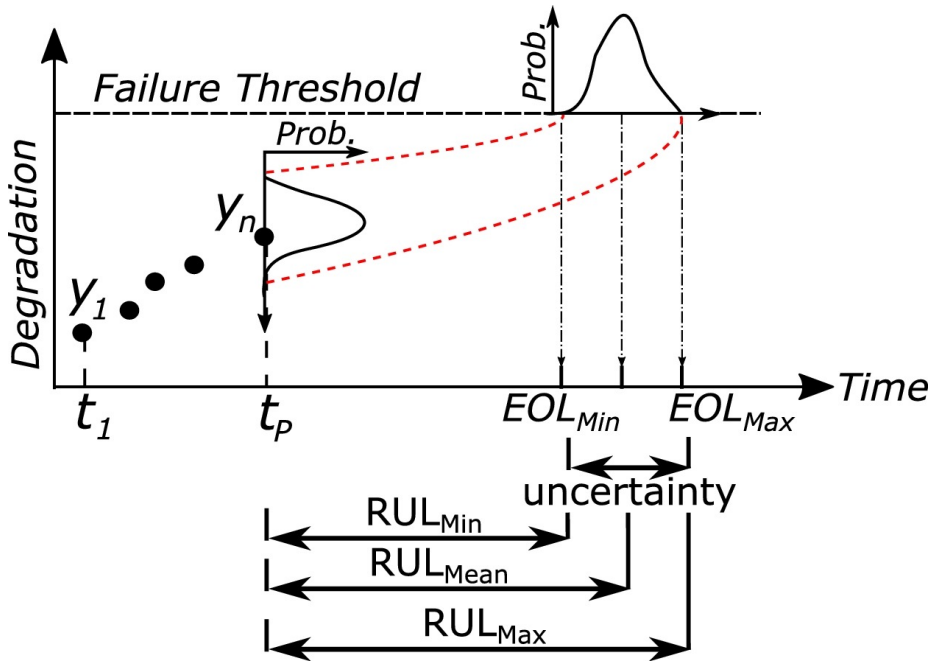


Figure 2.1.: Remaining Useful Life prediction [1].

The figure demonstrates that the initial health-state estimate at t_p is uncertain, represented by a Probability Density Function (PDF). The progression of this PDF and the distance to EoL depend on the degradation process and the failure threshold. An example of a credible confidence interval for RUL is visualized (by EOL_{Min} to EOL_{Max}), indicating the possible range of EoL values, characterizing the variability in predictions.

Generally, prognostic models can be divided into three approaches, i.e., model-based, data-driven and hybrid. The RUL predictions are the outcome of a prognostic model. Model-based approaches rely on Physics-of-Failure (PoF) models, which describe the degradation process. These models use mathematical representations, material properties and operational conditions to simulate the deterioration process over time. When limited monitoring data is available, model-based approaches are considered to be more favorable, as they have been proven effective for engineering systems when the degradation process is well understood [25]. The key advantage of model-based approaches is that once a PoF model has been developed, it can be generalized and applied to different systems that follow the same underlying degradation mechanisms. This makes model-based approaches versatile for predicting the RUL in various assets that are experiencing similar

types of wear or failure processes, such as fatigue or corrosion [26]. Data-driven approaches utilize historical data, sensor measurements, and degradation-related features to estimate RUL by learning patterns directly from the data. These methods typically rely on ML techniques to model the relationship between observed conditions and future states, without requiring explicit physical representations of the degradation process. Data driven computational techniques such as, neural networks, deep learning, and support vector machines are widely used [17, 27], especially when large amounts of sensor data are available. This makes them well-suited for complex systems where precise physical modeling is challenging or impractical [24]. Finally a hybrid approach combines aspects of both model-based and data-driven models. It integrates the physical insights from PoF with the flexibility of ML techniques. Thereby leveraging the strength from both approaches, improving prediction reliability, particularly for assets, where neither model alone can fully capture the complexity of the degradation process [28].

2.1.1. Uncertainty in prognostics

Uncertainty is inherent in prognostics due to the predictive nature of estimating the RUL of assets. Addressing this uncertainty is critical for ensuring the reliability of prognostic models. A comprehensive approach to model prognostic uncertainty involves three key steps: Uncertainty Representation (UR), Uncertainty Quantification (UQ) and Uncertainty Management (UM). These three processes are often confused and used interchangeably with one another, but are in fact distinctly different processes and should be dealt with accordingly [15].

The first step, UR, involves establishing a mathematical framework to represent, measure, and express uncertainty in an interpretable manner. This step focuses on identifying and defining the types of uncertainty present and ensuring that these uncertainties are well-understood within the context of the model. The "classical" categorization of uncertainty sources are divided in: aleatoric uncertainty—the intrinsic randomness inherent to the system (e.g., variations in material properties or operating conditions), and epistemic uncertainty, which arises from limited knowledge (e.g., model assumptions or incomplete data [29]). This provides the foundation for subsequent processes by creating a structured way to represent uncertainty consistently. The aleatoric/epistemic categorization is widely used in literature for uncertainty representation, which will be reflected upon in Section 2.2.

Building on UR, UQ measures the extent of the identified uncertainties by assigning probabilistic values to RUL predictions. This step typically uses PDFs or confidence intervals to characterize the magnitude of uncertainty. UQ is essential for understanding the reliability of RUL estimates [11].

Finally, UM builds upon the previous steps by implementing strategies to mitigate the identified uncertainties and improve prediction reliability. UM focuses on reducing uncertainty through methods such as refining models, incorporating adaptive algorithms that use real-time data [30], or enhancing data quality through better sensors or diagnostics. Additionally, UM involves assessing risk tolerance for specific scenarios and scheduling maintenance actions accordingly, whether in real-time (online) or scheduled (offline) [31]. This step is particularly valuable in critical industries like aerospace, where reducing uncertainty enhances informed decision-making in maintenance scheduling and operational cost-efficiency [32].

By clearly differentiating these three steps, this study provides a structured framework for addressing uncertainty in prognostics. While all steps are correlated, this research focuses specifically on UR and UQ and its integration with UM, aiming to reduce the impact of uncertainty and enhance the reliability of prognostic models in the aerospace industry.

2.2. Sources of Uncertainty

While the distinction between aleatoric and epistemic uncertainty is fundamental to understanding uncertainty in prognostics, it offers only a coarse categorization. Aleatoric uncertainty captures inherent randomness, whereas epistemic uncertainty reflects limitations in knowledge. However, these broad classes often fall short in guiding actionable strategies for managing uncertainty in complex prognostic applications. To enable targeted uncertainty management, a more nuanced classification is required. This section introduces a broader categorization of uncertainty divided in past, present, future, model, and prediction method uncertainty, that better aligns with the data-driven workflows used in aerospace PHM.

2.2.1. State of the art prognostics in uncertainty

It is difficult to definitively state which method "best" accounts for uncertainty in prognostics. The ideal approach depends on factors like the specific application, available data, computational resources, and desired level of reliability. As mentioned in section 2.1, three approaches are commonly used to obtain RUL predictions, i.e., model-based, data-driven and hybrid models. State-of-the-art prognostic models involve a variety of approaches and considerations, each addressing specific types and sources of uncertainty that influence the prediction of an asset' RUL. The focus will be on Data Driven Models (DDMs), as addressing each modeling method is beyond the scope of this review and would require extensive knowledge from many different fields.

DDM approaches are particularly valuable when a detailed physical model of the degradation process is unavailable, for example, when considering non-linear anisotropic materials like composites [25]. DDM approaches utilize both historical and real-time sensor data for prognostics. Where historical data forms the basis for training these models, providing the degradation model needed to predict RUL and real-time sensor data is essential for making predictions as the asset operates. These approaches utilize various probabilistic models and ML techniques to quantify uncertainty directly.

One approach, for example, involves using stochastic models, which study the evolution of random variables over time through processes such as Gaussian, Markov, and Lévy processes [33]. In [11], the authors highlight the use of Hidden Markov Models (HMMs) for their inherent ability to capture aleatoric uncertainty. When the posterior distribution has a closed form, meaning it can be represented by an exact mathematical formula that is explicitly solvable without iterative or approximate methods, HMMs can directly capture aleatoric uncertainty, including its quantification through the prognostic measure. Additionally, HMMs can incorporate epistemic uncertainty, as seen in Generalized Hidden Markov Models (GHMMs) [34], though this often requires computationally intensive techniques like sensitivity analysis or the use of imprecise probabilities. Other adaptations and extensions of HMMs, such as Similarity Learning Hidden Semi-Markov Models (SLHSMMs) [35] and Adaptive Non-Homogeneous Hidden Semi-Markov Models (ANHHSMMs) [25], excel at handling unforeseen data and unexpected phenomena, respectively, demonstrating the versatility of HMMs. On top, the Similarity Learning feature of the SLHSMM allows the model to manage uncertainties.

While stochastic models generally require less training data due to their reliance on explicit physical or statistical degradation representations, ML techniques often need larger datasets to effectively learn model parameters and uncertainty distributions. In data-driven prognostic models, recent developments have focused on improving RUL predictions by addressing both aleatoric and epistemic uncertainties. Two notable ML techniques include Bayesian Neural Networks (BNNs) and the Long-Short-Term Memory (LSTM) model.

BNNs offer a probabilistic framework in estimating epistemic uncertainty. This is achieved by introducing a prior distribution over model parameters, which is then updated with training data to produce a posterior distribution. Moreover, BNNs naturally incorporate aleatoric uncertainty through their probabilistic approach, enabling them to include the uncertainty stemming from limited data availability or gaps in the model structure. This is particularly beneficial in fields like aerospace, where operating conditions vary widely and large datasets for training, such as from full-scale crash testing, are extremely costly and logistically challenging to obtain [36]. However, the main drawback is that BNNs have convergence issues. The BNN approach involves treating each parameter within the neural network as having their own probability distribution. This significantly increases the complexity of the model, complicating the training process and hindering convergence. This aspect of BNNs limits their practical application.

On the other hand, LSTM networks have shown success in capturing temporal dependencies within data, making them suitable for time-series-based prognostic tasks. Specifically due to their accuracy of the mean RUL predictions [37]. While LSTMs are typically deterministic, uncertainty can be introduced by using Monte Carlo dropout, creating an ensemble approach that produces multiple RUL predictions to generate a credible interval for the predictions. This technique is useful when handling epistemic uncertainty in scenarios where noise and variability in sensor data affect the accuracy of RUL predictions [11]. However, Monte Carlo dropout only provides an approximation of the uncertainty, more specifically a profile of the epistemic uncertainty [11, 38]. Therefore limiting the reliability of the predictions.

2.2.2. Epistemic and aleatoric

In most prognostic practices, uncertainty is typically categorized into two types: aleatoric and epistemic uncertainty. However, in many prognostic papers, the assessment of uncertainty is inconsistent. Some studies overlook uncertainty entirely, presenting RUL predictions without acknowledging the inherent variability in their models. For instance, in [39], the authors utilize Artificial Neural Networks (ANNs) to predict RUL based solely on historical failure data. They present point estimates without explicitly modeling uncertainty. Mean Square Error (MSE) is commonly used as an overall prognostic performance measure of model accuracy, treating RUL as a deterministic value. While MSE effectively measures average prediction error, it does not capture the confidence intervals or probability distributions associated with individual predictions. Similarly, [40] and [41], give RUL predictions of assets and use RMSE as a measure to model accuracy but don't include or mention uncertainty at all.

This raises an important issue: although metrics like MSE and RMSE are appropriate for evaluating deterministic predictions, they do not characterize how well a model accounts for or quantifies uncertainty. Consequently, many prognostic studies that rely solely on these metrics do not provide insights into the reliability or confidence of their RUL predictions [39, 40, 41, 42]. To properly evaluate models that incorporate uncertainty, additional or alternative metrics should be employed to assess both accuracy and uncertainty quantification.

Even in cases where uncertainty is categorized into aleatoric and epistemic, researchers often fail to specify which type of uncertainty is being considered. For example, the authors in [37], presents a structured Bayesian approach to RUL prediction that accounts for both aleatoric and epistemic uncertainties, supporting the reliability of predictions by providing credible intervals. However, the complexity and computational demands of the model pose significant challenges for practical implementation. Additionally, while the methodology separates different uncertainty sources, the results do not explicitly differentiate between them. This generalized categorization under "epistemic" uncertainty does not effectively progress uncertainty management, as it groups all

reducible uncertainty sources into a broad category without offering insights on how to leverage data to mitigate their impact on RUL predictions.

Furthermore, even when both aleatoric and epistemic uncertainties are considered, the combined uncertainty could result in overly broad confidence intervals. While these wide intervals technically capture the range of possible outcomes, they may not be practical for decision-making. Firstly, because they are too uncertain, and secondly, because they do not provide interpretation of how to leverage data to better characterize the sources of uncertainty and reduce their impact on RUL predictions. In such cases, the RUL estimates are so uncertain that it becomes difficult to set effective maintenance schedules or plan interventions with confidence. This highlights the challenge in prognostics, balancing comprehensive uncertainty quantification with practical usability, ensuring that predictions are not only accurate but also reliable.

2.2.3. Broader categorization of uncertainty

While the previously given classical categorization provides a foundational framework for uncertainty representation, it falls short in facilitating effective uncertainty management. Given that time is an inherently random variable in prognostics, the uncertainties associated with the different stages of the prediction process cannot be adequately represented by the binary aleatoric/epistemic categorization. Prognostics, inherently linked to the time domain, demand a categorization that adequately reflects this time aspect. The work of S. Sankararaman [17] and later extended by N. Eleftheroglou [18], have proposed a more nuanced breakdown of uncertainty in prognostics. These proposals differentiate between uncertainties related to time and those related to the model. This refined categorization better integrates the time dynamics of prognostics, offering a more targeted approach to managing the uncertainty sources:

1. **Past Uncertainty:** Arises from the manufacturing or assembly processes, including material quality. Issues stemming from these early stages can affect future predictions, and while they may be partly addressed by characterizing manufacturing processes, they often contribute residual uncertainty to current predictions as they propagate through time.
2. **Present Uncertainty:** Relates to the lack of precise knowledge about the current degradation level of the system. For instance, sensor inaccuracies or noise can prevent exact determination of the system's current condition. This uncertainty can be minimized with advanced sensing technology but is rarely eliminated entirely.
3. **Future Uncertainty:** The future operating conditions, environmental factors, and load profiles are inherently unknown and can vary widely. This category, often the most challenging to handle, introduces significant uncertainty into long-term predictions, as it is impossible to foresee every change in operational conditions.
4. **Model Uncertainty:** This arises from limitations in the chosen degradation model, including assumptions made about degradation processes and simplifications in the model structure. Both model-based and data-driven models are subject to model uncertainty due to their underlying assumptions and limited capacity to represent all degradation mechanisms accurately.
5. **Prediction Method Uncertainty:** Related to the method used for making predictions, this includes errors introduced by the specific algorithms or techniques applied. In stochastic models this uncertainty is introduced due to the needed prognostic measure. In ML techniques, for example, the RUL is directly obtained via the model, making prediction method and model uncertainty a single source.

The proposed categorization, by Sankararaman [17] and Eleftheroglou [18], of uncertainty provides a more nuanced framework for managing the complexities inherent in advanced prognostic models. By distinguishing between past, present, and future uncertainties, as well as model and prediction method uncertainties, this framework acknowledges the multifaceted nature of uncertainty in prognostics. This detailed breakdown allows for a more targeted approach in uncertainty management, enabling decision-makers to address specific sources of uncertainty at different stages of the prediction process. Furthermore, it facilitates the development of reliable prognostic models by providing a comprehensive understanding of the various factors that influence the prediction process.

Among these five categories, past state uncertainty is an inherent yet often overlooked aspect of prognostics. Unlike present or future uncertainties, which are typically addressed through sensing or probabilistic modeling, past state uncertainty originates early in the asset's life-cycle, during manufacturing or assembly, and can persist unmitigated throughout the prediction horizon. Addressing this foundational uncertainty can therefore improve the initial reliability and overall trustworthiness of RUL predictions.

2.3. Identifying the paste state uncertainty

This section focuses specifically on past state uncertainty, a foundational yet underrepresented source of uncertainty in prognostic modeling. Unlike other uncertainty categories that emerge during operation, past state uncertainty stems from inherent variability in material quality, manufacturing processes, and assembly conditions. As it arises before degradation begins, in terms of operation, it defines the initial conditions upon which all RUL predictions are based. Despite its early influence, this uncertainty is rarely characterized in current literature, creating both a significant research gap and an promising opportunity for managing this source of uncertainty in prognostics. Especially when models transition from population-based predictions to asset-specific estimates, addressing past state uncertainty becomes increasingly important for enhancing prediction accuracy and confidence.

To address past state uncertainty, it is crucial to focus on the specific facets included in this category. As previously noted, past state uncertainty arises from manufacturing or assembly processes and is embedded in the material quality. While past and present uncertainties may seem intertwined, they are distinct in nature. Past state uncertainty reflects the variability introduced during manufacture, while present uncertainty relates to the current degradation level of the asset, often influenced by sensor noise or real-time measurement inaccuracies.

In the context of RUL prediction, the process begins with an initial confidence interval that reflects the uncertainty at the start of the prediction. The variability in material quality, particularly for composite materials with anisotropic properties, can significantly impact the initial confidence interval. Composite materials often exhibit varying quality levels throughout a structure, introducing inconsistencies that remain throughout the RUL prediction process. Providing detailed and accurate material quality data and how they impact system degradation can help reduce the initial confidence interval, thereby improving the model's overall reliability.

Figure 2.2 visualizes RUL prediction with normalized time ($t = 0$ to $t = 1$) and RUL ($RUL = 1$ to $RUL = 0$), representing the start to EoL. The blue interval ("Without UM") reflects predictions without UM, starting with a wider range and converging over time. The yellow interval ("Including UM") incorporates additional data, such as material quality characterization, resulting in a narrower initial range and improved reliability. Both intervals are hypothetically derived from the same model, handling other uncertainties identically. While conceptual, Figure 2.2 demonstrates the potential benefits of addressing past uncertainties in prognostic modeling.

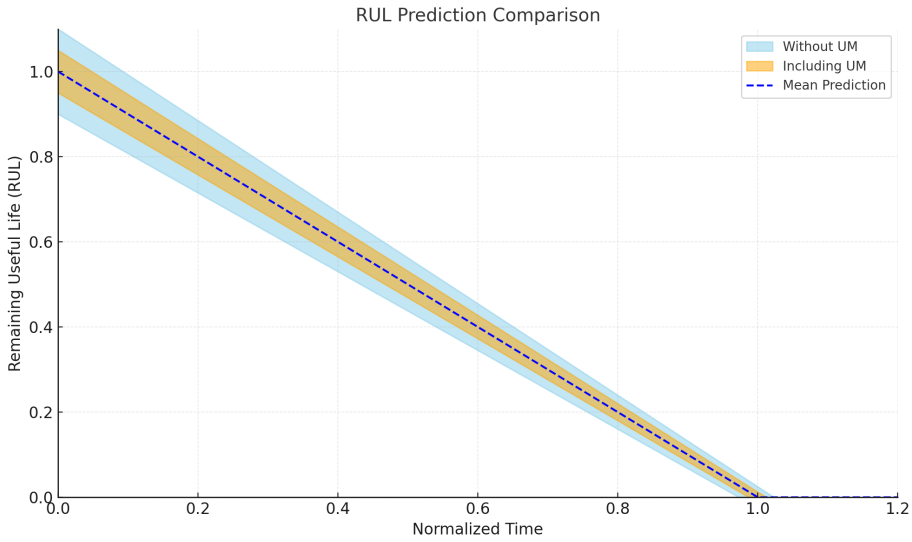


Figure 2.2.: Conceptual comparison of RUL predictions with and without UM

The next subsection explains the importance of NDT techniques in addressing past state uncertainty. Effective uncertainty management for past state uncertainties requires data that can characterize this specific source of uncertainty. Such data can be extracted using NDT methods, which provide detailed insights into material characteristics, defects, and anomalies. These insights serve as the foundation for transitioning from population-wide asset behavior models to predictions tailored to individual assets.

To enable this transition, similarity research is essential. Similarity research categorizes subsets of assets with shared material properties or quality levels, providing a more refined understanding of individual asset behavior. By grouping assets based on their similarities it becomes possible to improve the reliability of the predictions and reduce the uncertainty linked to manufacturing or assembly processes. Section 2.4 that follows introduces a few potential methods for performing similarity research.

2.3.1. Dolphicam as NDT technique for material quality inspection

NDT of composite materials plays a crucial role in high-demand industries such as aerospace. One of the challenges in using NDT methods for composites is the complex nature of these materials. They display anisotropic behavior that can complicate the detection of internal flaws such as delaminations, porosity, fiber misalignment and voids. A wide range of NDT techniques exists for assessing composite materials, including ultrasonic testing, Acoustic Emission (AE), Dolphicamera, infrared thermography, X-ray imaging, Digital Image Correlation (DIC) [19].

From these techniques the Dolphicam (by DolphiTech [20]) system was chosen as a suitable tool due to its high-resolution imaging capabilities, local availability and ease of use [21]. Specifically, the Dolphicam2 has received certifications from leading aircraft manufacturers, including Boeing and Airbus. Boeing has authorized specific NDT procedures for the 787 Dreamliner, while Airbus has accepted the Dolphicam2 for inspecting all its composite assets [22]. Thereby proving it to be a reliable and highly regarded NDT technique in the aerospace industry.

The Dolphicam2 is a matrix array ultrasonic technology system, that uses a matrix array of 16,384 transducers to perform high-resolution scans of composite materials. In operation, the system emits ultrasonic waves into the material, which propagate through the layers of the composite. The waves interact with the material's internal features, such as fiber orientation, resin distribution, and any internal anomalies, and are reflected back to the system's transducers [43]. These reflected waves provide critical data about the internal structure of the material, such as the thickness of layers, the presence of voids, or inconsistencies in fiber alignment. The Dolphicam2 captures this reflected data in the form of high-resolution images through A-scan, B-scan, and C-scan modes, which correspond to one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) images, respectively:

1. A-scan mode provides a simple profile of the reflected signals, showing the depth and intensity of the echoes along a specific line, useful for pinpointing the location of features or potential anomalies.
2. B-scan offers a cross-sectional view, helping to visualize internal layers and how they interact with the ultrasonic waves, revealing structural integrity.
3. C-scan generates a top-down, full-surface view, allowing for the inspection of material uniformity and detecting areas that may have quality issues, such as inconsistencies in the laminate structure.

Figure 2.3 and Figure 2.4 showcase a case study of barely visible impact damage on a CFRP aircraft skin [44]. A C-scan has been made for the material in which the damage throughout the different layers of composite has become visible.

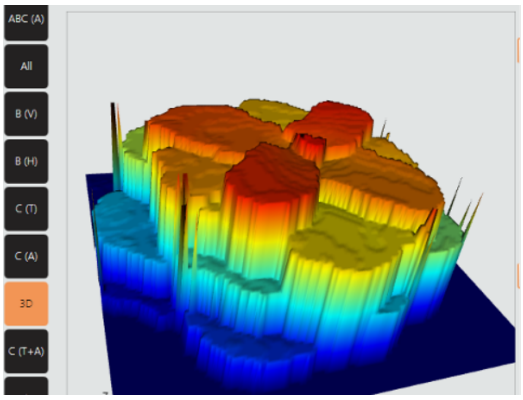


Figure 2.3.: Impact damage on CFRP, 3D view.

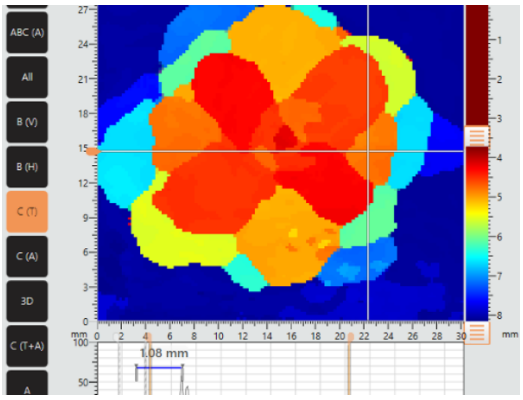


Figure 2.4.: Impact damage on CFRP, 2D view.

The output of the Dolphicam2 is a highly detailed visual representation of the material's internal state, providing an image that can "grade" the quality of the material. This output is valuable in assessing uniformity and detecting any inconsistencies in the composite's structure before it enters the service life cycle. Moreover, the use of image-based data generated by the Dolphicam allows for consistent and reproducible evaluations, making it easy to compare results across different inspection samples and over time. This consistency is key in reducing past uncertainties related to material quality.

2.4. Similarity research

Similarity learning methods offer a novel approach for managing past state uncertainty in prognostics by identifying and categorizing patterns within data that reflect differences in material quality [35]. In the context of composite materials, where internal structure and manufacturing variability can affect degradation behavior, these methods enable the grouping of similar materials based on features extracted from imaging data. This process supports the development of category-specific prognostic models, reducing the reliance on generalized assumptions and improving the reliability of RUL predictions.

To support this approach, images generated by the Dolphicam2 can be used to assess internal structural features such as fiber alignment, porosity, void content and embedded defects. Instead of averaging material quality across an entire composite laminate, specimens with similar characteristics can be grouped and analyzed independently. This enables a more targeted uncertainty management strategy. Several methods are commonly used for performing similarity analysis on image data. These include: Convolutional Neural Networks (CNNs) for learning hierarchical image features[45], Histogram-based similarity metrics for comparing pixel intensity distributions [46], and Structural Similarity Index Metric (SSIM)[47] and its extension Multiscale Structural Similarity Index Measure (MSSSIM)[48] for evaluating an image' structural information. The following subsections provide a theoretical overview of each method and their relevance to similarity-based uncertainty management.

2.4.1. Convolutional Neural Networks

CNNs are a class of deep learning models particularly well-suited for processing and analyzing image data. CNNs are composed of multiple layers, including the input layer, convolutional layers, pooling layers and fully connected layers [49].

- The input layer hold a set of pixels which represent the image.
- Convolutional layers apply learnable filters (kernels) across the input images to detect local patterns such as edges, textures or shapes. These filters generate feature maps that capture spatial hierarchies within the data, allowing deeper layers to recognize increasingly abstract features.
- Pooling layers reduce the spatial dimensions of the feature maps while preserving the most prominent information, making the network more computationally efficient.
- Fully connected layers at the end of the network map the extracted features to output predictions or classifications.

The learning process in CNNs is driven by activation functions and loss functions. Activation functions, e.g. the Rectified Linear Unit (ReLu) $y = max(x, 0)$, are applied after each convolutional layer to introduce non-linearity by zeroing out negative values. This enables the network to learn more complex patterns beyond simple linear transformations. The loss function quantifies the difference between the predicted output and the ground truth. For classification tasks, the cross-entropy loss is commonly used, for similarity learning or regression tasks, for example, contrastive loss, triplet loss or mean squared error can be used depending on the objective [50].

In the context of similarity learning for material quality, CNNs can be trained on a labeled dataset of composite images where different categories correspond to known material properties or defect types. The network learns to extract features that best separate these categories. Once trained,

it can embed new test samples into a feature space and classify them based on similarity to the learned categories.

CNNs are advantageous because they can learn abstract and complex features from raw image data. However, they typically require large datasets to generalize well. In cases where data is limited, overfitting can occur, where the network memorizes the training data rather than learning general patterns [49]. Moreover, CNNs are often viewed as "black box" models due to their layered architecture and lack of interpretability [51], posing challenges in interpreting how decisions are made or to explain the model's reasoning.

2.4.2. Histogram-based Similarity

Histogram-based similarity methods assess images based on the distribution of pixel intensities [46]. Each image is represented by a histogram that counts the frequency of pixel values across a defined range. Similarity between images is computed by comparing these histograms using metrics such as histogram intersection, correlation or Chi-squared distance [52].

This method is computationally lightweight and effective for capturing global brightness patterns, making it a fast and intuitive approach for grouping images. In the case of composite scans, histograms can reflect overall material density, for example, or the presence of inconsistent texture regions.

However, histogram-based methods are limited by their lack of spatial awareness, meaning, they treat the image as a bag of pixels without considering the arrangement or position of features. This can lead to incorrect similarity assessments if two structurally different images share similar intensity distributions [53]. Moreover, histograms are sensitive to changes in lighting, angle, or image scaling, making preprocessing and standardization critical.

2.4.3. Structural Similarity Index Metric

The SSIM is a metric designed to evaluate image similarity based on structural information. It combines three components, i.e., luminance, contrast and structural comparison, into a single similarity score between 0 and 1, where a score of 1 indicates perfect similarity [54].

In the context of analyzing composite scan outputs, SSIM may be a useful method, especially when the goal is to evaluate structural features like consistent fiber alignment or uniform resin distribution. By computing the SSIM between each test image and the training set, it is possible to group images based on their structural similarity.

Although SSIM is widely used for assessing image quality it does have some limitations. SSIM assumes that images are perfectly aligned and of equal size and scale, which is often impractical in real-world scenarios. Additionally, it does not effectively handle color images since it primarily evaluates luminance changes, overlooking chrominance variations [47].

To overcome the limitations of SSIM, particularly its ineffectiveness with color images, an extension known as MSSSIM can be utilized. MSSSIM extends the capabilities of SSIM by applying the metric at multiple scales, which helps in capturing both the fine details and the broader aspects of an image's structure [48]. This makes MSSSIM more resilient to changes in image resolution and more effective in capturing the perceptual quality of color variations.

2.4.4. Similarity-Driven Prognostics

Similarity-driven prognostics categorizes composite structures based on their internal structural characteristics. By quantitatively evaluating structural quality through carefully designed similarity metrics, composites with comparable initial states can be effectively grouped. This enables the development of targeted prognostic models, under the assumption that composites exhibiting similar initial structural conditions will follow analogous degradation trajectories under operational loading.

A significant advancement in similarity-driven prognostics is the introduction of the SLHSMM [35]. The SLHSMM enhances the Non-Homogeneous Hidden Semi-Markov Model (NHHSMM) [14], by integrating a dynamic similarity-informed weighting mechanism into its degradation modeling process. Specifically, the model employs a discrete similarity learning vector, which adaptively modifies degradation and observation parameters based on the degree of similarity between the test sample and available training degradation trajectories [55]. By assigning greater weight to similar performing training cases, the SLHSMM effectively improves prognostic reliability, allowing the model to better capture relevant degradation paths and reducing uncertainty in RUL predictions, particularly when encountering structural outliers or unexpected degradation behaviors [35, 55].

Eleftheroglou et al. [35] demonstrated the practical applicability of similarity-driven prognostics through experimental validation on aerospace-grade composite specimens. Their results indicate substantial improvements in RUL prediction reliability, reflected by reduced uncertainty intervals and enhanced model adaptability to unforeseen degradation events. Incorporating trajectory similarity features allowed the prognostic model to integrate performance insights directly into the prediction framework, thereby aligning predicted trajectories more closely with the actual observed degradation paths.

Extending upon these principles, this thesis explicitly addresses uncertainties originating from manufacturing-induced variability, such as embedded defects and inconsistencies in material quality, by quantifying past state conditions through advanced NDI methods. Incorporating these past state similarity features into the prognostic framework further strengthens prediction reliability, demonstrating the added value of similarity-informed strategies beyond conventional prognostic methods.

In summary, similarity-driven prognostics, particularly through models like the SLHSMM, enhances predictive maintenance strategies by allowing for uncertainty management. This targeted approach to uncertainty management facilitates more reliable prognostic outputs, enabling informed operational decision-making and potentially reducing maintenance costs and operational risk in high-performance aerospace applications.

2.5. Proposed methodology

This literature review has highlighted a significant gap in current prognostic techniques: the lack of explicit uncertainty representation and management of past state uncertainties. While various methods exist to quantify uncertainty during service, for example through sensing or probabilistic models, few, if any, integrate pre-service material quality data into the prognostic process. To the best of the authors knowledge no existing studies address the influence of past state uncertainties in the RUL predictions.

To address this gap, this thesis proposes a novel methodology that explicitly manages past state uncertainty in RUL prognostics through similarity learning. The Dolphicam2, a high-resolution

ultrasonic imaging tool validated in the aerospace industry, was leveraged to obtain detailed C-scan data of composite structures prior to service. These scans provide insight into the internal quality of the material, such as delaminations, fiber misalignment, voids and embedded defects. By quantifying these differences through similarity metrics, this information can be translated into meaningful input for prognostic models.

As there is currently no open-access database of Dolphicam scans, and given the unique data characteristics of this tool, it was necessary to develop and tailor the similarity learning methodology specifically for Dolphicam outputs. This decision is embedded in the research question and is also reflected in the structure of this thesis. Since the similarity method must be adapted to the specifics of the experimental data, the case study precedes the full method development, allowing the methodology to evolve in response to actual scan results. This sequence is also visible in the proposed workflow diagram, where material manufacturing and scanning precede the development of the similarity method.

The proposed methodology is visualized in Figure 2.5. This workflow includes the manufacturing of defect embedded composites, data acquisition, feature extraction, similarity-based grouping and stochastic prognostic modeling.

The process begin with the collection of two complementary datasets. The first consists of pre-service NDI scans, obtained from the Dolphicam2, capturing manufacturing-induced variability in material quality. The second involves in-service strain measurements, acquired via DIC, which provide condition monitoring data during fatigue testing. These two data streams are first processed independently: similarity features are extracted from the NDI scans and prognostic features from the DIC data.

Similarity features are derived using a range of methods including histogram-based metrics, SSIM, and deep learning embeddings from CNNs, as discussed in Section 2.4. These features enable clustering of specimens based on their internal material quality. Instead of training the prognostic model on the entire DIC Feature Dataset (DICF), the model is trained on subsets of data from specimens grouped by similarity. This results in a Similarity-Informed DIC Feature Subset (SIM-DICF), which allows the model to better account for variability in initial health state; something conventional approaches typically overlook.

Once these datasets are established, they serve as inputs to a stochastic prognostic model. This model estimates its parameters (θ_{Baseline} and $\theta_{\text{Similarity}}$) and computes probabilistic RUL predictions with associated confidence intervals. The model itself, a SLHSMM, was developed by the supervising research group. Since modeling the full stochastic process is beyond the scope of this thesis, the RUL computations were carried out by the research group. For further information about the implementation of the prognostic model, readers are referred to [35, 56].

Figure 2.5 illustrates the full process. The green branch highlights the thesis' main contribution: a similarity analysis framework for managing past state uncertainty leveraging NDI data. The blue branch captures the experimental campaign, including fatigue testing and DIC feature extraction. The orange branch shows the integration of these datasets into a stochastic prognostic model, resulting in both baseline ($\text{Rul}_{\text{Baseline}}$) and similarity-informed ($\text{Rul}_{\text{Similarity}}$) predictions.

This framework introduces a structured and reproducible method for managing past state uncertainties in prognostic predictions models. By linking test samples with training data from structurally similar specimens, it supports a more tailored and relevant training process. This approach is expected to improve the reliability of RUL predictions and reduce initial uncertainty, ultimately enabling more effective maintenance scheduling in aerospace and other high-performance applications.

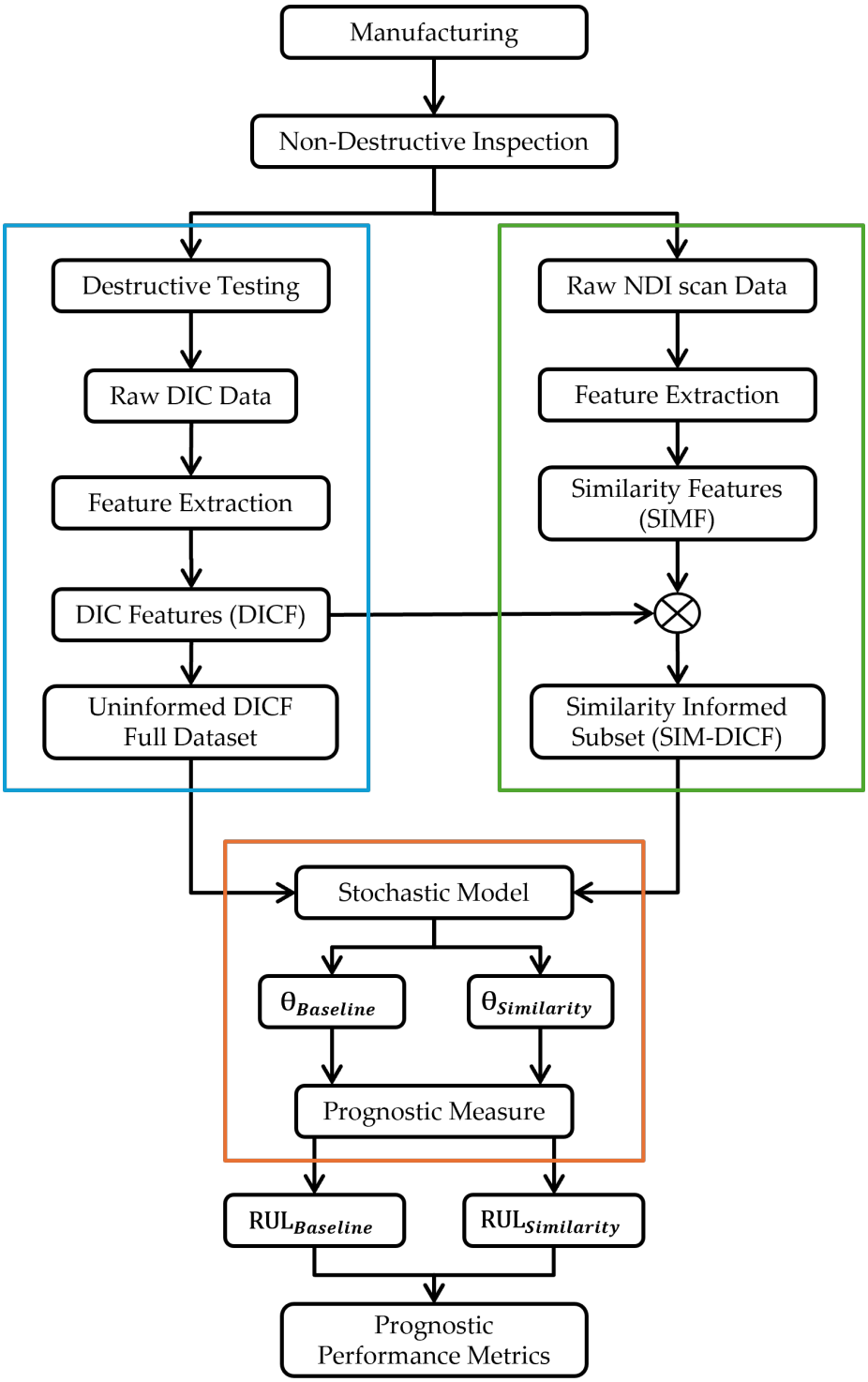


Figure 2.5.: Workflow of the proposed methodology.

2.5.1. Research Questions

From the literature review and proposed methodology, the thesis formulates the main research question:

How can uncertainty management for past state uncertainties be achieved through similarity learning techniques and Dolphicam data to improve reliability of Remaining Useful Life (RUL) predictions for composite materials?

This question is supported by the following sub-questions, which collectively guide the investigation and evaluation of the methodology:

1. What role does material quality variability play in RUL predictions, and how can Dolphicam data be used to quantify and address this variability?
2. What similarity method is best suited for leveraging Dolphicam data to group composite specimens accurately?
3. How does managing past state uncertainties in prognostic models affect the reliability of RUL predictions for composite materials?

To answer these questions, a case study is presented in Chapter 3 involving the fabrication of composite laminates with embedded defects and the acquisition of both Dolphicam and DIC data. These experimental results are used to tailor and validate the methodology under realistic conditions. This is followed by Chapter 4, which presents the detailed development and evaluation of the similarity-based methodology in detail.

3. Case Study

To validate the proposed similarity-informed methodology for managing past state uncertainties in RUL prediction, this chapter presents a structured experimental campaign. The methodology, introduced in Section 2.5, relies on the assumption that internal structural variations, stemming from manufacturing inconsistencies such as voids, contamination, or fiber misalignment, significantly influence the fatigue behavior of composite materials. By capturing these variations using NDI data and integrating them into the prognostic process, more tailored and reliable RUL predictions can be achieved.

This chapter is organized to reflect the experimental workflow used to develop and validate the proposed similarity-informed prognostic framework. Section 3.1 begins by detailing the material selection and fabrication process of the CFRP laminates, including the techniques used to embed controlled defects that simulate manufacturing variability. Following this, Section 3.2 outlines the specimen preparation steps such as machining, drilling of open holes, and surface preparation for DIC, which are essential for enabling consistent fatigue testing and strain field acquisition. Section 3.3 describes the ultrasonic NDI process using the Dolphicam2 system, which captures amplitude-based C-scan images that characterize the internal structural quality of each specimen. Finally, Section 3.4 presents the fatigue test setup, loading protocol, and observed degradation behavior, including the failure definition used to generate the RUL labels for model evaluation. Together, these procedures generate the image and strain datasets required for implementing and validating the similarity-informed modeling workflow introduced in Chapter 4.

3.1. Material and Manufacturing

This section details the composite manufacturing procedure used to introduce controlled variability into the dataset. Four CFRP laminates were fabricated using unidirectional prepreg with $\pm 45°$ layups. Each laminate included a specific type of embedded defect: void patches, oil contamination, or a combination, to simulate deviations in internal quality. The defect placement, curing cycle, and lay-up methods were documented to ensure traceability. This structural variability was later categorized via ultrasonic imaging and used for grouping specimens based on internal similarity.

3.1.1. Prepreg Handling and Lay-up

First, a high-quality structural DeltaPreg DT120 unidirectional prepreg roll was conditioned to room temperature for at least 6 hours to eliminate thermal gradients and ensure uniform resin viscosity. Plies were cut to 424 x 424 mm with fiber orientation of $\pm 45°$, using a CNC ply-cutter to guarantee dimensional and angle consistency. An aluminum mold was degreased and a release agent was applied, ensuring non-destructive demolding after curing. Eighteen plies were then hand-laid on the aluminum tool to the target sequence $[[+45, -45]_4 + 45]_s$. After each three-ply increment, the stack was vacuum-debulked for 9 min (3 min ply$^-$1), mitigating trapped air and waviness. During lay-up, defects were randomly distributed and embedded between

specific layers. Perforated release film was used to simulate localized porosity/delamination and silicon oil to mimic contamination. In total four laminates were fabricated. The summary of the embedded defects is visible in Table 3.1, where ply count is from tool side. Note that during the lay-up of laminate 4, an unintended defect occurred: layer 17 was oriented at +45°, instead of -45°. To recover global stacking symmetry, ply 18 was subsequently placed at -45°. Even though this unintended defect occurred, it still serves the purpose of the experimental campaign, as the goal is to account for manufacturing defects.

Table 3.1.: Defect configurations per laminate

Laminate	Release Film Patch	Silicon Oil Droplets
1	-	-
2	(10x10 mm) between plies 4-5 and 12-13	-
3	-	(0.1mL) droplets between plies 5-6 (0.1mL) droplets and fingerprints between 15-16
4	(10x10 mm) between plies 9-10 (7x10 mm) between plies 16-17	(0.1mL) droplets between plies 3-4 (0.1mL) droplets between plies 9-10

3.1.2. Vacuum Bagging and Autoclave Cure

After lay-up, a peel-ply was placed directly on the laminate surface, followed by a perforated release film. A breather was placed on top to ensure uniform vacuum distribution. Then, a vacuum valve and bag were applied and sealed to the tool using Tack-Tape. A vacuum was drawn and tested for any leakages. A schematic representation of the vacuum bagging is visible in Figure 3.1. Once the setup was properly sealed, the autoclave cure cycle was initiated. The laminates were cured under the manufacturer-recommended DT120 cycle: heat-up at $2°C\ \text{min}^{-1}$ to 120°C, hold for 1.5h at 6 bar autoclave pressure, then cool to 25°C at $2°C\ \text{min}^{-1}$ while maintaining vacuum. The entire bagging and cure sequence was repeated for each laminate. For comprehensive details on material selection, defect embedding, lay-up, and curing process, refer to Appendix A.

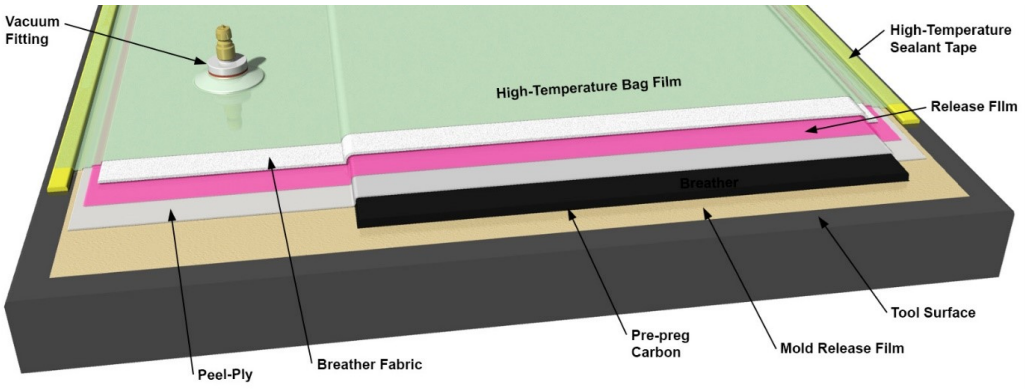


Figure 3.1.: Schematic representation of vacuum bag process [2].

3.2. Tooling of Laminates for Testing Campaign

After curing, each laminate was processed to produce individual specimens, including precision drilling of central holes and DIC surface preparation using a high-contrast speckle pattern. These steps are essential to ensure consistent geometry, controlled damage localization, and accurate full-field strain measurements during fatigue testing.

3.2.1. Cutting and Labeling

Open-hole specimens with dimensions of 400×45 mm and a central hole diameter of 10 mm were prepared for fatigue testing. Each cured laminates measured approximately $428 \times 428 \times 3.75$ mm, with the slight increase in dimensions due to resin and fiber bleed during autoclaving. Specimens were cut using a water-cooled CompCut ACS 600 to ensure consistent dimensions and clean edges. Each of the four laminates yielded eight specimens, which were immediately labeled to prevent misidentification. The labeling convention followed the format *SampleXY*, where X denotes the laminate number [1-4] and Y the specimen number [1-8]. For example, *Sample46* refers to the sixth specimen from laminate four, contains mixed defects. Figures 3.2 and 3.3 depict a subset of the labeled specimens.

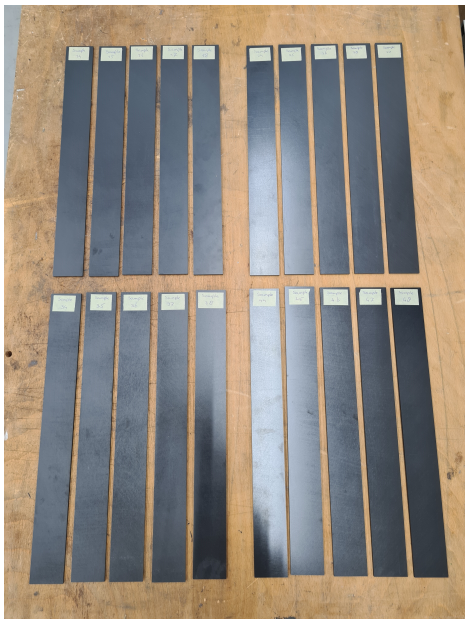


Figure 3.2.: Labeled samples, per laminate, from smooth side.

Figure 3.3.: Samples visible from textured surface

3.2.2. Drilling

After cutting, each specimen was drilled at the center using a Wörner pillar drill with a composite-specific bit, operated at a low speed (500 RPM) to minimize thermal damage, visualized in Figure 3.4. Hole locations were pre-marked, and drilling was done incrementally, approximately one ply at a time, to reduce heat buildup and delamination. Lubricant was applied to further limit

tool wear and friction. Despite precautions, minor delaminations and loose fibers were occasionally observed at the exit side, particularly through the final ply, as shown in Figure 3.5. To reduce surface damage on the measurement side and ensure optimal conditions for DIC, all holes were drilled from the textured surface.

Figure 3.4.: Manually drilling holes.

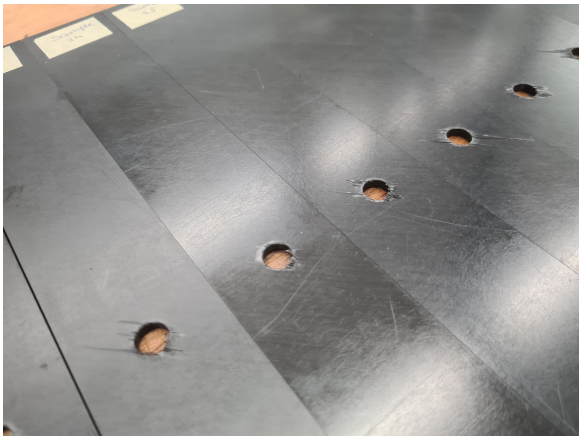


Figure 3.5.: Small delaminations and loose fibers on exit-side of drilling.

3.2.3. Speckle Painting

The textured surface created by the peel ply during curing was selected for DIC imaging due to its diffuse reflectivity and suitability for speckle patterning [57]. This surface minimizes glare and ensures consistent contrast. First, a double layer industrial-grade white base coat was applied to slightly smooth the surface while maintaining texture. After drying, a high-contrast random black speckle pattern was sprayed on top, enabling accurate strain tracking during DIC. Figures 3.6 and 3.7 show the prepared DIC surface. The outer 50 mm at both ends of each specimen were masked during painting to leave uncoated grip areas, to reduce the risk of slippage during fatigue testing.

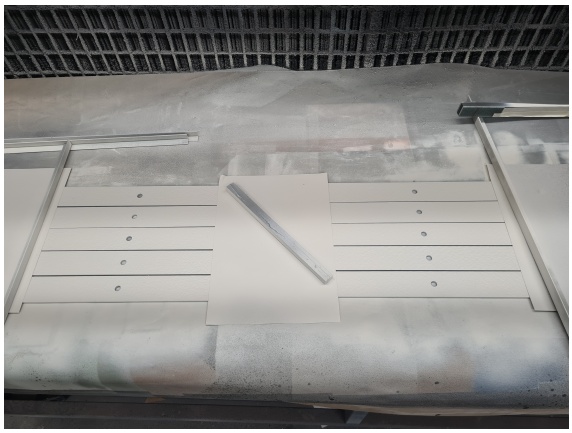


Figure 3.6.: Two-layer industrial-grade white paint coating.

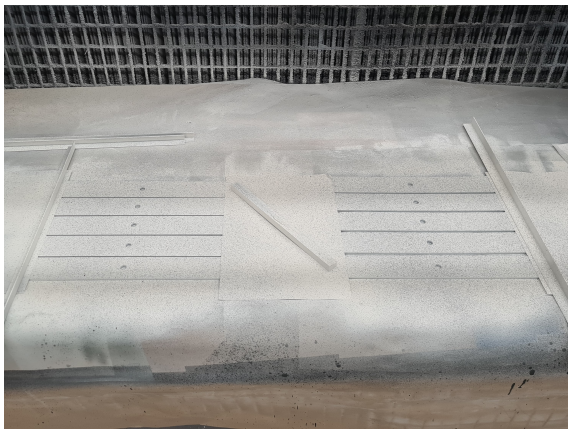


Figure 3.7.: Randomly distributed black speckle pattern.

3.3. Ultrasonic C-Scan Inspection Using Dolphicam2

Prior to fatigue testing, all specimens were inspected using the Dolphicam2 NDI system to capture internal structural features and embedded defects. The system uses a high-resolution matrix Transducer-Receiver Module (TRM), in this case equipped with an 8 MHz broadband transducer, suited for thin composite laminates.

3.3.1. Scanning Procedure and Specimen Configuration

Each 400 mm specimen was systematically scanned in eight predefined locations using a 30×30 mm aperture. Although the TRM has a nominal active area of 32×32 mm, a reduced transducer aperture size was chosen to avoid high amplitude distortions observed at the edges. The scan naming followed the format $SampleXY\_Z$, where X is laminate number [1–4], Y the specimen number [1–8], and Z the scan location [1–9], depicted in Figure 3.8. For example, Sample26_5 refers to the central-hole region of specimen 6 from laminate 2. Due to labeling and mechanical constraints, the top section "9" was not scanned. This area would be used for clamping and is therefore not part of the critical structure where failure could occur. Scans from location 1 are later disregarded in post-processing. Notably, what is also shown in Figure 3.8, one of the specimens (rightmost) displays a +45°, unlike the -45°orientation observed in others. This deviation is the result of the manual lay-up defect, previously discussed in Section 3.1.1.

Furthermore, adequate acoustic coupling between the transducer and the specimen surface is essential for reliable ultrasonic inspections. It was observed that delaminations near the hole, resulting from drilling, caused poor acoustic coupling on the smooth side of the laminate. Therefore, all regions were scanned from the smooth side, except the hole area, which was scanned from the textured peel-ply side where coupling conditions were found to be better. A water-soluble ultrasonic gel was used consistently to enhance signal transmission.

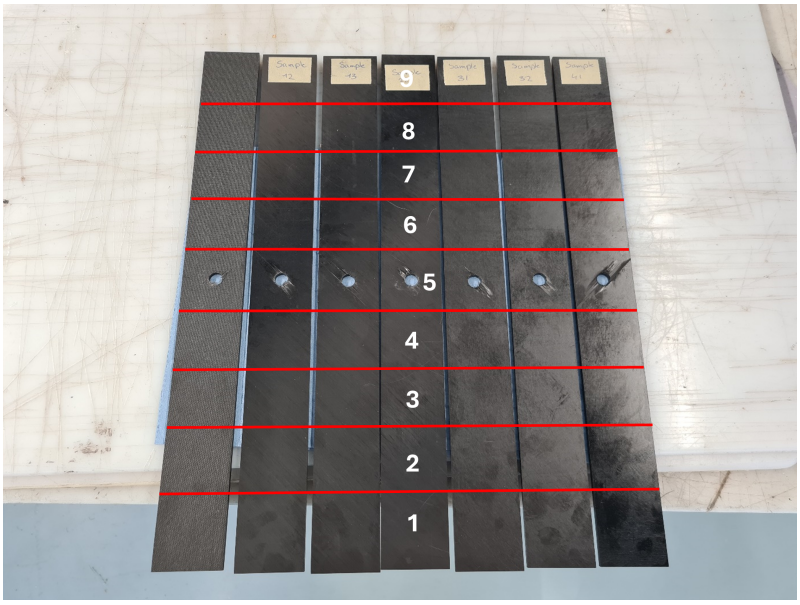
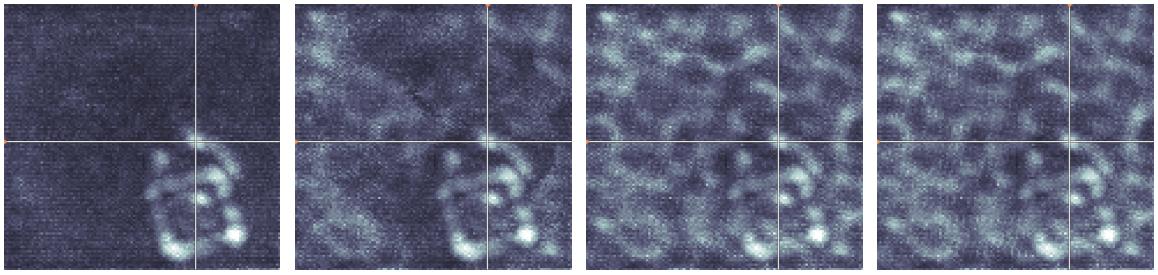


Figure 3.8.: Labeled scan locations across all specimens. Fiber pull-out and slight delamination is visible on drill exit side. Rightmost specimen shows manual lay-up error in +45°.

3.3.2. Gate Selection and Optimization

Gate settings were selected to balance signal clarity and depth resolution. Surface noise increased beyond 3.01 mm, corresponding to reflections from the backside texture. Therefore, for smooth-side scans, the gate was limited to 0.00–3.01 mm, ≈ up to ply 15, to capture internal features while minimizing noise. Figure 3.9 compares signal quality across various gate depths.



(a) Gate 0–3.01 mm (b) Gate 0–3.10 mm (c) Gate 0–3.38 mm (d) Gate 0–3.75 mm

Figure 3.9.: Comparison of gate depth selections for Sample27_3. Noise levels increase with gate depth due to scattered echoes.

3.3.3. Determination of Electronic Parameters

Optimizing the electronic acquisition parameters was essential to ensure high-resolution, low-noise ultrasonic C-scan imaging. The most critical settings are summarized in Table 3.2. These parameters were fine-tuned through a combination of theoretical guidelines and empirical tuning to match the specific material and thickness used in this study. In particular:

- **Acoustic Velocity (c):** was set to 3070 m/s to accurately convert time-of-flight into depth.
- **Pulse Length (**375 ns**)** and **Transducer Pitch (**0.25 mm**)** balanced axial and lateral resolution.
- **Trigger Delay (**7.4 μ**s)** and **Pulse Re-Trigger Time (**10 μ**s)** ensured clean signal separation.
- **Signal Averaging (**8×**)** reduced random noise.
- **Transmit Aperture)** of 4 elements yielded sufficient depth penetration without degrading lateral resolution.

These optimized parameters collectively enhanced the clarity, contrast, and interpretability of the ultrasonic amplitude C-scans. The selection of pitch, frequency, and gate timing was especially critical to accurately resolve embedded manufacturing defects and minimize noise from surface roughness or coupling inconsistencies. For detailed parameter derivations, refer to Appendix B.

3.3.4. NDI Results

The selected acquisition parameters provided optimal imaging conditions for detecting embedded defects such as artificial voids and silicon oil contamination. These settings were established through a combination of manufacturer and theoretical guidelines, and iterative empirical tuning tailored to the specific geometry, lay-up, and material properties of the fabricated CFRP laminates.

Table 3.2.: Summary of Dolphicam2 scanning parameters used for amplitude C-scans

Parameter	Symbol / Description	Value
Acoustic velocity	c	3070 m/s
Transducer pitch	p	0.25 mm
Pulse length	$\tau = \frac{N_{\text{cycles}}}{f_0}$	375 ns (3 cycles at 8 MHz)
Pulse re-trigger time	T_{rep}	10 µs
Trigger delay	t_{delay}	7.4 µs
Transmit aperture	Tx elements	4
Signal averaging	N_{avg}	8
Gate depth (smooth side)	$t \in [t_1, t_2]$	0.00 – 3.01 mm
Gate depth (textured side)	$t \in [t_1, t_2]$	0.40 – 3.75 mm
Scanning aperture	Active area	30×30 mm^2
Spatial resolution (approx.)	$\sim 2p$	~ 0.5 mm

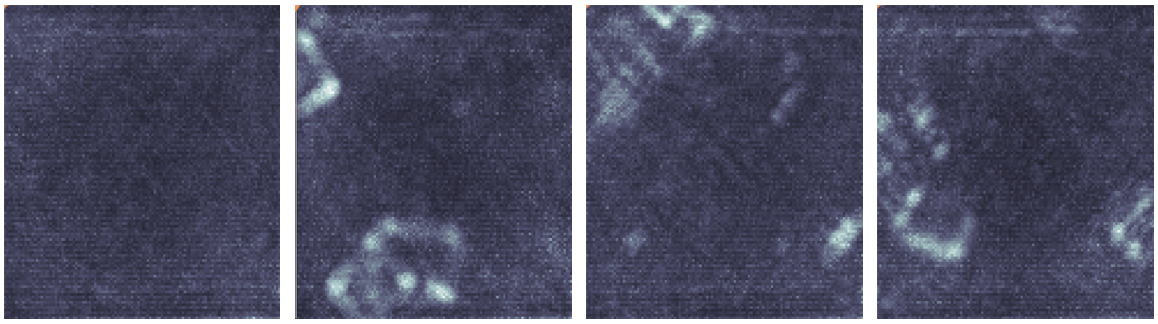
While effective for this configuration, parameters like gate timing, pulse length, and signal averaging remain dependent on laminate thickness, fiber orientation, and matrix properties, and would require adjustment for other composite systems.

Out of the 32 fabricated specimens, 25 were scanned prior to fatigue testing. The remaining specimens were used for mechanical setup, system calibration, and parameter verification. This NDI campaign yielded 200 scans across eight locations per specimen, following the naming and scan protocol described in Section 3.3.1.

Each scan included an amplitude C-scan, Time-of-Flight C-scan, and a representative A-scan waveform. To facilitate post-processing, a custom Python script was developed to extract and normalize the amplitude C-scan, ensuring consistent grayscale scaling across the dataset. The scans were organized by sample and location for efficient use in downstream analysis.

The amplitude C-scans showed clear contrast between pristine and defective regions, with consistent lateral resolution across all samples. Representative examples of each defect category are shown in Figure 3.10, where features such as attenuation patterns and geometric outlines allowed reliable distinction between voids, patch inserts, and oil contamination.

With each specimen's internal quality now quantified, the study transitions to fatigue testing and strain field acquisition using DIC, described in the next section.



(a) Sample13_7_AMP: No defect (b) Sample24_7_AMP: Patch inserts (c) Sample35_4_AMP: Oil contamination (d) Sample43_6_AMP: Mixed defects

Figure 3.10.: Representative normalized amplitude C-scans illustrating various embedded defect types.

3.4. Fatigue Test-Setup and Results

This section describes the experimental setup used for cyclic loading and strain acquisition. It includes the mechanical configuration, DIC system, and loading protocol. The fatigue behavior of each specimen is evaluated, and failure is defined based on a normalized strain threshold, providing the degradation trajectories used for prognostic modeling.

3.4.1. Overview

An overview of the experimental setup is shown in Figure 3.11, with key equipment highlighted using colored boxes for clarity. The system integrates DIC measurements, and mechanical load application via a servo-hydraulic fatigue testing machine. Note that, the AE system is depicted in the setup, but acoustic data is not used in this thesis. The components are briefly described as background information since they will appear in images of the specimens and setup.

Figure 3.11.: Overview of experimental setup; Yellow: AE system. Green: DIC system. Red: MTS fatigue machine and control interface.

Acoustic Emission System The AE system, depicted in yellow, consisted of an 8-channel Vallen AMSY-6 system operated via Vallen VisualAE software. Two Vallen VS900-M passive sensors (100–900 kHz frequency range) were attached to the specimens using clamps and coupling gel for optimal signal transmission. Signals were amplified with 34 dB pre-amplifiers before reaching the data acquisition hardware. The AE system was synchronized with the MTS controller, embedding force and displacement data into the AE stream. While not used in this study, the data was recorded for potential future analysis.

Digital Image Correlation In green, the components of the DIC system are highlighted. Surface strain measurements were performed using a stereo DIC system composed of two FLIR 5 MP industrial cameras, each equipped with a 50 mm fixed focal length lens. The cameras were positioned approximately 600 mm from the sample to ensure a full view of the central defect region. The field of view was approximately 120 mm, which captured the critical zone around the 10 mm diameter open hole, which is expected to be the initiation site for damage and crack propagation.

The VIC-3D software environment was used to perform real-time DIC tracking and data storage. To maximize speckle contrast, a diffuse LED light source was directed at the specimen from behind the cameras at an angle that would not cause direct reflections into the lenses, also to minimize fluctuations in ambient lighting, particularly during extended night-time testing. To further suppress background reflections and lighting variation, a matte black backdrop was placed behind the specimen.

Fatigue Testing Machine Finally, in red, the servo-hydraulic MTS fatigue system with a load capacity of 100 kN and MTS control console are depicted. The clamping force was hydraulically controlled and could be adjusted up to 200 bar. Test parameters such as frequency, load ratio, and cycle count were configured via the MTS control console. This console also initiated and monitored the test progression throughout the fatigue loading campaign.

This integrated setup ensured synchronized acquisition of visual (DIC), and mechanical data, forming the basis for the multi-modal assessment of fatigue damage evolution in the composite specimens.

3.4.2. Fatigue Performance

To determine baseline mechanical properties, representative specimens from each laminate were subjected to quasi-static tension tests to determine their Ultimate Tensile Strength (UTS) and failure strain. The tests were displacement-controlled at a rate of 2 mm/min. Results are presented in Table 3.3. The pristine laminate (Laminate 1) had the highest UTS of 28.96 kN, while the most defective laminate (Laminate 4) reached 26.48 kN. The average failure strain across all laminates was approximately 11.4%, consistent with high deformability associated with the $\pm 45^{\circ}$ cross-ply layup.

Table 3.3.: Quasi-static tension test results for representative specimens

Sample ID	UTS (kN)	Failure strain (%)	Control Mode
Sample11	28.96	11.97	2mm/min
Sample22	28.29	11.60	2mm/min
Sample31	27.86	11.68	2mm/min
Sample41	26.48	10.51	2mm/min
Average	27.90	11.44	-

Based on these results, fatigue parameters were selected to ensure measurable degradation in the weakest laminate without excessively extending test durations for the pristine ones. A peak load of 16 kN ($\approx$ 57% of the average UTS), was applied with a load ratio of $R = 0.1$ at 7 Hz using a 100 kN servo-hydraulic MTS machine. Although a constant-amplitude loading was intended, the large strain deformation and conservative PID control resulted in slight amplitude variability. Consequently, the fatigue loading followed a variable-amplitude waveform, depicted in Figure 3.12. Exiting a DIC cycle, the 7 Hz, 1.6-16 kN force command, depicted in blue is visible.

In red, the measured force, has a converging trend towards the defined amplitude. It reaches 15 kN after $\approx$ 10 seconds and needs an additional $\approx$ 20 seconds to reach a stable 1.6-16 kN amplitude loading.

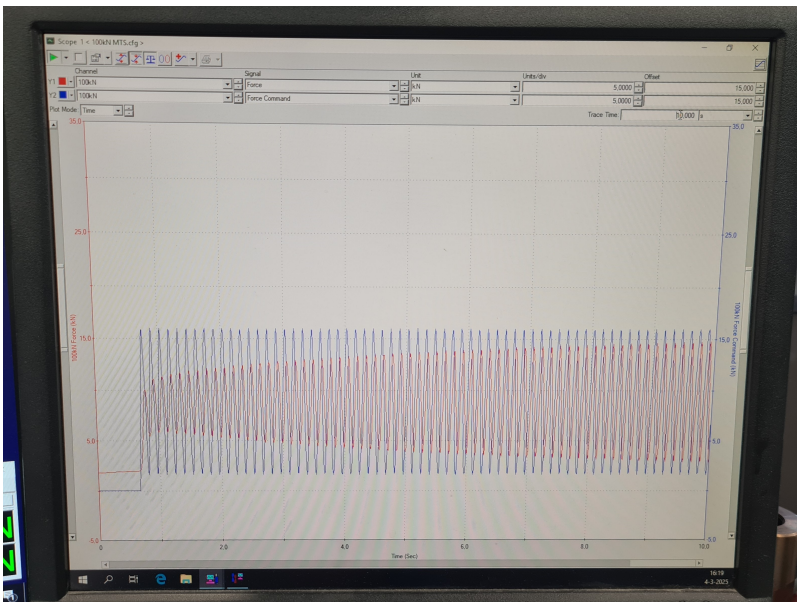


Figure 3.12.: Converging trend visible in fatigue amplitude loading. In blue, the force command and in red, the measured force is depicted.

Notably, upon initial loading to 16 kN, all specimens exhibited audible and visible cracking due to the vulnerability of the cross-plies. The central hole deformed into an elliptical shape, and slight necking was observed around the hole edge. Despite this early-stage damage, the loading protocol was maintained and testing continued.

DIC images were acquired every 625 cycles. At each interval the MTS ramped to the peak load of 16 kN, displacement controlled at 15 mm min^{-1}, held for 0.5 s for system stabilization, acquired a DIC image, wait another 0.5 , ramp down at 15 mm min^{-1} and return to cyclic loading. Each block thus included 624 fatigue cycles plus 1 quasi-static DIC cycle.

A summary of the fatigue test conditions and observations is presented in Table 3.4. Twelve fatigue tests, three from each laminate, were selected for further analysis. Some samples were not tested until failure due to time constraints. These samples are marked with an asterisk (*). Closer examination of the fatigue results reveals a consistent trend linked to the manufacturing-induced defects. Samples from laminate 3 and 4, which contained embedded oil, demonstrated significantly reduced fatigue life compared to the other groups. Except for Sample26, all specimens from laminate 1 and 2, constructed without embedded oil, exceeded 800,000 cycles or did not fail within the test duration. In contrast, most samples from laminate 3 and 4 failed after fewer than 100,000 cycles. This contrast strongly suggests that the presence of oil substantially degraded interlaminar bonding and accelerated damage initiation and propagation.

Table 3.4.: Summary of fatigue tests and test conditions

Sample ID	Force (kN)	Ratio (-)	Frequency (Hz)	DIC Interval	Cycles
Sample14*	16.0	0.1	7.0	1/624	364k
Sample15*	16.0	0.1	7.0	1/624	488k
Sample16*	16.0	0.1	7.0	1/624	495k
Sample26	16.0	0.1	7.0	1/624	823k
Sample27*	16.0	0.1	7.0	1/624	1.38M
Sample28*	16.0	0.1	7.0	1/624	1.40M
Sample35	16.0	0.1	7.0	1/624	84k
Sample37	16.0	0.1	7.0	1/624	24k
Sample38	16.0	0.1	7.0	1/624	22k
Sample46	16.0	0.1	7.0	1/624	78k
Sample47	16.0	0.1	7.0	1/624	195k
Sample48	16.0	0.1	7.0	1/624	69k

3.4.3. Prognostic Health Indicator and Strain Feature Extraction

The objective of post-processing DIC data was to extract a suitable prognostic *Health Indicator (HI)*, a feature that correlates with damage accumulation and can be used to estimate the RUL of the structure. As defined by Eleftheroglou et al. [58], a prognostic indicator is defined as a measurable quantity derived from Structural Health Monitoring (SHM) data that exhibits a consistent, ideally monotonic trend, is robust to noise and operational variability and reflects the underlying damage mechanisms.

Eleftheroglou et al. emphasize three key metrics: *monotonicity*, *prognosability*, and *trendability*. In this thesis, monotonicity was prioritized, based on the assumption that fatigue-induced degradation should follow a clear unidirectional progression. A strain region that exhibits consistent monotonic progression under cyclic loading was thus considered suitable for prognostic modeling.

Damage Behavior and Speckle Loss

Visible damage occurred early in fatigue testing. Around the hole, local necking, paint cracking and delaminations were observed from the first load cycle, as illustrated in Figure 3.13. These effects led to partial speckle loss and degraded DIC tracking. Final failure typically occurred through $\pm 45^{\circ}$ matrix cracks and delaminations between the interfaces, as seen in Figure 3.14, along with a release film patch sticking out.

Figure 3.15 shows how local deformation along the $+45^{\circ}$ failure path caused progressive DIC correlation loss, especially near the side notches of the hole. After several trials, only the region directly beneath the hole exhibited consistent, monotonic strain development across all specimens. This region remained free of early delamination and preserved speckle quality throughout testing.

An uniform Region Of Interest (ROI) was therefore selected beneath the hole for all specimens (Figure 3.16). To reduce noise from surface settling and transient early-cycle effects, the first 10 fatigue blocks (6250 cycles) were excluded. Subsequent axial strain values were normalized relative to this post-transient baseline, allowing comparison across different specimens. Note, this normalization approach remains compatible with real-time application. The baseline is established after the initial stabilization phase during operation, following standard practice in SHM.

As a result, all subsequent strain measurements can be tracked relative to this in-service reference point, allowing the feature to be used effectively in online prognostic systems.

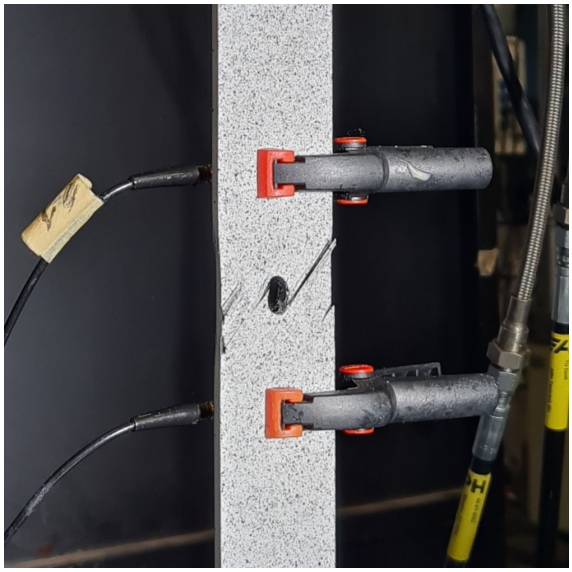


Figure 3.13.: Elliptical hole deformation, necking and paint cracking during early fatigue cycles.

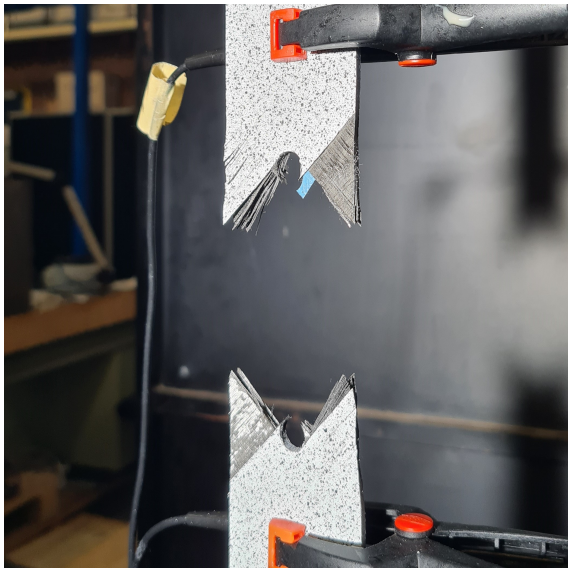


Figure 3.14.: Failure along $\pm 45°$ due to matrix cracking and delaminations with a patch defect sticking out.

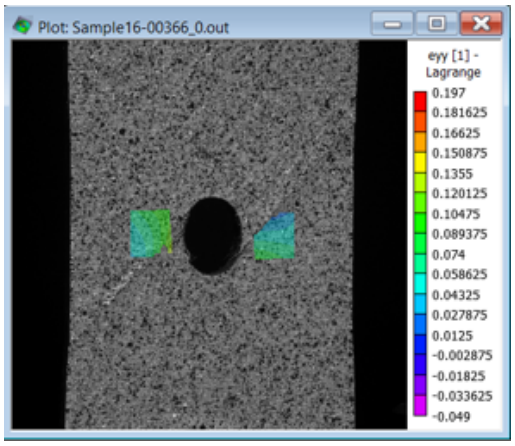


Figure 3.15.: Loss of DIC tracking along the $+45°$ failure path.

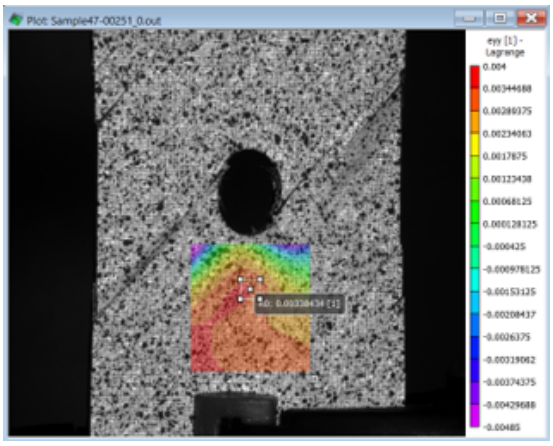


Figure 3.16.: Stable region below the hole, where ROI is used for strain extraction.

Strain Evolution and Prognostic Feature Definition

The normalized axial strain curves are presented in Figure 3.17. Under tension-tension fatigue loading, the samples exhibit cyclic variation between axial and transverse strain components. While all specimens were subjected to identical loading conditions (Table 3.4), not all were tested until failure. These curves primarily serve to demonstrate the consistency and suitability of the extracted strain feature for prognostic modeling.

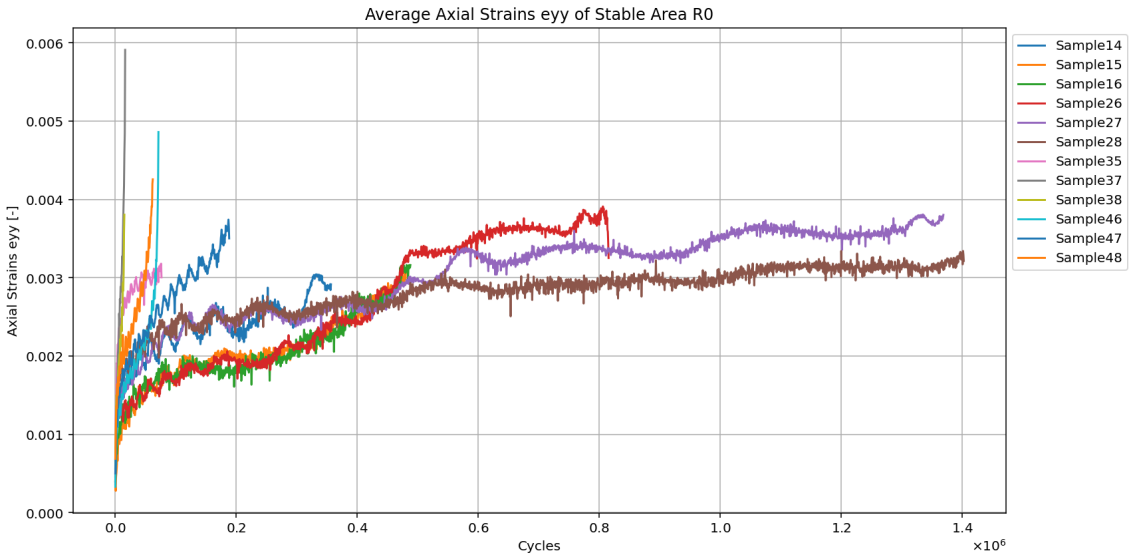


Figure 3.17.: Normalized axial strain curves for all tested samples under identical fatigue loading.

Although irreversible damage, such as matrix cracking and delamination, occurred early in life, this was not defined as EoL in this study. In aerospace applications the design principle in composite structures is the "no damage or zero tolerance" approach. Such a definition is overly conservative in the context of long-term fatigue prediction. Several specimens exceeded one million cycles without reaching structural failure, highlighting that visible damage does not always correspond to EoL from a mechanical standpoint. Instead, EoL was defined by a critical reduction in stiffness, which corresponds to increased strain under constant loading. Given the uniform specimen geometry, a predefined axial strain value can be used as a surrogate for stiffness loss.

A strain threshold of 0.003 (normalized axial strain) was adopted. This value corresponds to a consistent point of stiffness reduction across specimens, while still falling within the monotonic region of the strain curve. Once a specimen exceeded this threshold, it was considered to have reached failure for prognostic purposes.

Figure 3.18 shows the strain curves clipped at this failure threshold. These curves, with a clear monotonic trend and a uniform EoL condition, were then discretized and used as input for the prognostic modeling process. The results of this modeling phase are presented in Chapter 5.

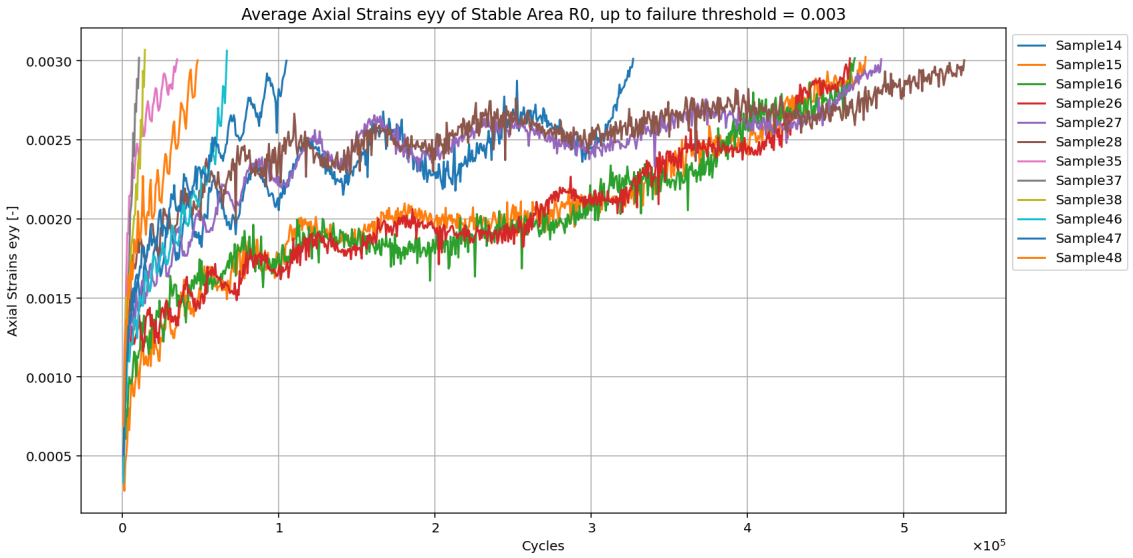


Figure 3.18.: Normalized axial strain curves for all tested samples up to axial strain threshold of 0.003.

4. Similarity Methodology

As outlined in Section 2.5, this thesis proposes a methodology for managing past state uncertainties in RUL predictions through similarity-informed modeling. This chapter builds upon the case study presented in the previous chapter by using the obtained Dolphicam2 C-scan images to develop and evaluate a set of tailored similarity methods. The aim is to compute meaningful similarity scores across specimens based on their internal structural characteristics.

A core assumption underlying the proposed methodology is that composite specimens exhibiting similar internal material quality (e.g. defect types, void density, or fiber misalignment), are likely to degrade in a similar manner and reach failure under similar operational conditions. This relationship forms the basis for incorporating structural similarity as prior knowledge into the prognostic process. Instead of training the prognostic model on a broad dataset comprising all available specimens, which may include structurally dissimilar cases, the approach conditions the model on a subset of training specimens that are most similar to the test samples in terms of their internal quality. By assigning more weight to these structurally similar specimens, the model is expected to generalize more effectively to the individual case at hand.

The rationale for this approach is grounded in the belief that degradation behavior is inherently influenced by the initial structural state of the material. Thus, leveraging structurally similar cases enables the model to learn from degradation paths that more closely reflect the test sample's actual behavior, leading to a more reliable estimate of its RUL. This similarity-informed strategy also reduces the impact of variability introduced by unrelated samples, thereby improving prediction accuracy and narrowing the associated uncertainty bounds. In essence, this methodology shifts the focus from population-based predictions to specimen-specific estimates, aligning the training context with the physical condition of the test specimen.

This chapter presents the development and evaluation of three image-based similarity methods: the Structural Similarity Index (SSIM), a Spatial Pyramid Histogram (SPH)-based similarity method, and a CNN-based embedding method trained using triplet loss. These methods were initially identified in the literature review (Section 2.4) due to their proven effectiveness in quantifying structural and textural differences in image-based applications. Given the lack of publicly available Dolphicam datasets, and considering the unique characteristics of Dolphicam ultrasonic imaging data, it was first necessary to experimentally generate scan data, as detailed in the preceding case study. With the Dolphicam scans now available from Chapter 3, a detailed development of these similarity methods is realized. The methods discussed in this chapter have thus been specifically tailored and optimized based on actual ultrasonic imaging data, ensuring that the computed similarity scores effectively capture the internal structural characteristics relevant to the degradation behavior of aerospace-grade composites.

4.1. Structural Similarity Index

SSIM is a well-established method in the field of image quality assessment [45, 47]. Unlike simple intensity-based metrics such as MSE, SSIM compares local patterns of pixel intensities that have

been normalized for luminance and contrast [54]. The SSIM value between two image patches x and y is computed as [47]:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.1}$$

where μ denotes the local mean, σ^2 the variance, and σ_{xy} the covariance between image patches. The constants C_1 and C_2 are used to stabilize the division and prevent numerical instability when denominators are near zero.

SSIM is typically used to assess image quality degradation by comparing a distorted image against a high-quality reference. It provides values between 0 and 1, where 1 denotes perfect similarity. Its strength lies in capturing structural information such as edges, textures, and local correlations [47]. Based on the capabilities, SSIM was initially considered a promising candidate for comparing Dolphicam C-scans. However, during implementation, several key limitations emerged that rendered SSIM unsuitable for global similarity assessments in this context:

- Internal defects in the scanned laminates are randomly distributed, with no consistent reference position across specimens.
- Feature alignment is infeasible due to variability in defect location, shape and orientation.
- SSIM is highly sensitive to small spatial misalignments and scale variations, both of which are inherent in the dataset.

Although SSIM showed distinctive similarities scores in assessing intra-sample consistency (e.g., comparing different cropped regions of the same specimen), its effectiveness collapsed when comparing global scan structure across specimens. Preliminary experiments, including the generation of a SSIM similarity matrix heatmap (Appendix C.1), revealed that SSIM was sensitive to local variations and alignment inconsistencies across the scans. These effects undermined its reliability in capturing meaningful, scan-wide structural similarity across specimens with diverse internal features. The similarity scores were low and failed to show meaningful grouping or distinguishability between laminates with similar internal features.

While SSIM could offer an approach for localized application, such as segmenting and comparing individual defect regions, this would require consistent image segmentation and defect isolation algorithms, which were considered beyond the scope of this thesis. Given these limitations, the focus shifted toward histogram-based similarity and deep learning methods, which are better suited to capture global intensity distributions [46] and learn structural patterns directly from data [49]. These alternative approaches are more resilient to spatial misalignment.

4.2. Histogram Based Similarity Using Spatial Pyramid Histograms

Histogram-based methods are widely used in image analysis due to their simplicity, computational efficiency and interpretability. However, traditional global histograms compress all pixel intensity values into a single frequency distribution, discarding spatial information that may be critical when analyzing localized defects in composite structures. This spatial insensitivity makes them inadequate for capturing fine-grained material quality differences in ultrasonic C-scans.

To address this limitation, this work adopts the SPH technique [59], which retains the core advantages of histogram-based methods while introducing hierarchical spatial encoding. SPH allows for the representation of both global and local pixel intensity distributions, making it a suitable middle ground between simplistic global histograms and more complex, data-hungry deep learning methods. SPH does not require model training and is therefore more transparent and less data-intensive.

4.2.1. Spatial Pyramid Histogram Methodology

The SPH technique improves upon global histograms by partitioning an image into grids at multiple levels of resolution and computes local histograms within each grid cell. These localized histograms are then concatenated into a single feature vector that encodes both global and local intensity distribution information [60]. This layered decomposition maintains a level of spatial awareness that improves over traditional global histograms. At pyramid level L, the image is divided into $2^L \times 2^L$ grids. Let $H_{l,i}$ be the histogram of the i^{th} cell at level l. The final feature vector $\mathbf{H}_{\text{SPH}}$ is given by:

$$\mathbf{H}_{\text{SPH}} = \bigcup_{l=0}^{L} \bigcup_{i=1}^{4^l} H_{l,i} \tag{4.2}$$

Figure 4.1 and Figure 4.2 visualize pyramid level 2 ($2^2 \times 2^2$) and level 3 ($2^3 \times 2^3$) grid overlays on representative amplitude scans. As the pyramid level increases, the spatial resolution becomes finer, enabling the method to capture more localized variations in image texture and brightness.

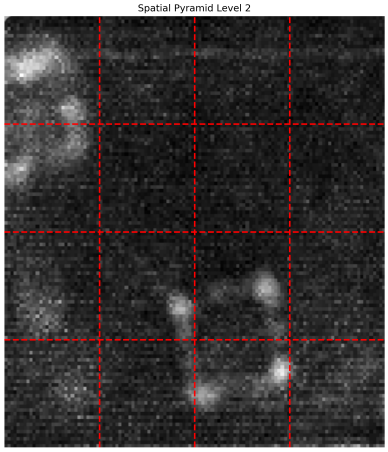


Figure 4.1.: Spatial Pyramid Histogram grid overlay at level 2.

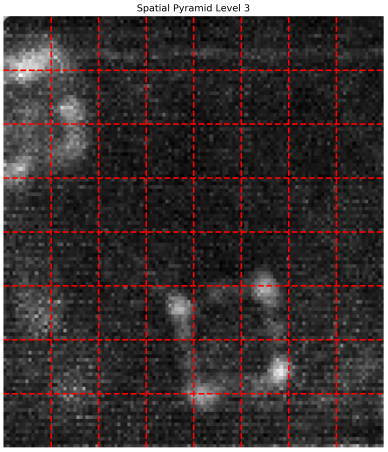


Figure 4.2.: Spatial Pyramid Histogram grid overlay at level 3.

4.2.2. Preprocessing for Histogram Computation

To enhance the sensitivity of the histogram-based comparison to defect patterns while reducing sensitivity to texture noise, several preprocessing steps were applied:

1. **Grayscale conversion and normalization**: All amplitude scans were first converted to grayscale and normalized to the [0,1] intensity range. This step ensured consistent contrast and intensity scaling across the dataset, eliminating variability introduced by scan-to-scan brightness differences. Normalization was essential to make histogram comparisons meaningful and reliable across specimens.
2. **Gaussian blurring**: A Gaussian blur was applied to suppress fine-grained texture and reduce noise, while preserving higher-intensity regions typically associated with defects (e.g., voids, delaminations, or contamination). This step reduces the influence of speckle and measurement noise on the histogram. The kernel size and standard deviation were selected iteratively through visual inspection, balancing two goals: (1) preserving the bright, spatially diffuse features typically associated with embedded defects, and (2) attenuating small-scale texture that did not correlate with material quality variation.
3. **Threshold masking**: To further focus histogram computation on potentially defect-related regions, a global intensity threshold was applied. Pixels below this threshold were excluded from the histogram, under the assumption that low-intensity values primarily reflect background or structurally sound regions. The threshold value was set empirically based on analysis of pristine versus defective specimens. As shown in Figure 4.3, pristine scans (e.g., Sample17_6_AMP) show nearly all information concentrated in the lower intensity range [0-0.4], while defective samples (e.g., Sample24_7_AMP) exhibit a broader intensity spread, particularly above 0.4. This guided the threshold selection to emphasize higher-intensity, defect-prone regions.

Focusing the analysis on the higher-intensity (defect-relevant) region was essential for enabling meaningful similarity computation. Without threshold masking, histogram comparisons were dominated by low-intensity background regions shared across all specimens, resulting in artificially high similarity scores and poor discriminative power. By isolating the brighter, structurally informative parts of the scan, this preprocessing step allowed the similarity method to distinguish pristine from defective laminates. When combined with the spatial encoding of the SPH method, this focused representation enabled differentiation not only by presence of defects, but also by defect type and spatial distribution.

Figure 4.4 illustrates this pipeline on a representative scans, highlighting the effect of normalization, blurring, and threshold masking. This tailored preprocessing ensures that the resulting SPHs reflect meaningful structural differences between specimens, rather than superficial intensity variations or acquisition noise. By carefully tuning each step to align with the physical characteristics of the scanned material, the method improves the interpretability and relevance of the computed similarity scores.

4.2.3. Similarity Metric

The extracted SPH feature vectors were compared using multiple similarity, including histogram intersection [61], cosine similarity, Bhattacharyya distance, and chi-squared distance [52]. Among these, histogram intersection consistently produced the most interpretable and discriminative results across the dataset. Histogram intersection $S(H_1, H_2)$ quantifies the degree of overlap between two normalized histograms H_1 and H_2, and is defined as:

$$S(H_1, H_2) = \sum_{i=1}^{n} \min(H_1(i), H_2(i)) \tag{4.3}$$

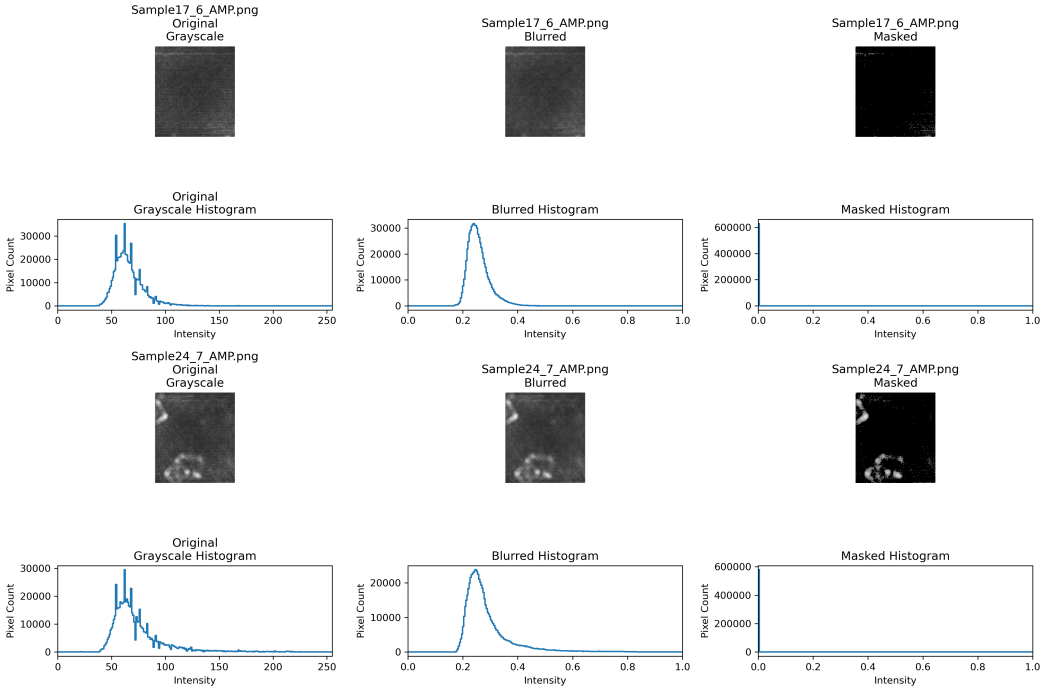


Figure 4.3.: Effect of preprocessing on grayscale histograms: original, blurred, and masked histograms shown for two samples.

This method directly quantifies the overlap in normalized frequency between two SPH feature vectors. It remained sensitive to structural differences in the data, whereas other metrics failed to provide discriminative results. As nearly the entire gray-level mass of every SPH histogram lies in a set of identical low-intensity bins, other overlap-based distances such as cosine similarity, Bhattacharyya distance, and chi-squared distance saturated to high similarity scores across all specimens (see Appendix C.2). In contrast, histogram intersection remained effective because, in this implementation, a brightness threshold was applied to discard low-information bins prior to comparison. This effectively filtered out the common background intensities and emphasized structural variations relevant to defect characterization.

Finally, Equation 4.4 defines the histogram intersection similarity score as the sum of the minimum values of corresponding bins in the SPH feature vectors of two images. Each SPH vector, $\mathbf{H}_{\text{SPH}}$, is constructed by concatenating local histograms $H_{l,i}$ computed over spatial grid cells at multiple pyramid levels, as shown in Equation 4.2.

$$S\left(\mathbf{H}_{\text{SPH}}^{(1)}, \mathbf{H}_{\text{SPH}}^{(2)}\right) = \sum_{j=1}^{n} \min\left(H_{\text{SPH}}^{(1)}(j),\ H_{\text{SPH}}^{(2)}(j)\right) \tag{4.4}$$

Here, n denotes the total number of bins across all pyramid levels and grid cells, and $H_{\text{SPH}}^{(k)}(j)$ is the j-th bin of the SPH feature vector for specimen k. This overlap formulation emphasizes regions where both specimens exhibit similar intensity patterns, making it well-suited for identifying structurally similar defect characteristics across C-scan images.

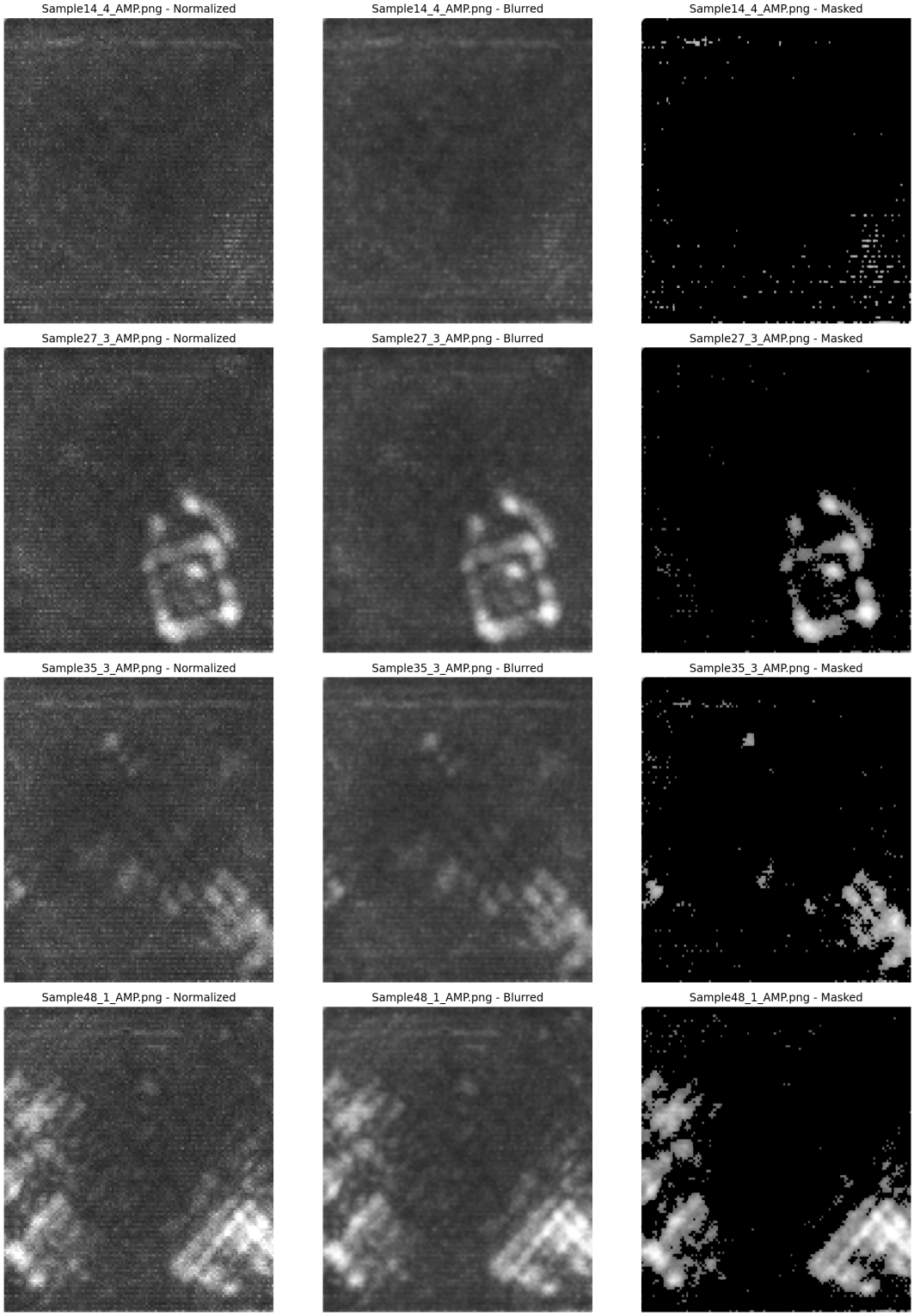


Figure 4.4.: Preprocessing pipeline applied to amplitude scans: normalized, blurred, and masked versions.

4.2.4. Limitations and Exclusions

Scans containing the open-hole region (i.e., scan index _5) were found to contain either overly dominant bright zones or noise-induced high-amplitude regions, which significantly skewed histogram distributions. Including these scans resulted in disproportionate influence on global similarity scores. To mitigate this, hole scans were excluded from SPH computation. Instead, scan index _1, representing a structurally relevant but non-critical area in fatigue testing, as this is the clamping area, was included to better capture global material characteristics.

4.2.5. Final Parameter Selection and Similarity Matrix Evaluation

The final configuration for the SPH-based similarity analysis was determined through iterative tuning, aiming to balance discriminative power, interpretability, and computational efficiency. Each parameter was selected to enhance the method's ability to capture meaningful structural variation while suppressing noise and minimizing bias from dominant background regions:

- **Intensity threshold (0.4):** As identified in Figure 4.3, pixel intensities below 0.4 represent the non-informative background in both pristine and defective samples. By masking out this low-intensity range, the similarity computation was focused on the [0.4–1.0] domain, which contains the features most indicative of internal defects.
- **Gaussian blur kernel size (25×25):** This kernel size was chosen to suppress pixel-level texture noise while preserving broader brightness patterns associated with defects. It provided a balance between detail preservation and smoothness, considering the original resolution of the Dolphicam amplitude scans (752 × 853 pixels).
- **Gaussian sigma (3.0):** A standard deviation of 3 pixels produced moderate smoothing, effective in reducing fine surface irregularities without erasing the intensity gradients of small but significant defect regions.
- **Pyramid level (3):** A level 3 subdivision (8×8 grid) was selected as it provided a balanced trade-off between capturing sufficient spatial detail and avoiding overly sparse descriptors. Lower pyramid levels (e.g., level 2) lacked the resolution to detect localized defect regions, while higher levels (e.g., level 4) introduced excessive sensitivity to noise and substantially increased computational complexity.

These parameter selections reflect a trade-off between maximizing sensitivity to internal structural variation and minimizing computational complexity. The derived SPH vectors are compact and informative, enabling reproducible representation of material quality across specimens.

Using this finalized setup, similarity scores were computed using histogram intersection for all pairwise combinations of SPH feature vectors. The resulting raw similarity matrix is shown in Figure 4.5, and the Min–Max row-normalized matrix in Figure 4.6. Normalization was applied per row to highlight relative similarity patterns across specimens, excluding self-similarity to avoid bias in interpretation. Higher values indicate stronger similarity.

Given a vector $v = \{v_1, v_2, \ldots, v_n\}$, the Min–Max normalized value v_i^{norm} is computed as:

$$v_i^{\text{norm}} = \frac{v_i - \min(v)}{\max(v) - \min(v)} \tag{4.5}$$

In Figure 4.5, the row-wise Min–Max normalized similarity matrix derived from the SPH method reveals notable clustering behavior. Strong intra-laminate similarity is evident for both laminate 1 (Samples14–16) and laminate 4 (Samples46–48), as indicated by consistently high similarity scores within these groups. This demonstrates the SPH method's ability to reliably group structurally similar samples based on their internal C-scan characteristics.

However, the matrix also reveals high inter-laminate similarity between certain groups. For instance, laminate 1 samples exhibit elevated similarity scores with laminate 3 (Samples37–38), while laminate 2 (Samples26–28) appears highly similar to laminate 4. These cross-group similarities can be reasonably attributed to the nature of the embedded defects. Both laminate 2 and laminate 4 contained patch-based defects, which likely produced similar amplitude histogram features. In contrast, laminate 3, which included oil contamination, demonstrated structural characteristics that resembled the pristine features of laminate 1 in the scanned intensity domain.

It is important to note that due to the limitations of the scan depth, specifically, the inability to detect top-layer defects beneath textured surfaces, as discussed in Section 3.3.2, some defect features may not have been fully captured. As a result, certain laminate-specific characteristics were underrepresented in the similarity computation. This scanning limitation likely contributed to the observed inter-laminate clustering.

While these similarities are justifiable based on the known material states, the ideal outcome of the SPH similarity method would be the emergence of strictly intra-laminate clusters. Such grouping would indicate that the method is not only sensitive to structural patterns but also capable of isolating laminate-specific defects. The presence of partially mixed clusters thus reflects both the strengths and limitations of the current scanning and preprocessing configuration.

Finally, to define a criterion for structural similarity, a threshold of 0.80 was applied: samples with similarity scores above this value were considered structurally similar and assigned to the same training subset for prognostic modeling. However, to ensure adequate sample representation for stochastic modeling and to maintain reliable convergence of the prognostic model, each specimen was assigned at least three training specimens. If fewer than three specimens met the similarity threshold, the three most similar ones were selected regardless of absolute score.

The resulting similarity profiles form the basis of the SIM-DICF, which will be used in the prognostic modeling phase described in the next chapter. This dataset reflects a tailored grouping strategy representing similar internal material characteristics, aimed at reducing uncertainty in RUL predictions by aligning test samples with degradation trajectories of structurally similar specimens.

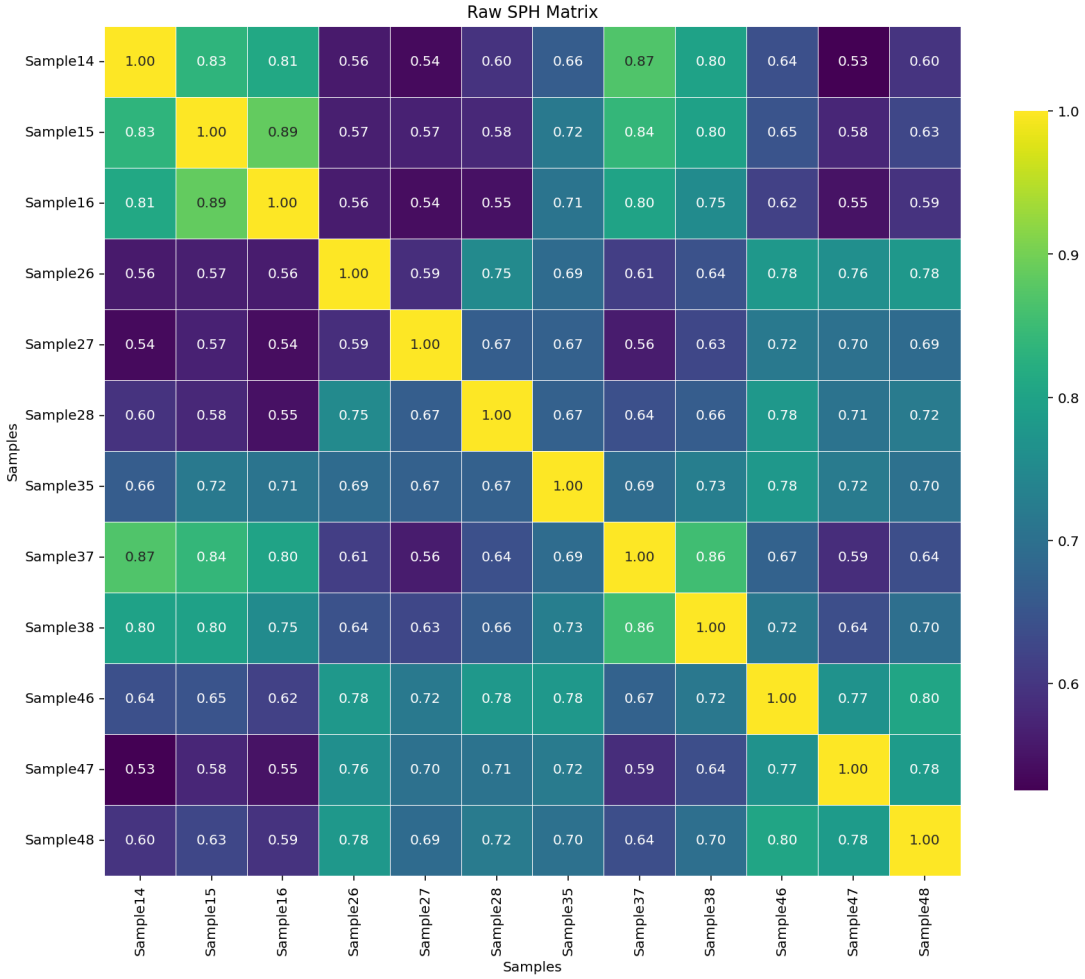


Figure 4.5.: Raw similarity matrix computed using histogram intersection on SPH features.

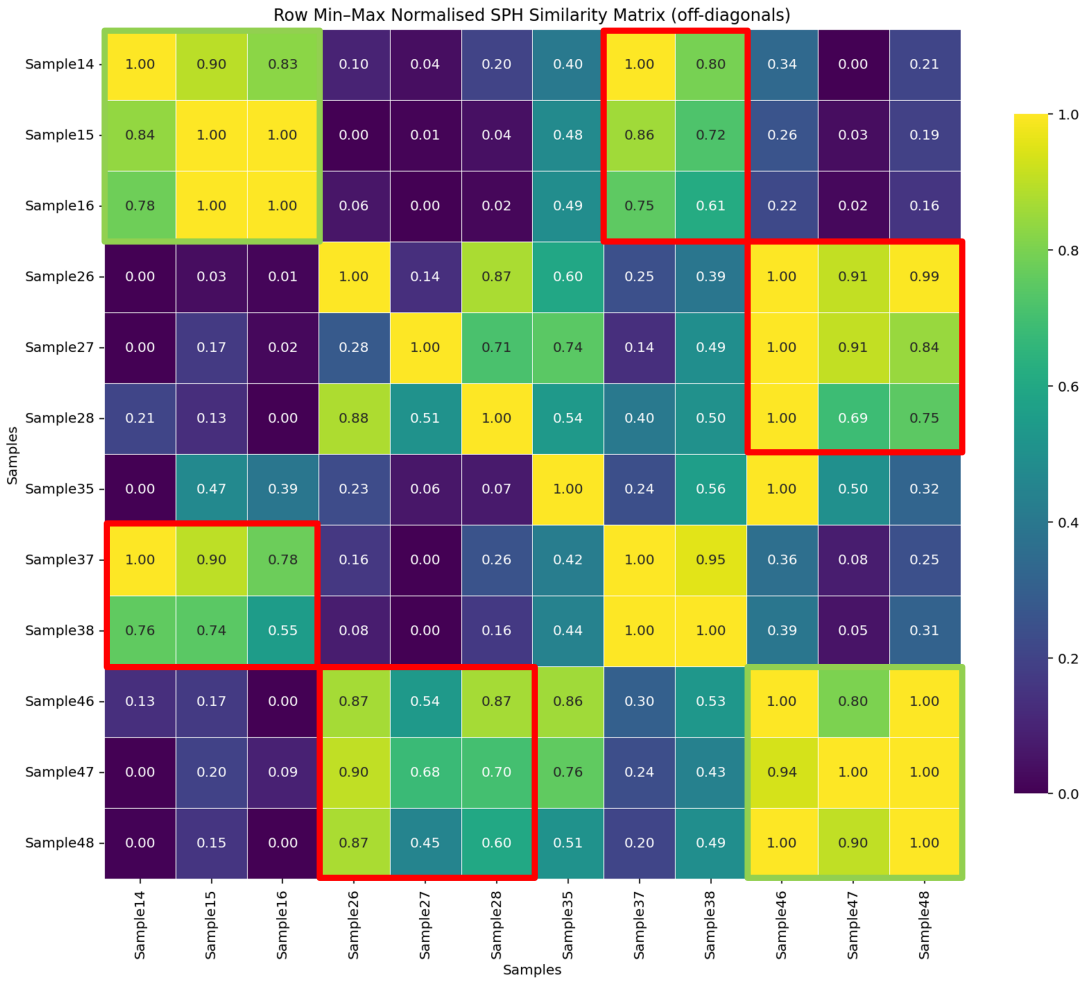


Figure 4.6.: Row-wise Min–Max normalized SPH similarity matrix.

4.3. CNN-Based Similarity Using Triplet-Loss Embedding

While histogram-based techniques provide a transparent and computationally efficient way to compare ultrasonic scans, they are limited in their ability to capture subtle spatial patterns and defect morphologies within high-resolution amplitude images. To address these limitations, this section introduces a CNN-based similarity learning approach. The objective is to learn an image embedding space where samples with similar internal structural quality cluster together, forming the basis for uncertainty-informed prognostic modeling.

Unlike traditional classification approaches, this methodology focuses on similarity learning: instead of assigning fixed labels to each scan, the model learns to evaluate how alike two scans are. A triplet-loss architecture is employed to train the network, as this approach has been successfully used in various computer vision tasks requiring fine-grained grouping or retrieval of similar images, including face recognition, medical imaging, and materials analysis [62, 63, 64, 65].

4.3.1. Theoretical Background

Convolutional Neural Networks

CNNs are a class of deep learning models specialized for image processing. Their architecture comprises multiple hierarchical layers that automatically learn spatially local patterns from pixel data, starting from low-level features like edges and textures and progressing to more abstract representations in deeper layers [49, 65]. A typical CNN consists of:

- **Convolutional layers,** where learnable filters are applied across the image to extract features;
- **Activation functions,** which introduce non-linearity;
- **Pooling layers,** which downsample feature maps to reduce dimensionality;
- **Fully connected layers,** which integrate features for classification or embedding.

CNNs are particularly powerful because they exploit spatial hierarchies, making them well-suited for detecting both global structure and local anomalies in images, crucial for evaluating internal laminate quality in C-scans. As demonstrated by Li et al. [42], CNNs have successfully been used in prognostics to estimate RUL from degradation patterns in sensor data.

Similarity Learning with CNNs

Rather than directly classifying defects, this thesis applies CNNs to a similarity learning task: grouping images based on shared material characteristics. This strategy aligns with work in image retrieval and quality assessment, where the goal is to map inputs to a feature space that reflects perceptual or structural similarity [45, 66]. Within this space, geometrical proximity between vectors corresponds to similarity in internal structure.

Triplet Loss and Siamese Networks

To train a CNN for similarity, a triplet-loss function is employed, embedded within a Siamese Network architecture [67]. Unlike standard classification losses, triplet loss is designed to enforce relative similarity constraints among triplets of samples; an anchor A, a positive P (same class), and a negative N (different class). The goal is to learn an embedding function $f(\cdot)$ that satisfies:

$$\|f(A) - f(P)\|_2^2 + \alpha < \|f(A) - f(N)\|_2^2 \tag{4.6}$$

Here α is the margin, and the network is penalized whenever the negative sample is not sufficiently farther from the anchor than the positive. This encourages intra-class compactness and inter-class separation in the embedding space. The corresponding triplet loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \max\left(\|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha, 0\right) \tag{4.7}$$

Triplet-loss is favored over alternatives such as contrastive loss, motivated by its ability to enforce relative distance constraints, making it especially well-suited for datasets with subtle variations, such as those seen in manufacturing defects in composites. Triplet-loss architectures have become standard in tasks like facial recognition [62], material classification [65], and image similarity-based retrieval, where distinguishing between visually similar yet distinct images is essential [63, 64].

Their effectiveness lies in the ability to shape the embedding space such that similar images are placed closer together while dissimilar ones are pushed apart, see Figure 4.7. This allows models to capture subtle differences in structural features even when no explicit labels are available, making the method suitable for similarity learning in material quality assessment.

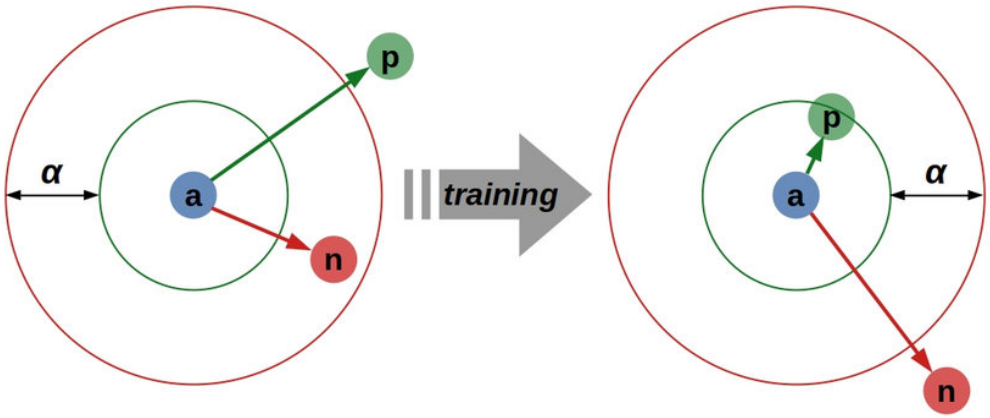


Figure 4.7.: Visualization of the triplet loss constraint: anchor (blue), positive (green, same class), and negative (red, different class) embeddings before and after training [3].

Embedding Space and Similarity Metric

The final layer of the CNN outputs a 128-dimensional vector in $\mathbb{R}^{128}$, capturing the learned features of the input image. These embeddings are then L2- normalized to unit length, projecting them onto the surface of a hypersphere $\mathbb{S}^{127}$, Equation 4.8. This normalization ensures scale-invariance and allows similarity to be computed via the cosine similarity , which simplifies to a dot product between normalized vectors, Equation 4.9:

$$\|f(x)\|_2 = 1 \quad \Rightarrow \quad f(x) \in \mathbb{S}^{127} \tag{4.8}$$

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = x \cdot y, \quad \text{for} \quad \|x\| = \|y\| = 1 \tag{4.9}$$

This distance metric is widely used in embedding-based models due to its efficiency and ranking consistency with Euclidean distance on normalized vectors [68]. This property is particularly useful for image retrieval tasks, where sorting based on similarity is often more important than the exact numerical value of the score.

ResNet18 Backbone

The backbone of a CNN refers to the core feature extractor, typically a pretrained convolutional architecture that transforms raw input images into intermediate feature maps. These maps capture spatial hierarchies and are then passed to additional layers for specialized tasks such as classification, regression or similarity learning. In this study ResNet18 was selected as the backbone due to its favorable balance between network depth and computationally efficiency [69].

A major challenge in training deep networks is the vanishing gradient problem, where gradients used for weight updates become increasingly small as they are propagated backward through many layers. As a result, earlier layers fail to update effectively, hindering the network's ability to learn deeper hierarchical features. This slows training or can even halt it entirely, especially in early layers. ResNet, short for Residual Network, addresses this issue by introducing residual (shortcut) connections. These allow the network to learn residual mappings $F(x) = H(x) - x$, reformulated as $H(x) = F(x) + x$, where $H(x)$ is the desired mapping and x is the input. These shortcut connections improve gradient flow, maintain meaningful weight updates (even in deeper layers) and accelerate convergence during training [69].

Although deeper versions of ResNet (e.g., ResNet50 or ResNet101) can learn more abstract features, ResNet18 offers a favorable balance for small datasets like the one in this thesis. Its relatively shallow depth minimizes the risk of overfitting, while still providing sufficient capacity to encode patterns in C-scan images. This includes detection of subtle structural differences that distinguish pristine and defective laminates.

ResNet18 has demonstrated strong generalization capabilities in fields such as materials analysis [69], industrial surface inspection [70], and medical image classification [71]. These cross-domain successes suggest that the architecture generalizes well to different imaging processes, including ultrasonic C-scans. Moreover, ResNet18 benefits from hierarchical feature learning. The earlier layers capture low-level characteristics such as edges, gradients, and textures, while deeper layers encode more abstract spatial and structural information, including shape outlines and defect morphology. These multi-scale representations are particularly suitable for similarity learning

in composite materials, where variations in defect structure and location, rather than high-level labels, define the internal similarity across specimens. Making it a well-suited backbone for similarity learning in this thesis.

Network Flow and Similarity Computation Summary

In the final architecture, each input image is first passed through the ResNet18 backbone, which extracts multi-level spatial features ranging from edges and textures to higher-level defect patterns. These features are then projected by a two-layer fully connected head into a 128-dimensional embedding vector. The embeddings are L2-normalized to ensure scale invariance and enable meaningful distance comparisons. During training, the network optimizes these embeddings using a triplet-loss function, ensuring that structurally similar scans are positioned closer together than dissimilar ones by at least a fixed margin. At inference, similarity between any two C-scans is computed via cosine similarity (dot product) between their normalized embeddings, allowing for quantitative grouping of specimens based on internal structural quality. This learned embedding space enables more tailored and informed selection of similar training samples for past state uncertainty-aware prognostic modeling.

4.3.2. Data Augmentation and Training Procedure

Given the limited size of the dataset, 200 scans from 25 specimens, extensive data augmentation was essential to increase sample variability and prevent overfitting. Deep learning models, particularly CNNs, are known to overfit small datasets by memorizing training patterns instead of learning generalizable features [49, 72]. The augmentation strategy focused on generating synthetic variations that preserved the physical integrity of defect features while introducing variation in scale, orientation, and intensity. The following augmentations were applied:

- Random affine transformations (rotation, translation, zoom) to simulate different spatial perspectives;
- Horizontal and vertical flips to account for symmetry and eliminate directional bias;
- Brightness and contrast jittering to simulate changes in scanning or lighting conditions;
- Gaussian blur to suppress noisy high amplitude pixels.

All images were zero-padded to square shape and resized to 224×224 pixels, conforming to match ResNet input requirements.

4.3.3. Training and Early Stopping

To track generalization and avoid overfitting, 1 specimen per laminate (4 in total) was withheld for validation. The model was trained for a maximum of 30 epochs using the Adam optimizer [73], with early stopping triggered if validation loss failed to improve for 5 epochs. An epoch refers to one complete pass over the training dataset. As seen in Figure 4.8, training loss gradually decreased from $\approx$ 0.20 to $\approx$ 0.11, indicating increasing ability of the network to distinguish between similar and dissimilar scans. Validation loss, however, reached its minimum ($\approx$ 0.07) at epoch 23 before rising again, an indication that overfitting began, despite continued training loss reduction. Early stopping at this point ensured that the selected model generalizes well without memorizing training data.

This behavior reflects typical instability seen in training deep learning models on small datasets, where stochastic variations and high model capacity can lead to volatility in validation performance [49, 50]. The selected checkpoint at minimum validation loss was therefore used for inference.

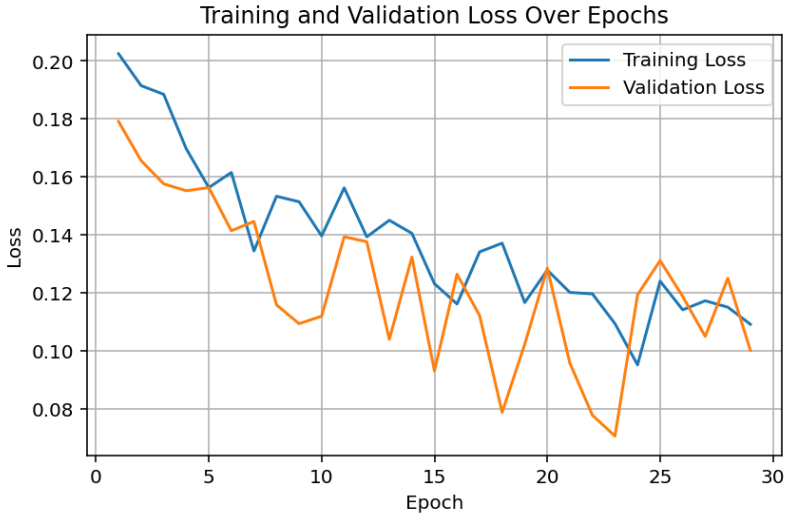


Figure 4.8.: Training and validation loss for CNN triplet-loss embedding model. Minimum validation loss occurs at epoch 23; continued training increases validation loss, indicating overfitting.

4.3.4. Final Parameter Configuration and Similarity Matrix Evaluation

The final model configuration was selected through iterative experimentation and informed by established practices in triplet-loss-based image similarity learning [63, 64]. The goal was to balance representational capacity, training stability, and inference efficiency. The configuration included:

- **Backbone:** ResNet18 (pre-trained on ImageNet [74]),
- **Embedding dimension:** 128-dimensional feature vector,
- **Triplet loss margin:** $\alpha = 0.20$, chosen iteratively within the established range [0.1, 0.3] shown effective in related literature [62, 63, 64],
- **Optimizer:** Adam, with learning rate 1×10^{-4} [73],
- **Early stopping:** Patience of 5 epochs (model reverts to best checkpoint),
- **Validation:** 1 specimen withheld per laminate (4 in total).

After convergence, the full dataset was passed through the trained model to generate embeddings for each scan. Because the training dataset was small, all data was required to adequately train the model. The entire dataset was therefore used to compute the full similarity matrix based on cosine distance between L2-normalized embeddings (Figure 4.9). For performance analysis, the matrix was subsetted to include only the rows and columns corresponding to test samples. This subset was then Min–Max normalized on a per-row basis to highlight relative similarity patterns

(Figure 4.10). The normalization procedure mirrors the one used in the SPH method and facilitates comparison across methods.

In Figure 4.10, the row-wise Min–Max normalized CNN similarity matrix reveals distinct intra-laminate clustering for laminate 1 and laminate 4. These results are consistent with expectations, as laminate 1 represents the pristine condition and laminate 4 contains the most severe manufacturing defects, making them the most dissimilar from the other groups. Additionally, moderate inter-laminate similarity is observed between laminate 1 and laminate 3, suggesting the network struggles to fully disentangle structural patterns in cases where defect types (e.g., oil contamination in laminate 3) are partially obscured in the available C-scan data. Compared to the SPH method, the CNN model demonstrates improved clustering specificity for laminate 2, with clearer separation from laminate 4. This indicates that the deep features learned by the CNN are more sensitive to certain structural differences, but still face challenges in cases of limited defect visibility or overlapping feature representations across laminates.

Samples exceeding a similarity threshold of 0.80 were considered structurally similar. To ensure stable downstream model training, if a sample had fewer than three matches, the top three most similar samples were used regardless of threshold. This process yielded the SIM-DICF, which serves as input for the prognostic modeling discussed in Chapter 5.

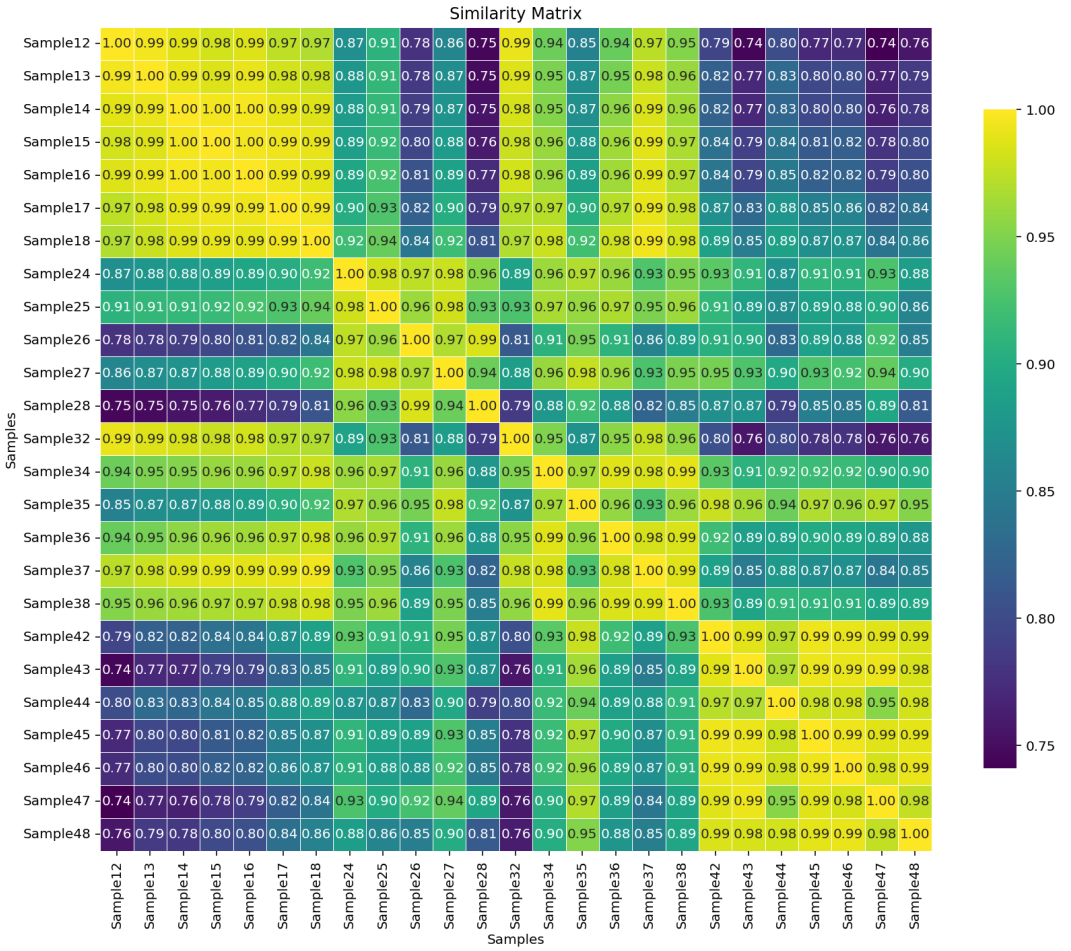


Figure 4.9.: Raw similarity matrix of full dataset, computed using cosine similarity between CNN embeddings.

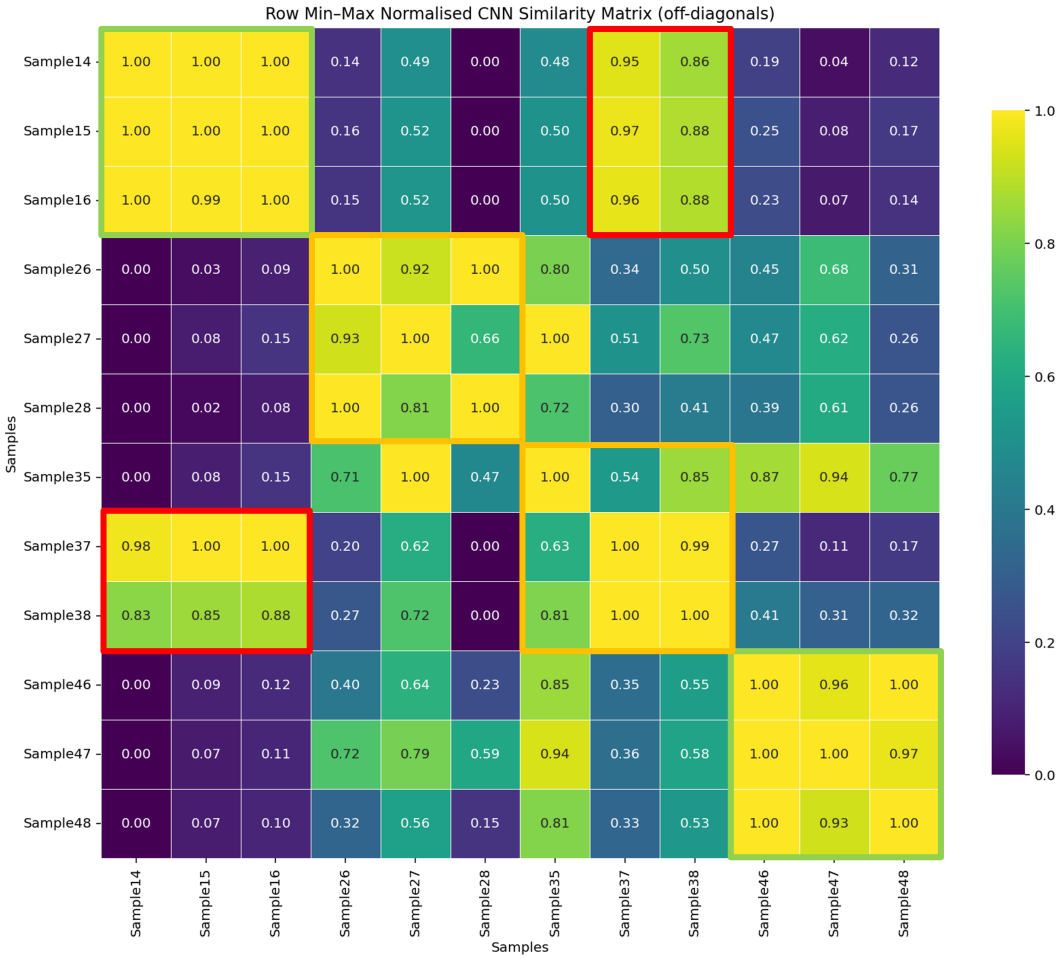


Figure 4.10.: Row-wise Min–Max normalized CNN similarity matrix.

5. Prognostic Results

This chapter presents the results of the proposed similarity-informed methodology, with a focus on evaluating the impact of incorporating past state information, derived from NDI, on RUL predictions of composite specimens. It evaluates the performance of the SIM-DICF, obtained from the SPH or CNN methodology, against the baseline approach that ignores material quality variability. The following sections analyze the derived similarity matrices from Chapter 4 and quantify their influence on prognostic uncertainty reduction using an existing stochastic model.

To generate RUL estimates, the SLHSMM was used [35]. This model builds on the foundations of the Hidden Semi-Markov Model (HSMM) [75], which is well-suited for modeling degradation processes due to its ability to account for variable-duration states, a critical feature when modeling time-to-failure behavior. Unlike traditional HMMs, which assume that state durations follow a geometric distribution, HSMMs allow for more realistic, explicitly defined duration distributions. This added flexibility makes HSMMs particularly effective for capturing complex time-based dynamics such as progressive damage accumulation in composite materials [76].

The SLHSMM introduces and additional layer by integrating similarity information directly into the model's transition and emission mechanisms. This similarity prior, derived from either SPH or CNN clustering, is used to inform the model about which degradation trajectories are most relevant for each test sample. This categorization enables the model to weigh similar degradation histories more heavily, improving both predictive accuracy and confidence estimation.

All SLHSMM modeling and training tasks were carried out externally by the supervising research group. This thesis focuses on the experimental case study, similarity analysis, and interpretation of the results. For a complete explanation of the SLHSMM framework and its mathematical details, the reader is referred to [35, 56].

The results in this chapter aim to assess whether integrating past state knowledge through similarity learning contributes to more accurate and reliable RUL predictions. Two similarity-informed datasets, one derived using SPH features and the other using CNN embeddings, were used to train separate SLHSMM models. These were then compared to a baseline model trained on the DICF without any similarity filtering.

An example of a discretized similarity-informed training dataset is illustrated in Figure 5.1. This figure depicts the degradation trajectories of the three most similar training samples selected for Test Sample28, based on the SPH similarity matrix (Figure 4.6). As fewer than three training samples exceeded the predefined similarity threshold of > 0.80, the top three most similar specimens: Sample26, Sample46, and Sample48 were selected to ensure convergence of the SLHSMM model. Each curve represents the discrete health state evolution as input to the SLHSMM, with the test sample shown in dashed red for reference.

This discretized input serves as the conditionally tailored training set for the SLHSMM, which uses structural similarity to weight degradation histories. The model subsequently produces a probabilistic RUL estimate for the test sample by learning from patterns within this subset. This procedure was applied consistently for both the SPH and CNN methods, forming the basis for the comparative performance evaluations presented in this chapter.

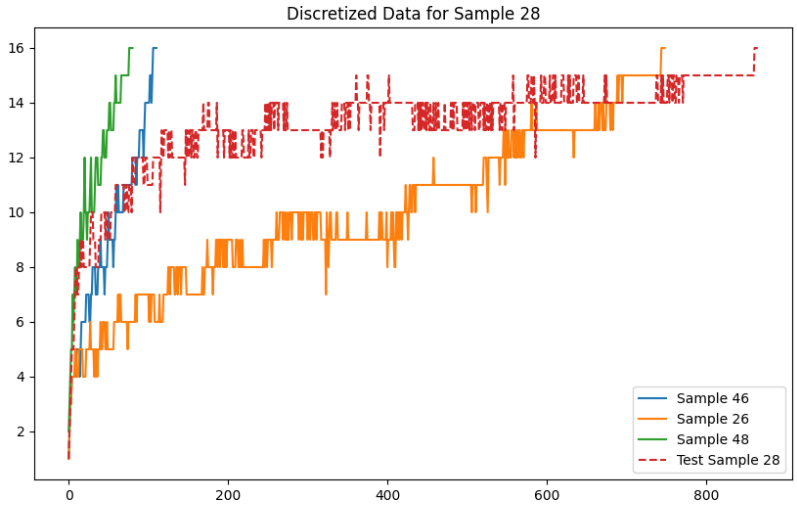


Figure 5.1.: Discretized degradation histories to train the SLHSMM for Sample28, based on SPH similarity. The dashed red line shows the test trajectory.

5.1. RUL Prediction Results Using Similarity-Informed Prognostics

This section presents the RUL prediction results obtained using the similarity-informed methodology and the SLHSMM. For each of the 12 selected test specimens, the RUL evolution is visualized as a function of fatigue life and compared against a baseline model trained on DICF without similarity clustering. Each plot contains three key elements:

- The **True RUL**, shown as a dashed black line, represents the actual remaining cycles until EoL for each test specimen, based on the determined threshold in Section 3.4.3. This degradation history is fixed and identical across both models.
- The **Baseline prediction**, shown in blue, indicates the mean predicted RUL as computed by the stochastic model trained on the complete training dataset, regardless of internal material similarity. The shaded blue area represents the 95% confidence interval associated with the prediction.
- The **Similarity-informed prediction (SPH or CNN)**, shown in orange or green respectively, represents the mean RUL estimate and confidence interval computed using the SIM-DICF, which incorporates past state similarity groupings based on SPH or CNN metrics. This approach aims to improve the reliability of the RUL predictions by conditioning the model on data from structurally similar specimens.

The comparison of these methods illustrates how similarity-informed priors affect the trajectory of predicted RUL and its associated uncertainty. Initial observations show that the SPH-informed method (Figure D.1–D.12) exhibits tighter confidence bounds and improved alignment with the true RUL trajectory for most cases. The CNN method (Figure D.13–D.24) shows mixed results, some predictions seem promising, but also deal with convergence issues, likely due to overfitting risks or limitations in the SLHSMM. The comparison also provides a critical view on the trade-off

between model complexity and performance (Figure 5.2–5.13). To put the results into a comparative perspective the next section details two prognostic performance metrics that were used to quantitatively analyze the improvements of the RUL predictions.

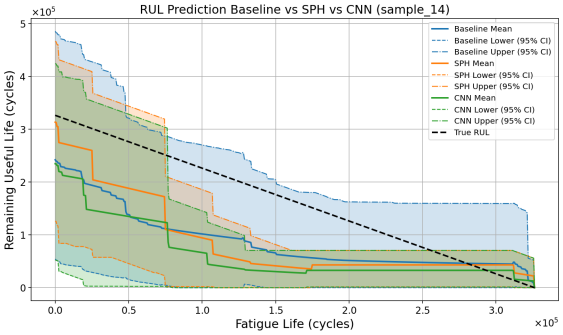


Figure 5.2.: RUL comparison baseline vs SPH vs CNN for Sample14

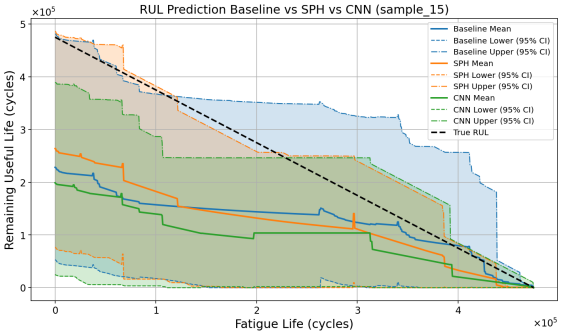


Figure 5.3.: RUL comparison baseline vs SPH vs CNN for Sample15

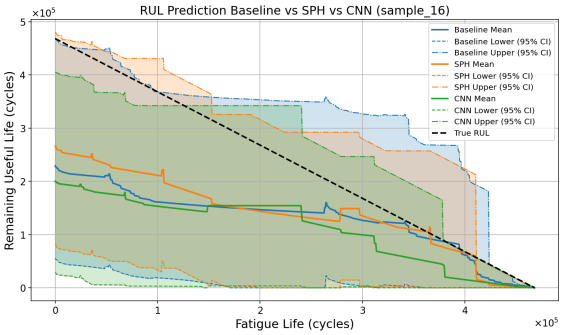


Figure 5.4.: RUL comparison baseline vs SPH vs CNN for Sample16

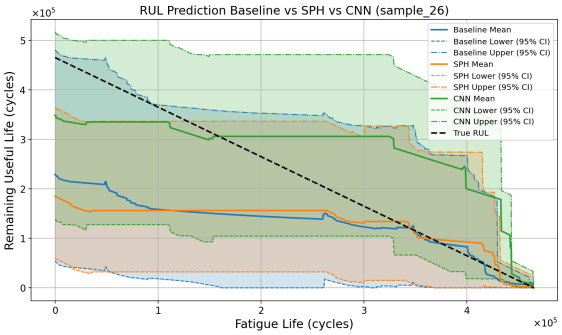


Figure 5.5.: RUL comparison baseline vs SPH vs CNN for Sample26

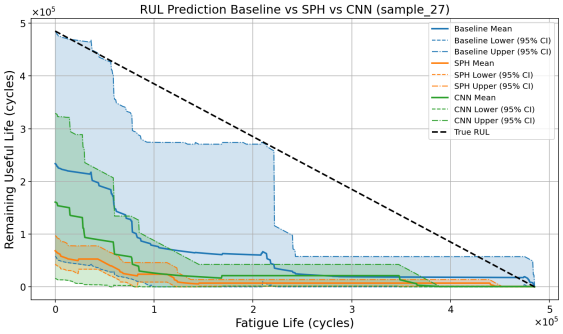


Figure 5.6.: RUL comparison baseline vs SPH vs CNN for Sample27

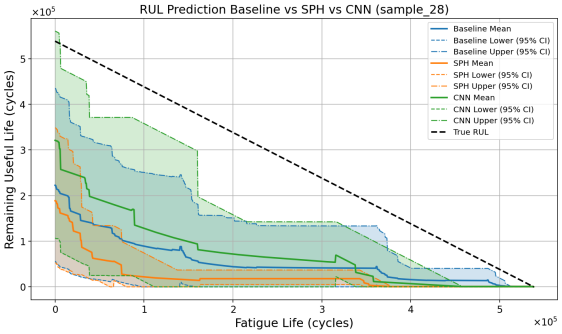


Figure 5.7.: RUL comparison baseline vs SPH vs CNN for Sample28

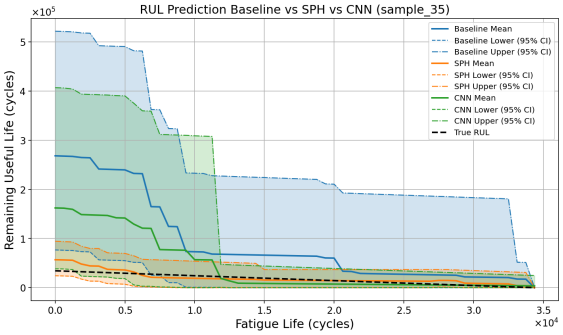


Figure 5.8.: RUL comparison baseline vs SPH vs CNN for Sample35

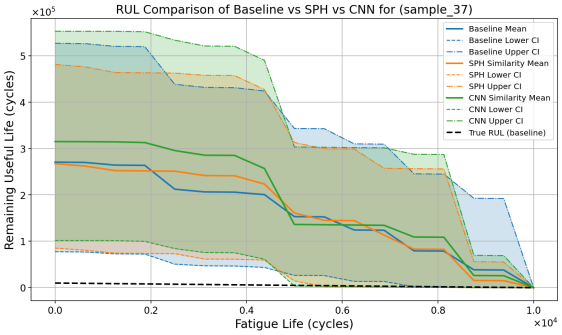


Figure 5.9.: RUL comparison baseline vs SPH vs CNN for Sample37

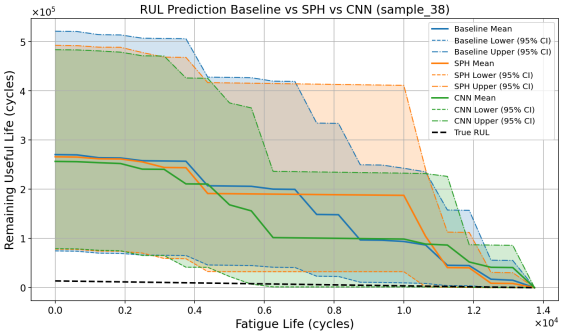


Figure 5.10.: RUL comparison baseline vs SPH vs CNN for Sample38

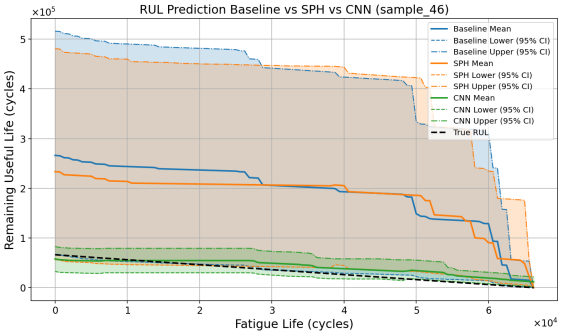


Figure 5.11.: RUL comparison baseline vs SPH vs CNN for Sample46

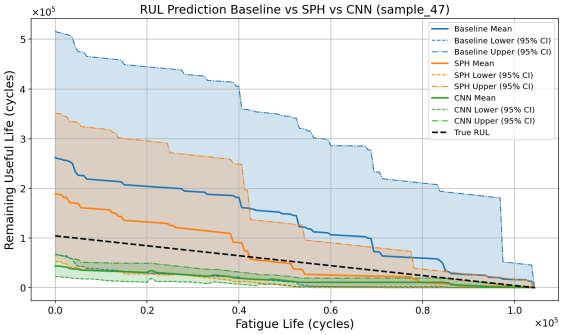


Figure 5.12.: RUL comparison baseline vs SPH vs CNN for Sample47

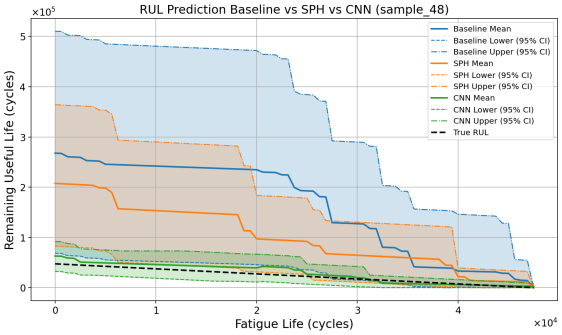


Figure 5.13.: RUL comparison baseline vs SPH vs CNN for Sample48

5.2. Prognostic Performance Metrics

As introduced in Chapter 1, the reliability of a prognostic model is judged not only by its predictive accuracy but also by its ability to express uncertainty, maintain interpretability, and support explainable decision-making. This section focuses on evaluating the RUL predictions produced by the SLHSMM using two complementary metrics: RMSE and Continuous Ranked Probability Score (CRPS). Together, these quantify both prediction accuracy and the probabilistic quality of the forecast.

Root Mean Square Error

RMSE is a widely-used metric for measuring point-prediction accuracy [77]. It captures the average squared deviation between predicted RUL and the true RUL values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5.1}$$

Where $\hat{y}_i$ is the predicted RUL and y_i is the true RUL for the test instance i. RMSE penalizes larger errors more severely than smaller ones, making it sensitive to major deviations. It is suitable for capturing the alignment of predicted trajectories with ground truth [78, 79].

Continuous Ranked Probability Score

While the RMSE evaluates the mean prediction accuracy, CRPS measures the quality of the full probabilistic forecast. It evaluates both:

- **Accuracy:** how closely the predicted RUL distribution is centered around the true RUL value,
- **Sharpness:** how confidently (i.e. narrowly) the model predicts the future RUL.

For a test sample with predicted cumulative distribution function $F_i(x)$ and true RUL y_i, CRPS is given by [80]:

$$\text{CRPS}_i = \int_{-\infty}^{\infty} \left(F_i(x) - \mathbb{I}\{x \geq y_i\}\right)^2 dx \tag{5.2}$$

The indicator function $\mathbb{I}\{\cdot\}$ equals 1 if $x \geq y_i$, and 0 otherwise. Lower CRPS values indicate better-calibrated probabilistic outputs and reduced uncertainty [77, 80]. Unlike RMSE, CRPS rewards predictions that are both accurate and confident, making it suitable for evaluating uncertainty-aware prognostic models.

Performance Overview

To compare the relative performance clearly, Tables 5.1 and 5.2 summarize the improvements achieved by both SPH and CNN methods relative to the baseline. The next section will discuss the results in more detail, highlighting the improvements and trade-offs.

Table 5.1.: RMSE performance score between Baseline, SPH-method and CNN-method.

ID	RMSE Base	RMSE SPH	RMSE CNN	Δ RMSE % (Base vs SPH)	Δ RMSE % (Base vs CNN)
Sample14	91847	92147	115409	-0.33	-25.65
Sample15	141778	133380	167651	5.92	-18.25
Sample16	133955	117709	146512	12.13	-9.37
Sample26	133706	144423	102589	-8.02	23.27
Sample27	204078	262107	242428	-28.43	-18.79
Sample28	243324	279846	215562	-15.01	11.41
Sample35	106581	7476	55180	**92.99**	48.23
Sample37	173772	180714	212320	-3.99	-22.18
Sample38	175713	182844	154766	-4.06	11.92
Sample46	165285	153576	11773	7.08	**92.88**
Sample47	89249	34847	39637	**60.96**	**55.59**
Sample48	157786	90423	8091	42.69	**94.87**
Mean				+13.49	+20.33
Median				+ 2.80	+11.67

Table 5.2.: CRPS performance score between Baseline, SPH-method and CNN-method.

ID	CRPS Base	CRPS SPH	CRPS CNN	Δ CRPS % (Base to SPH)	Δ CRPS % (Base to CNN)
Sample14	55689	63469	82417	-13.97	**-48.0**
Sample15	81520	81973	112018	-0.56	-37.41
Sample16	76532	67383	90739	11.95	-18.56
Sample26	75781	85021	59770	-12.19	21.13
Sample27	156403	225484	203655	**-44.17**	-30.21
Sample28	193240	239086	170874	-23.72	11.57
Sample35	50984	4099	21055	**91.96**	**58.7**
Sample37	110299	117177	137773	-6.24	-24.91
Sample38	110451	118996	97517	-7.74	11.71
Sample46	108749	100904	7437	7.21	**93.16**
Sample47	47912	16774	31467	**64.99**	34.32
Sample48	95967	54385	4266	43.33	**95.55**
Mean				+ 9.24	+13.91
Median				- 3.40	+11.64

5.3. Interpretation and Evaluation of Prognostic Metric Results

To quantify the benefits of managing past state similarity in a SLHSMM prognostic framework, two complementary performance metrics were employed: the RMSE and the CRPS. The RMSE measures the accuracy of the predicted mean RUL, indicating how closely the predicted trajectory aligns with the true RUL. The CRPS, on the other hand, assesses both accuracy and uncertainty, quantifying the sharpness and calibration of probabilistic predictions [77, 80]. By combining these metrics, this study evaluates not only the alignment with the true RUL of the predictions but also their reliability and confidence, crucial for effective decision-making in real-world aerospace prognostics.

5.3.1. Interpretation of SPH Results

The SPH-informed similarity method shows consistent improvements in predictive performance when compared to the baseline (see Table 5.1 and Table 5.2). Notably, the SPH-based approach achieves an average RMSE improvement of approximately 13.49 %, indicating that incorporating histogram-based similarity effectively aligns the training data with structurally similar degradation paths. Particularly significant improvements are observed for specimens such as Sample35, Sample47, and Sample48, where the reduction in RMSE exceeds 40 %. These improvements are attributed to the SPH method's ability to capture meaningful internal structural differences from the Dolphicam images, resulting in more targeted and representative training subsets.

However, some samples, notably Sample27 and Sample28, show substantial deterioration (-28.43 % and -15.01 %, respectively). This negative performance can likely be traced back to the limited depth visibility in Dolphicam scans, especially in laminates containing defects placed at deeper ply interfaces. Consequently, specimens from laminate 3, which appeared superficially similar to the pristine laminate 1 due to incomplete scanning, were incorrectly grouped, adversely affecting performance.

In terms of probabilistic predictions, the CRPS scores reveal a similarly positive yet nuanced picture. The mean CRPS improvement is 9.24 %, reflecting overall better probabilistic calibration. However, the median slightly worsens (-3.40 %), highlighting that improvements are not uniformly distributed. Samples with distinct and well-captured internal defects benefit significantly from sharper probabilistic forecasts (Sample35, Sample47, and Sample48), while others show slight deterioration due to insufficient defect detection or similarity misclassifications. Thus, the SPH method, while broadly beneficial, emphasizes the importance of high-quality, comprehensive NDI scans for effective uncertainty management.

5.3.2. Interpretation of CNN Results

The CNN-informed method, leveraging deep learning for structural similarity, exhibits a more volatile but generally positive improvement trend. The average RMSE improvement stands at 20.33 %, surpassing the SPH-based method. Notable improvements, particularly in Sample46 (92.88 %) and Sample48 (94.87 %), demonstrate the CNN's powerful capability to identify structurally distinct and highly representative subsets, enabling highly accurate mean predictions. However, the performance is not consistent, with also significant deteriorations observed in cases such as Sample14 (-25.65 %) and Sample27 (-18.79 %). These negative outcomes suggest sensitivity of the CNN method to dataset size, data quality, and the effectiveness of augmentation, potentially causing misclassification or embedding instability.

The probabilistic calibration, as measured by the CRPS, follows a similar trend: CNN achieves an average improvement of 13.91 %, yet individual sample results are uneven. For samples where the CNN successfully identified meaningful structural clusters (Samples35, 46, and 48), probabilistic predictions dramatically improved, reflected by sharp confidence intervals and high accuracy. Conversely, significant increases in CRPS for Sample14 (-48.00 %), Sample15 (-37.41 %), and Sample27 (-30.21 %) indicate overconfident or misaligned predictions, likely due to sparse training data and limited variability within the CNN-learned embedding space. Overall, these results underscore the CNN approach's strong potential but highlight its sensitivity and the need for structured data augmentation or more extensive training sets.

6. Discussion

6.1. Manufacturing and Scanning Process

During the manufacturing and scanning stages, several practical considerations emerged that influenced the overall experimental outcomes and the subsequent prognostic analyses. Firstly, an unintended asymmetry in the ply stacking of laminate 4 was created. Despite this minor misalignment, no significant influence on the observed fatigue life or the computed similarity scores was detected, suggesting a tolerance for small geometric deviations in composite manufacturing for prognostics applications.

Another critical aspect involved the trade-off between surface texture and measurement quality. The textured surface greatly enhanced speckle application for DIC, providing clear displacement measurements [57]. However, this texture negatively impacted ultrasound imaging during NDI scanning, significantly reducing defect visibility in deeper layers. Notably, defects between plies 15 to 18 became indistinguishable in the scans. Especially for laminates 3 and 4, containing defects in these plies, this led to limited defect detection and potentially incorrect similarity groupings, influencing predictive reliability. For future work, ensuring smooth laminate surfaces on both sides or performing multiple scans from opposing surfaces could substantially enhance the reliability and depth of ultrasound inspections.

Furthermore, delamination and fiber pull-outs resulting from the mechanical drilling process were observed, predominantly at the exit surface of drilled holes. Although the observed delaminations were minor and unlikely to compromise structural integrity significantly, they could still introduce local uncertainties in DIC and fatigue performance assessments. Using water-cooled CNC drilling methods is recommended for future experiments to mitigate fiber damage, further enhancing measurement quality, improving specimen consistency and reducing local stress concentrations around drilled areas [5].

6.2. Fatigue Testing

The fatigue testing campaign provided crucial insights into the degradation behavior of composite materials under cyclic loading. Although constant amplitude loading is typically preferred for generating reliable and consistent fatigue data, practical challenges arose from the PID settings of the fatigue testing machine. These settings were insufficiently tuned, leading to slight variability in loading amplitude. This underscores the need for more precise PID optimization to achieve stable loading conditions, ensuring that fatigue performance can be consistently evaluated and accurately related to material quality differences [12].

A notable finding from the fatigue experiments was the strong correlation of degradation with cumulative high-amplitude cycles rather than merely the total number of fatigue cycles. Specifically, high-amplitude loading dominated the early onset of material degradation, significantly accelerating damage progression. This phenomenon was illustrated by substantial lifetime variability within identical laminate specimens under nominally similar loading conditions. For example, the

peak load amplitude of 16 kN was reached only after approximately 30 seconds, corresponding to around $30s \times 7Hz \approx 210$ initial subcritical cycles (see Figure 3.12). While determining fatigue parameters, identical laminates tested under comparable conditions ($16kN$ amplitude, $R = 0.1$, $f = 7\ Hz$) produced markedly different lifetimes; Sample44 exceeded 218,000 cycles (DIC image every 500 cycles), whereas Sample43 reached only approximately 10,000 cycles (DIC image every 1000 cycles).

Moreover, variability was also noted between specimens within the same laminate under identical conditions: Sample26 failed after approximately 823,000 cycles, while Sample28, from the same laminate, exceeded 1.4 million cycles without failure (see Table 3.4). Such variances illustrate both the cumulative damage effects of frequent high-amplitude cycles and the inherent stochasticity within composite fatigue behavior [13]. Thus, carefully optimizing test parameters, particularly PID settings and the frequency interval for DIC measurements, is essential to capture representative fatigue progression curves reliably.

An important fundamental assumption underpinning the developed methodology was that materials with similar structural quality would exhibit comparable degradation behavior. The fatigue testing outcomes provided valuable insights regarding this assumption. Initially, it was anticipated that laminate 4, containing a mixture of defect types (including perforated patches leading to localized delaminations and voids, alongside oil contamination affecting interlaminar bonding), would exhibit the worst fatigue performance. However, the results revealed laminate 3, containing exclusively oil defects, clearly exhibited the poorest fatigue performance. This strongly indicates that oil contamination is the dominant factor accelerating material degradation. Moreover, laminate 4, despite containing both oil and perforated patch defects, performed better than laminate 3 due to the lower quantity of oil defects.

Notably, perforated patches alone did not significantly impact fatigue performance. Samples 27 and 28 from laminate 2 (containing only patch defects) endured around 1.4 million cycles without failure, further confirming the lesser impact of these defects compared to oil contamination. Although, aside from Sample26, other specimens from laminates 1 and 2 were not tested until failure, preliminary observations strongly suggest that patch defects alone have a significantly smaller impact on fatigue performance compared to oil-related defects.

Overall, these findings emphasize that oil contamination strongly correlates with accelerated degradation. Therefore, the assumption that similar structural qualities lead to analogous degradation behaviors remains valid but requires careful consideration of defect type and relative severity. This nuanced understanding reinforces the importance of accurately quantifying internal structural characteristics through NDI to enhance similarity-driven prognostic reliability.

6.3. NDI Results

The Dolphicam ultrasonic scanning highlighted essential factors for successful defect detection and similarity quantification. High consistency across samples was achieved by adhering to identical scan parameters. A critical limitation emerged in detecting shallow defects beneath textured surfaces. The textured face introduced wave scattering and energy attenuation, particularly in the upper layers, limiting the ability to capture all the defects in the scans.

Going forward, establishing an optimized scanning protocol that maximizes defect detection across all layers and maintains high consistency across specimens is recommended to enhance the similarity-based prognostic methods. Future studies should adopt a standardized scanning protocol that balances spatial coverage with scan depth resolution, preferably on smooth surfaces.

Multi-angle or bi-face scanning could improve defect visibility, although this would require more extensive scanning efforts.

6.4. Similarity Methods

The comparative evaluation of similarity techniques highlighted both opportunities and constraints for image-based grouping in composite health monitoring. Three distinct methods: SSIM, SPH, and CNN were explored for quantifying internal structural similarity. SSIM, while conceptually attractive, proved limited in practice due to its sensitivity to global misalignment, performing poorly when analyzing entire C-scans.

SPH performance heavily depended on careful preprocessing involving Gaussian blur and thresholding, critical to suppress noise and enhance defect contrast. Additionally, masking the scans prevented the similarity computation from becoming skewed by dominant low-intensity regions, thus focusing the analysis on defect-specific areas [52, 59]. However, the SPH results showed a interpretable and stable performance, especially given the limited size of the available dataset. Clear performance gains were observed when structural differences could be quantified effectively through histogram analysis. Nonetheless, SPH's reliance on manually selected parameters presents a limitation for broader applicability, requiring retuning for different materials or new scanning protocols.

The CNN-based similarity method demonstrated effective clustering of laminates and provided substantial performance gains for samples with well-defined structural similarity. Unlike the SPH method, which relies on manually tuned histogram parameters that may not generalize across datasets, the CNN model offers greater adaptability. As it learns to recognize patterns directly from the data, it can be more readily applied to different datasets without requiring extensive parameter retuning, offering a clear advantage in scalability and reuse.

However, this flexibility comes with trade-offs. The CNN approach exhibited higher volatility in its predictions, primarily due to the sensitivity of the learned embeddings to subtle changes in the training distribution. Although overfitting was mitigated through extensive data augmentation, performance variability was still observed. Further improvements could be achieved by training on domain-specific datasets instead of relying solely on ImageNet-pretrained weights. Nevertheless, in many prognostic applications, acquiring large, high-quality datasets is often infeasible due to experimental constraints, limited failure cases, and costly inspection procedures. This data scarcity remains a key limitation in fully leveraging the potential of CNN-based similarity learning in practical prognostic applications.

Future research may benefit from exploring hybrid approaches that integrate Spatial Pyramids within CNN's, to combine grid pooling with pattern recognition, leveraging both their respective strengths [64, 66].

6.5. Prognostic Modeling

The SLHSMM effectively leveraged similarity information to condition its RUL predictions, clearly outperforming the baseline method when appropriate similarity subsets were selected. Both SPH and CNN methods demonstrated improved prediction reliability, validated by the reductions in RMSE and CRPS scores. However, CNN exhibited superior mean and median performance improvements alongside notable volatility, highlighting the risks associated with relying on potentially unstable similarity clusters.

The SLHSMM offered powerful flexibility but remained essentially a black-box model. Given that all modeling work was carried out externally by the supervising group, insights into internal parameter behavior were limited. Additionally, a minimum of three training histories was required for reliable model convergence. In cases of extremely tight clusters (Samples35, 46, and 48 for CNN-method), the model became overly confident, overshooting RUL estimates and producing non-converging uncertainty bounds. While the similarity-informed SLHSMM approach clearly improves RUL prediction reliability, its potential is currently limited by convergence constraints and sensitivity to training data characteristics.

Notably, performance was severely impacted by incomplete NDI layer coverage in defect regions, leading to misclassifications of laminate 3 samples being similar to laminate 1, while performances were 20 k versus 400 k respectively. This underscores the critical importance of complete, reliable NDI data for similarity-informed prognostic modeling.

7. Conclusions and Recommendations

This thesis developed and validated a novel uncertainty management framework addressing past state uncertainties in aerospace-grade composite materials through similarity learning techniques. The inherent complexity and variability introduced by manufacturing processes, particularly in composite structures, pose significant challenges to reliable RUL predictions. Addressing this gap, the study introduced an approach leveraging advanced ultrasonic NDI via Dolphicam2 to quantify internal material defects such as voids and oil contaminations. Subsequently, two distinct similarity methods-SPH and CNN-were employed to effectively categorize specimens based on structural similarity. These similarity informed categorizations were then integrated into the SLHSMM to enhance predictive accuracy and reduce uncertainty in prognostic models, ultimately providing more reliable decision-making for aerospace maintenance scheduling.

The major contributions of this thesis include:

- A structured experimental campaign validating the novel similarity-informed prognostics methodology using realistic defect scenarios.
- Development and detailed assessment of SPH and CNN similarity methods optimized specifically for Dolphicam imaging data.
- Demonstration of the benefits of incorporating similarity-informed past state uncertainty management into stochastic prognostic modeling, notably reducing predictive uncertainty and improving accuracy.
- A comprehensive analysis of the strengths and limitations of SPH and CNN, providing clear guidance for their practical deployment.

7.1. Conclusions

To comprehensively evaluate how uncertainty management for past state uncertainties could be achieved through similarity learning techniques and Dolphicam data to improve the reliability of RUL predictions, three sub-questions were defined (Section 2.5.1). These sub-questions break down the main challenge into targeted areas of investigation, focusing respectively on: the influence and characterization of material quality variability, the identification of a suitable similarity method and the effect of incorporating past state uncertainty management on prognostic reliability. By addressing each of these sub-questions in detail, a clear and evidence-based answer to the main research question can be formulated.

Material quality variability, specifically oil contamination, emerged as a critical determinant of fatigue performance. Oil-related defects significantly accelerated degradation compared to other defect types such as artificial voids. Dolphicam data provided an effective means for identifying and quantifying internal structural variations arising from manufacturing defects. Accurate characterization of these variations enabled for effective clustering of structurally similar specimens, directly informing and enhancing the predictive reliability of the prognostic model. Nevertheless, Dolphicam scanning limitations, particularly in detecting defects beneath textured surfaces

highlighted the need for refined scanning protocols to maximize comprehensive defect detection. Importantly, while managing past state uncertainties clearly enhances the reliability of RUL predictions, it remains crucial to recognize that in real-life aerospace applications, operational loading conditions play an equally—if not more—critical role in fatigue performance. Variations in load amplitude, frequency, and environmental exposure can significantly influence degradation behavior. Therefore, future prognostic frameworks should complement past state uncertainty management with well-defined strategies for addressing operational and environmental variability to ensure truly reliable life predictions in-service.

In exploring the most suitable similarity method for grouping composite specimens based on Dolphicam data, the study demonstrated that both SPH and CNN methodologies effectively leveraged Dolphicam imaging data, each presenting distinct strengths and limitations. The SPH method offered consistent performance across various cases, particularly suitable for datasets with clearly distinguishable structural differences. The CNN approach, despite exhibiting higher volatility due to limited dataset size and sensitivity to training data quality, demonstrated great potential for capturing complex structural variations, achieving substantial accuracy improvements in selected cases. The CNN-based similarity method is favored due to its inherent adaptability. Unlike the SPH method, which relies heavily on manually tuned histogram parameters that may not generalize across different datasets, the CNN method learns patterns directly from the data. This ability allows CNN to more readily adapt to varying datasets without extensive parameter retuning, offering a clear advantage in terms of scalability and reusability, provided sufficient data is available or feasible to obtain. Although CNN exhibited higher sensitivity to data size limitations, its powerful feature extraction capabilities, when adequately supported by larger datasets, outperformed SPH in capturing complex structural variations and achieving notable predictive improvements.

Finally, the study examined how managing past state uncertainties through similarity-informed modeling affected the reliability of RUL predictions. The comparative analysis demonstrated clear reductions in prediction error (RMSE improvements averaging 13.49% for SPH and 20.33% for CNN) and improved probabilistic forecasting (CRPS improvements averaging 9.24% for SPH and 13.91% for CNN). These results underscore the substantial benefits of integrating internal structural quality characterization into prognostic modeling, emphasizing that reliable RUL predictions can be achieved by effectively addressing past state uncertainties.

Together, these findings demonstrate that similarity learning techniques, supported by high-resolution NDI data from Dolphicam, provide an effective strategy for managing past state uncertainties and enhancing RUL prediction reliability for aerospace composites. The integration of structural similarity into the SLHSMM enabled the selection of more representative training trajectories, leading to more accurate and confident prognostic outcomes.

In conclusion, the research shows that uncertainty management for past state uncertainties can be effectively realized through a similarity-informed framework that combines ultrasonic imaging with adaptable learning methods like CNN embeddings. This approach enables tailored, reliable RUL predictions by incorporating prior knowledge of material quality into prognostic modeling. Consequently, similarity-informed past state uncertainty management stands out as a valuable advancement in data-driven prognostics, providing significant potential for enhancing aerospace composite maintenance scheduling in practical, real-world applications.

7.2. Recommendations

The findings and limitations identified throughout this thesis offer several directions for future research and development. While the proposed similarity-informed framework demonstrated significant improvements in specifically managing past state uncertainties, there remains considerable potential for refinement and broader applicability. The following recommendations aim to strengthen the methodology and support its practical implementation in real-world aerospace maintenance scenarios.

- **Expanded and Diverse Datasets:** Expand the scope and volume of experimental data, capturing a broader range of defects and structural variations. More extensive datasets will significantly improve CNN generalization, and predictive reliability. (Provided that they are feasible to obtain).
- **Domain Specific Pretraining:** Exploring domain-specific pretraining strategies for CNN-based embeddings, using aerospace composite datasets, to further improve similarity identification and clustering stability.
- **Enhanced Ultrasonic Inspection Protocols:** Develop dual-sided or multi-angle ultrasonic inspection methods to overcome current limitations in defect detection, ensuring comprehensive internal structural characterization across all composite layers.
- **Hybrid Similarity Models:** Explore hybrid similarity methods combining the interpretability of histogram-based features with CNN's advanced pattern recognition to leverage complementary strengths, optimizing both predictive performance and interpretability.
- **Explore model-agnostic approaches:** As the SLHSMM was treated as a black box in this thesis, future work should investigate more transparent and flexible frameworks that can accommodate similarity-based inputs independently of the specific model architecture. This would enhance the general applicability of the methodology and allow for broader experimentation with different prognostic models.
- **Real-world Validation and Industrial Integration:** Conduct extensive validation in real-world aerospace scenarios and operational environments. Demonstrating the CNN model's adaptability and scalability across different structural contexts will confirm its practical utility.

Together, these recommendations lay the foundation for advancing the CNN-based similarity-informed framework into a scalable, and reliable method for managing past state uncertainties. By enabling more reliable RUL predictions, thesis offers a meaningful step towards more intelligent, data-driven, and uncertainty-aware decision-making in aerospace health management.

7.3. Acknowledgments

This research was conducted with the support of the Hydrogen Aircraft Powertrain and Storage Systems (HAPSS) project, which is part of the research and innovation programme Luchtvaart in Transitie, co-funded by the Netherlands National Growth Fund and the European Union.

A. Manufacturing Plan

A.1. Laminate Specifications and Material Selection

In this thesis, 32 cross-ply open-hole composites with embedded defects were fabricated. Each sample measures 400 × 45 mm and includes a centrally located hole with a diameter of 10 mm. The samples were machined from symmetric CFRP plates with an epoxy matrix plates, initially produced as square laminates of 424 × 424 mm. Symmetry layup was chosen to prevent warping and residual stresses due difference in thermal expansion during curing. The panel dimensions selected to maximize the usable length of a 600 mm wide unidirectional 0° prepreg roll, while allowing the required ±45° ply orientations. This sizing also enabled the maximum number of test specimens to be extracted from each laminate. The layup consists of 18 plies arranged in a symmetric stacking sequence: $[[+45, -45]_4 + 45]_s$, deliberately excluding 0° plies. This configuration was chosen to achieve minimal thickness suitable for NDI, enough plies to embed defects and to reduce axial fatigue strength in tension-tension fatigue testing. In such tests, fatigue performance is primarily dominated by the 0° plies, which are absent in this design. Instead, the ±45° laminate provides in-plane shear strength but limited axial strength, as there are no continuous axial fibers to bear load. As as result, the epoxy matrix plays a dominant role in transferring load between fibers. will be the load transferring dominant factor. The material choices listed below are driven by the selection of the prepreg roll and the curing temperature.

The materials selected for the manufacturing process include:

Table A.1.: Overview of materials used in laminate fabrication

Material Type	**Product Name**	**Temperature Range**
Mold build-up sheet	Aluminum plate	Up to 300 °C
Release liquid	Marbocoat 227	Up to 400 °C
Prepreg sheet	DT120 DeltaTech®	Curing range: 80–135 °C
Peel ply	Stitch Ply A	Up to 204 °C
Release film	Perforated WL3700	Up to 121 °C
Breather blanket	Air Weave N10	Up to 204 °C
Seal tape	Black LTS90B	Up to 150–180 °C
Vacuum bagging film	Clear WL5400	Up to 177 °C

A.2. Defect Embedding Methods

In total four laminates were manufactured, of which three laminates were embedded with different defect types to evaluate the impact on laminate behavior:

1. **Pristine:** No defects.
2. **Contaminants:** Localized droplets of silicon oil were applied between plies using a precision micro-liter pipette to ensure consistent volume and spatial separation, minimizing large-scale debonding.
3. **Artificial voids:** Perforated release films were placed between specific ply interfaces to trap air and simulate local delaminations.
4. **Mixed defects:** A combination of silicon oil droplets and perforated release film patches was introduced to create severely contaminated laminates. Though this scenario is less realistic in practice, it serves to investigate defect categorization during NDI.

Defect-inducing materials consist of:

Table A.2.: Overview of defect-inducing materials

Defect Type	Material / Product Name
Contaminants	PMX-200 Silicon Fluid 100cst (Silpak®)
Artificial voids and local delamination	Perforated WL3700

A.3. Manufacturing Process

The manufacturing procedure follows established industry standards to ensure consistency and reproducibility:

1. Thaw the DT120 prepreg roll to room temperature. This typically takes approximately 6 hours.
2. Cut the plies to the required dimensions and orientations ($\pm 45°$) using a Gerber CNC cutting machine to ensure consistency and accuracy.
3. Clean the aluminum mold plate thoroughly using propanol or acetone to remove any dust, grease, or other surface contaminants.
4. Apply a double layer of Marbocoat 227 release agent to the mold surface. Wait at least 10 minutes between coats to ensure complete drying and effective demolding after curing.
5. Lay up the DT120 prepreg layers according to desired stacking sequence $[[+45, -45]_4, +45]_s$.
   - A maximum of 3 layers should be placed before debulking.
   - Debulking should be performed under vacuum for 3 minutes per layer. For example, 3 layers require 9 minutes of vacuum debulking to remove entrapped air.
6. Introduce defects in the designated laminates, as specified for the experimental campaign:
   - **Contaminants**: Apply PMX-200 silicon oil at selected ply interfaces.

- **Artificial voids**: Insert perforated release film patches to locally interrupt bonding and trap air.

7. Cover the entire laminate with a peel ply (Stitch Ply A) to create a textured surface suitable for DIC during mechanical testing.
8. Cut and place a perforated release film (WL3700) over the peel ply. Ensure it extends slightly beyond the laminate edges to prevent resin flow into the breather fabric.
9. Lay a breather fabric (Air Weave N10) over the release film. The breather should:
   - Fully cover the laminate.
   - Extend to the vacuum valve location to ensure uniform pressure distribution.
10. Prepare and apply the vacuum bagging film (Clear WL5400):
    - Cut the film large enough to completely enclose the laminate, leaving room for edge sealing.
    - Insert the vacuum valve through a pre-cut hole. Ensure the hole is snug to maintain sealing integrity.
    - Lay the film smoothly over the layup, avoiding folds or wrinkles.
    - Do not apply tension to the film, to prevent stress formation during vacuum application.
    - Seal the edges to the mold plate using LTS90B sealant tape, applying firm pressure to ensure an airtight seal.
11. Connect the vacuum valve to a vacuum pump and draw a vacuum.
    - Check for leaks by monitoring for a pressure drop after switching off the pump.
12. Transfer the fully bagged and sealed assembly into the autoclave for curing, following the prescribed temperature and pressure cycle for the DT120 prepreg.
    - Ramp-up: Heat at a rate of $2°C/\text{min}$ to $120°C$.
    - Dwell: Maintain $120°C$ for 1.5 hours under 6 bar autoclave pressure.
    - Cool down: Cool at $2°C/\text{min}$ to room temperature while maintaining vacuum.
13. Repeat this process to produce four different laminates:
    - **Laminate 1**: No defects (reference) Figure A.1.
    - **Laminate 2**: Contains artificial voids (perforated film patches) Figure A.2.
    - **Laminate 3**: Contains silicon oil contamination Figure A.3.
    - **Laminate 4**: Combination of voids and silicon oil defects Figure A.5.

Note: In Laminate 4, a human error caused the misplacement of ply 17 in the $+45°$ orientation instead of $-45°$, as viewed from the mold layup perspective. Ply 18 was placed therefore laid $-45°$ to recover global stacking symmetry.

A.4. Visual Documentation of Manufacturing Process

Figures A.1–A.5 illustrate the defect embedding process for the various laminate configurations. This is followed by the vacuum bagging and autoclave curing procedures. As shown in Figure A.6, the setup achieves a fully airtight seal with the vacuum pump detached. No visible wrinkling or folding is present in the bagging film, ensuring a uniform pressure distribution and minimizing the risk of stress concentrations during curing. After the autoclave cycle, Figure A.7 shows the cured laminate. Resin bleed is observed through the laminate thickness and along its edges, resulting from the use of a stitch ply and perforated release film. Both facilitate controlled resin flow and surface definition. Following removal of the vacuum bagging materials, the final laminate panel is revealed. In Figure A.8, the glossy mold-contact side of the laminate is shown after demolding. The observed fiber orientation appears to be -45 °, which is an effect of the viewing perspective. Lastly, Figure A.9 presents the top surface of the laminate, where a textured finish is visible. This texture, created by the peel ply, is essential for DIC in later testing [57]. Note that no external damage is induced on laminate.

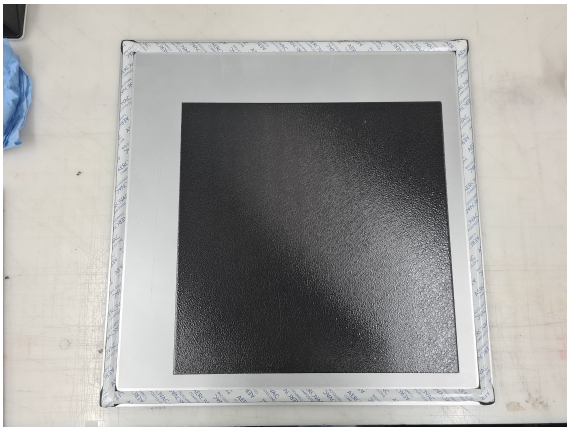


Figure A.1.: Smooth glossy surface after debulking plies of pristine laminate.

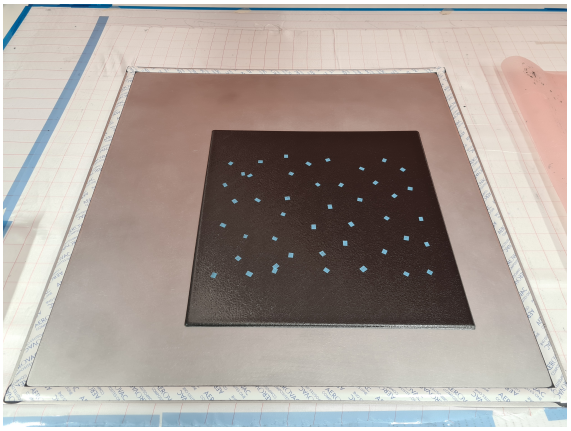


Figure A.2.: Application of perforated release film patches.

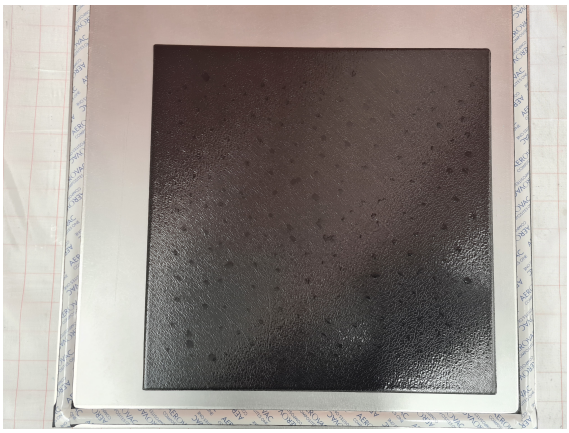


Figure A.3.: Silicon oil droplets applied via precision pipette for contamination.

Figure A.4.: Precision Pipette used to place droplets of silicon oil.

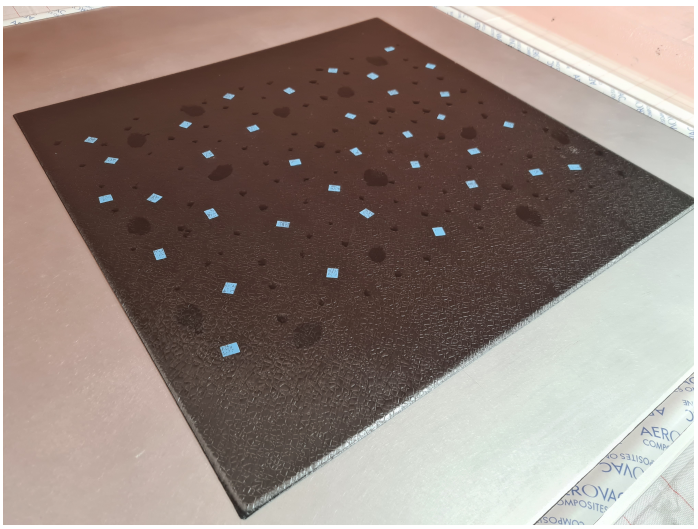


Figure A.5.: Mixed contamination: perforated patches, oil droplets, and fingerprints.

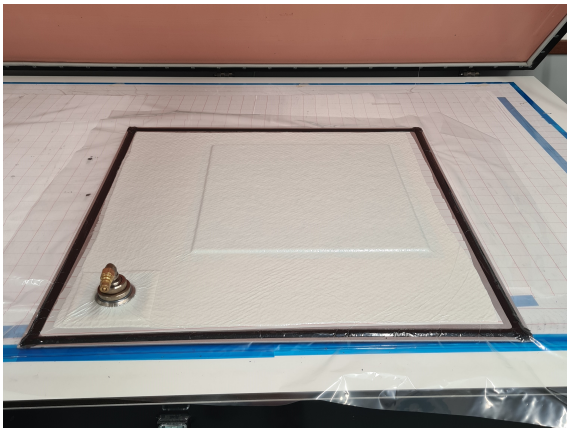


Figure A.6.: Vacuum bag applied and tested for airtightness before curing.

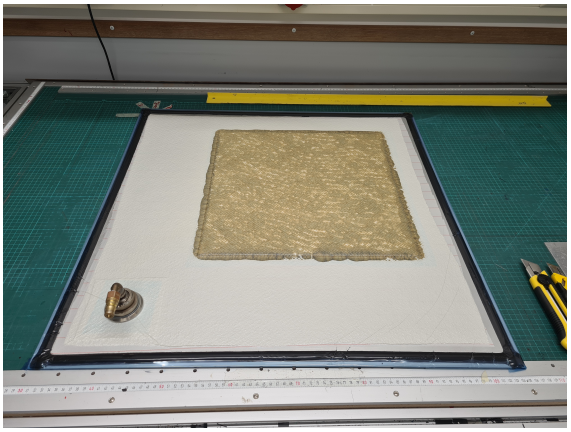


Figure A.7.: Autoclave cured composite laminate.

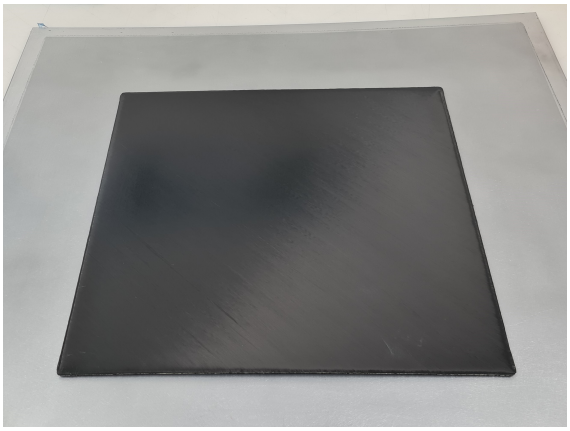


Figure A.8.: A glossy, smooth surface resulting from the mold-side.

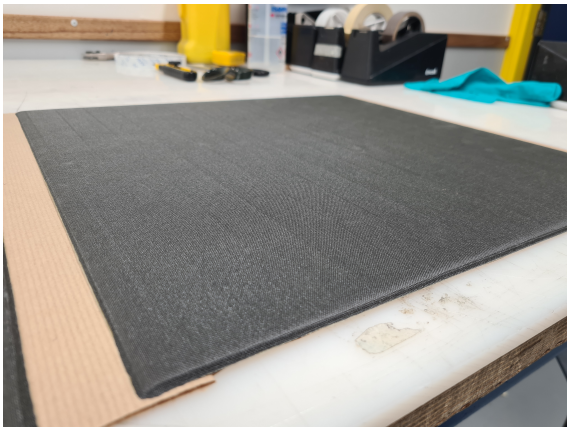


Figure A.9.: A textured, rough surface resulting from the peel-ply.

B. Detailed Dolphicam Parameter Derivation

- **Acoustic Velocity (c):** The sound velocity was set to 3070 m/s, a manufacturer-provided estimate for unidirectional CFRP with epoxy matrices. This value is used to convert Time-of-Flight measurements to physical depth using:

$$t = \frac{2d}{c} \tag{B.1}$$

  where d is the depth (in meters) and t is the round-trip time-of-flight. An accurate velocity value is critical for correct gate positioning and depth resolution in the resulting C-scan images.

- **Transducer Pitch:** The Dolphicam2 transducer uses a matrix array with a pitch of 0.25 mm, defined as the center-to-center distance between adjacent piezoelectric elements. This spacing controls the lateral spatial sampling of the scan. According to the Nyquist criterion, the smallest resolvable feature λ_{min} in the lateral plane is approximately:

$$\lambda_{\text{min}} \approx 2p \tag{B.2}$$

  where $p = 0.25$ mm. Thus, the spatial resolution is approximately 0.5 mm, sufficient to detect defects such as the 7–10 mm artificial voids and silicon oil inclusions introduced during manufacturing. A smaller pitch improves resolution but reduces the total field-of-view and may impact penetration depth.

- **Pulse Length (τ):** Defined as the duration of the ultrasonic burst, calculated by:

$$\tau = \frac{N_{\text{cycles}}}{f_0} \tag{B.3}$$

  where N_{cycles} is the number of sine wave cycles (typically 3) and f_0 is the transducer center frequency (8 MHz). This gives:

$$\tau = \frac{3}{8 \times 10^6} = 375 \text{ ns} \tag{B.4}$$

  Shorter pulse lengths yield better axial resolution but reduce signal energy. A 3-cycle burst provides a balanced trade-off.

- **Pulse Re-Trigger Time ($T_{\mathbf{rep}}$):** The re-pulse interval must be long enough for echoes from the previous pulse to fully decay before the next transmission. To avoid overlap or aliasing, the minimum requirement is:

$$T_{\text{rep}} > 2\frac{d_{\max}}{c} + \tau \tag{B.5}$$

  where $d_{\max} = 3.75$ mm and $\tau = 375$ ns, giving a minimum $T_{\text{rep}} \approx 2.8$ μs. A conservative value of 10 µs was selected to ensure clean signal separation.

- **Trigger Delay ($t_{\mathbf{delay}}$):** Set to 7.4 µs, one the minimum values allowed by the software, to synchronize the start of signal acquisition after transmission. This value balances the gate positioning for full-thickness scanning with system latency constraints.

- **Transmit Elements (Tx Aperture):** A transmit aperture of 4 elements was selected. Increasing the number of transmit elements boosts beam energy and depth penetration but reduces lateral resolution. A value of 4 was found to provide sufficient defect contrast without excessive signal spreading.

- **Signal Averaging ($N_{\mathbf{avg}}$):** Temporal averaging was used to suppress random noise. The improvement in signal-to-noise ratio (SNR) is approximately:

$$\text{SNR Improvement} \approx \sqrt{N_{\text{avg}}} = \sqrt{8} \approx 2.8 \tag{B.6}$$

C. Similarity Method Evaluations

C.1. Feasibility of SSIM

To assess the feasibility of the SSIM as a similarity metric for comparing Dolphicam C-scans, a dedicated Python script was developed that computed pairwise SSIM scores between all preprocessed images. Two forms of result visualization were generated:

1. A raw SSIM similarity matrix (Figure C.1),
2. A row-wise Min–Max normalized version of the same matrix, excluding self-similarity (Figure C.2).

The raw SSIM matrix, Figure C.1, revealed an extremely narrow range of similarity scores, with all off-diagonal values lying between 0.50 and 0.55. Such a narrow range undermines the discriminative power of SSIM in this context, as it fails to capture meaningful structural variations between scans.

To further explore interpretability, a normalized matrix was computed by applying row-wise Min–Max normalization to the off-diagonal elements (excluding self-similarity), see Equation C.1. This process mapped the minimum and maximum similarity scores within each sample row to [0, 1], as shown in Figure C.2.

Given a vector $v = \{v_1, v_2, \ldots, v_n\}$, the Min–Max normalized value v_i^{norm} is computed as:

$$v_i^{\text{norm}} = \frac{v_i - \min(v)}{\max(v) - \min(v)} \tag{C.1}$$

However, this transformation resulted in inflated differences, as the raw values were already clustered tightly. The normalization effectively masked the fact that absolute similarity remained statistically indistinguishable between all samples. These results confirm that SSIM is not a suitable method for global similarity comparison across samples, given the Dolphicam images in this study. The random spatial distribution of defects and absence of feature alignment across samples lead to non-informative SSIM scores. While SSIM could be a valuable tool for local, patch-based comparisons or quality degradation assessment, its application in this context was limited by the nature of the dataset.

Consequently, the SSIM-based similarity method was excluded from further development, and attention was redirected to more interpretable and robust methods, such as histogram-based similarity and CNN-based feature embeddings.

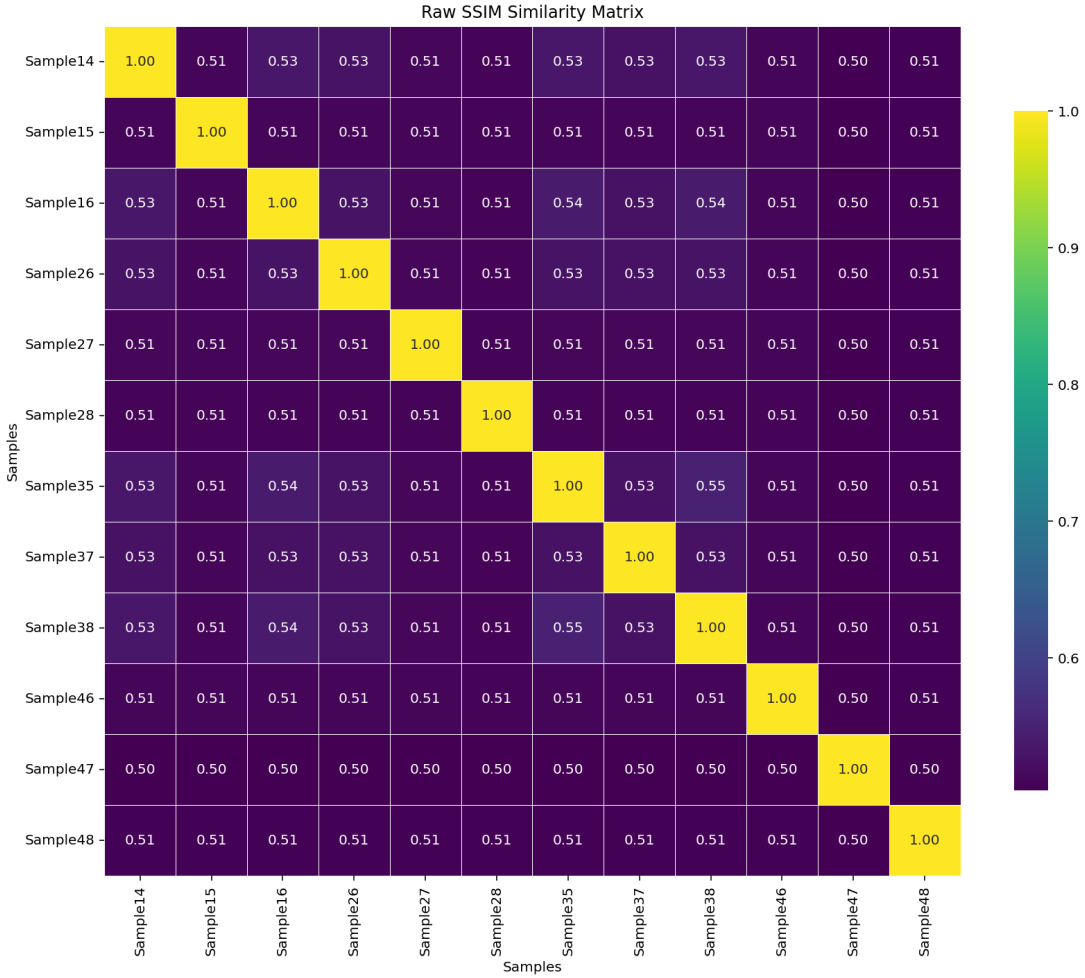


Figure C.1.: SSIM similarity matrix heatmap of all samples. Scores range from 0.50 to 0.55, suggesting minimal structural distinction across the dataset.

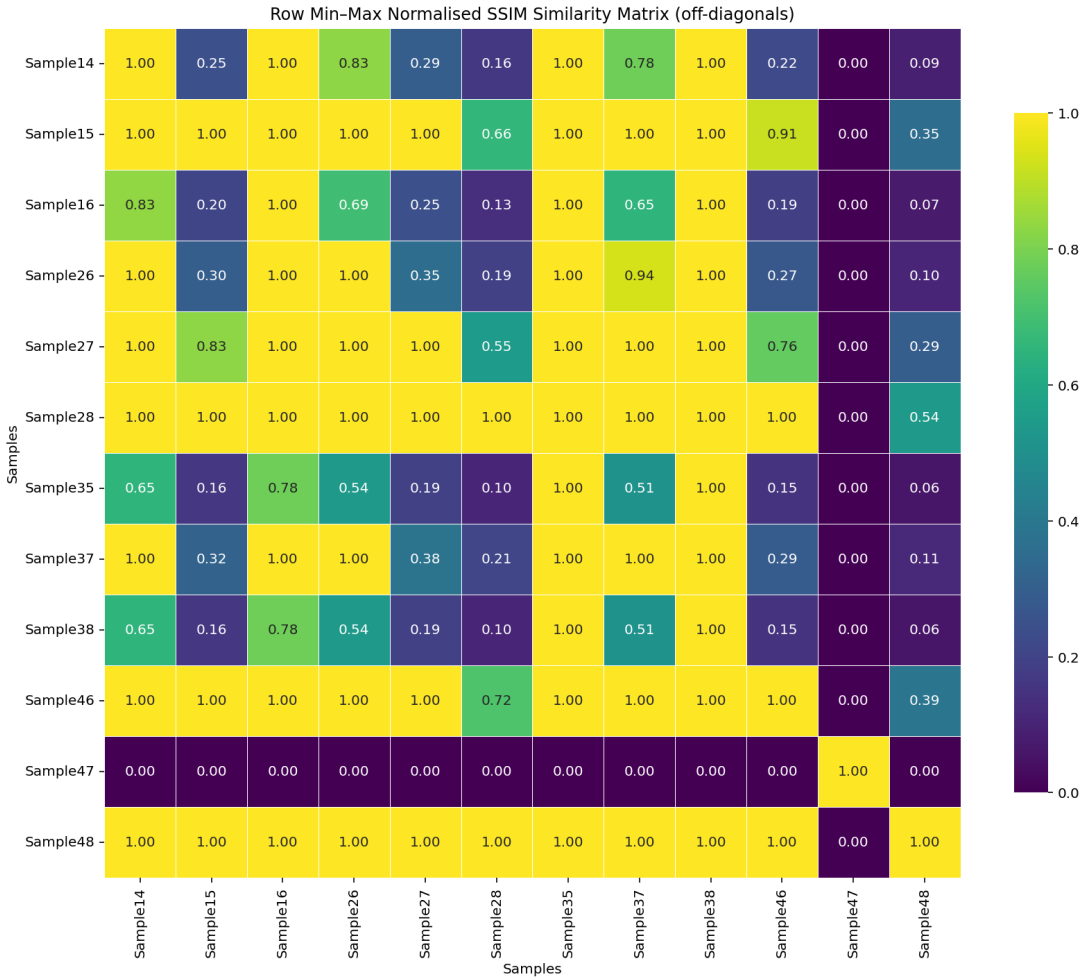


Figure C.2.: Row-wise Min–Max normalized SSIM similarity matrix excluding self-similarity. The normalization distorts relative ranking by artificially amplifying marginal differences.

C.2. Alternative SPH Similarity Metrics

In addition to histogram intersection, several alternative similarity metrics were evaluated for comparing SPH feature vectors, including Bhattacharyya distance (Figure C.3), chi-squared distance (Figure C.4), and cosine similarity (Figure C.5) [52]. However, these methods proved unsuitable for this application. Across the scanned dataset, all three metrics consistently produced near-uniform similarity scores, resulting in similarity matrices with minimal variation and limited discriminative power.

This behavior is attributed to their reliance on global intensity distributions, which in this case were dominated by the low-intensity background shared by all specimens. As a result, these metrics became overly sensitive to shared non-informative features, masking the subtle yet important differences in defect-related regions. Despite their theoretical suitability for normalized histograms, these metrics failed to reflect known structural variations between specimens and were therefore excluded from further analysis.

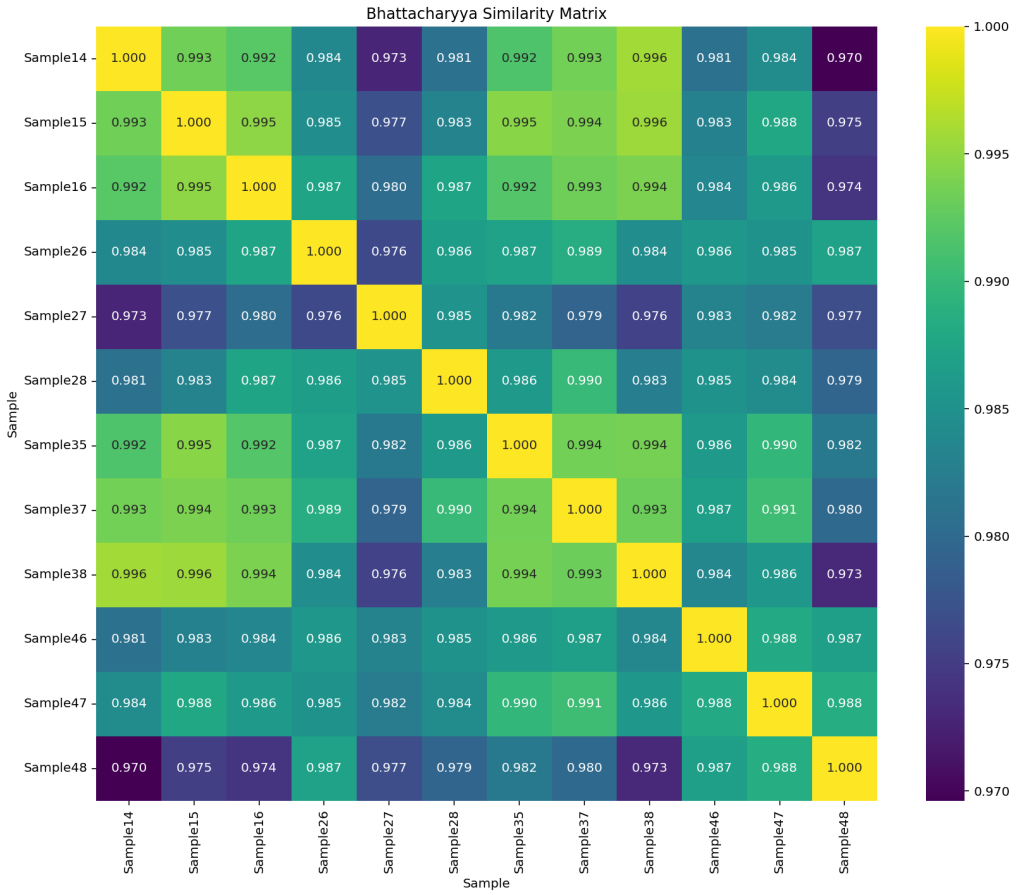


Figure C.3.: SPH similarity matrix heatmap of all samples using the Bhattacharyya distance metric.

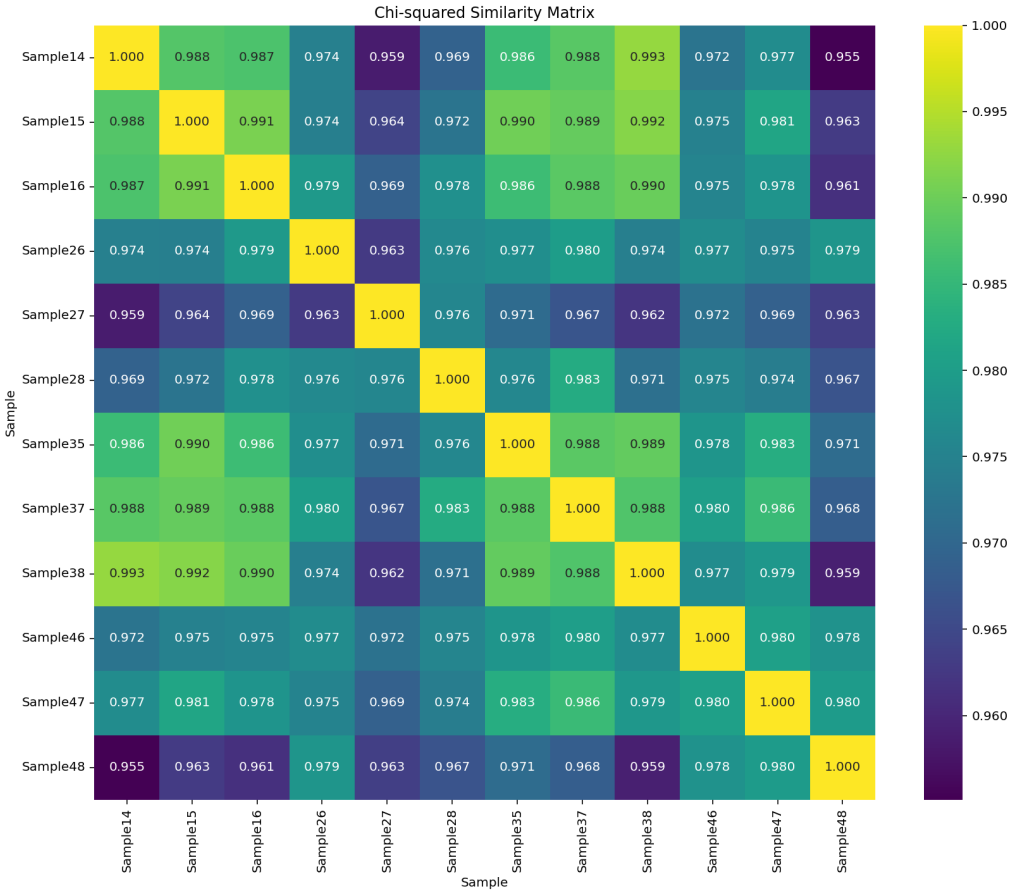


Figure C.4.: SPH similarity matrix heatmap of all samples using the Chi-squared distance metric.

Figure C.5.: SPH similarity matrix heatmap of all samples using the cosine distance metric.

D. RUL Similarity Predictions

For completeness, this appendix presents the individual RUL prediction results using the SPH-informed and CNN-informed SLHSMM methods. These figures complement the main results shown in Figure 5.2–5.13, which provide a direct side-by-side comparison between the two similarity-informed approaches.

Figure D.1–D.12 display the SPH-informed predictions, highlighting generally tighter confidence intervals and better alignment with the actual degradation trajectory in several cases. Figure D.13–D.24 show the CNN-informed predictions, which exhibit excellent accuracy in some bases but also occasional convergence challenges. These individual visualizations support the analysis in Chapter 5, allowing for a more granular inspection of model behavior per specimen.

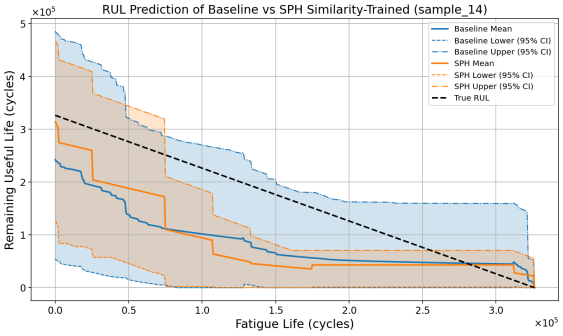


Figure D.1.: RUL comparison baseline vs SPH-method for Sample14

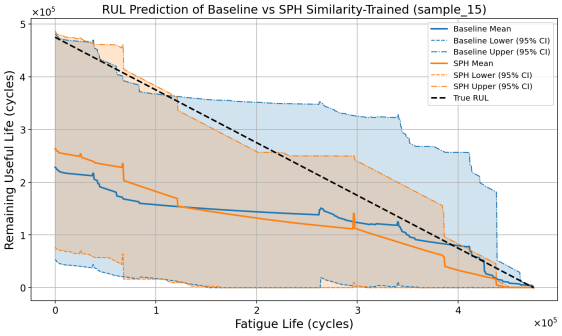


Figure D.2.: RUL comparison baseline vs SPH-method for Sample15

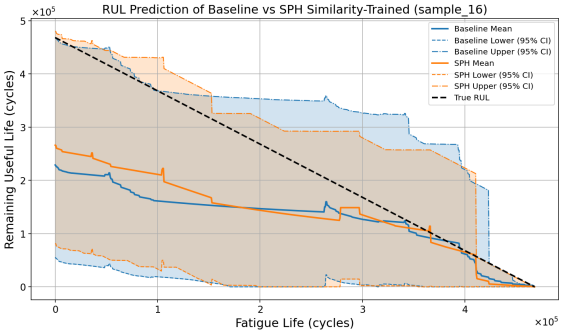


Figure D.3.: RUL comparison baseline vs SPH-method for Sample16

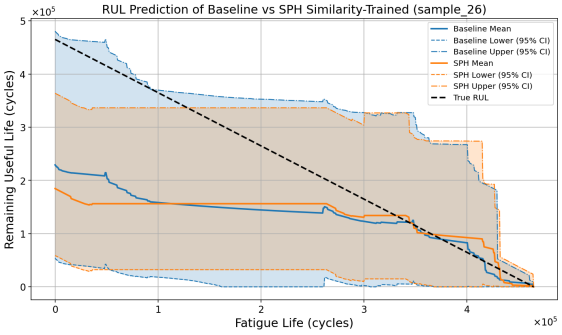


Figure D.4.: RUL comparison baseline vs SPH-method for Sample26

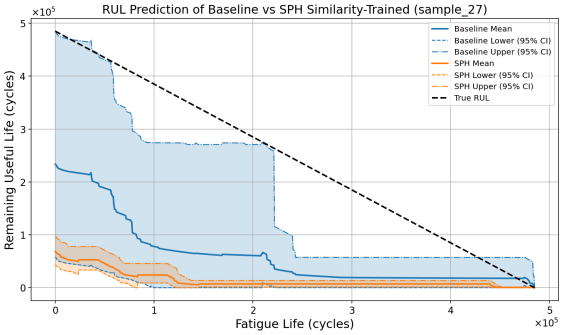


Figure D.5.: RUL comparison baseline vs SPH-method for Sample27

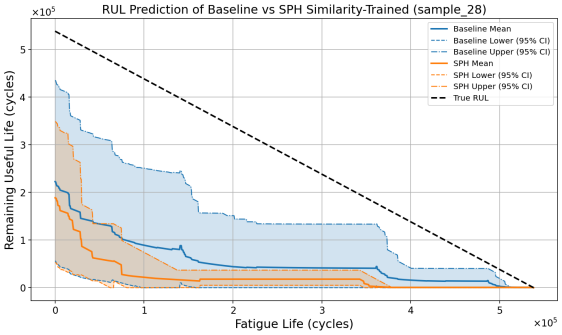


Figure D.6.: RUL comparison baseline vs SPH-method for Sample28

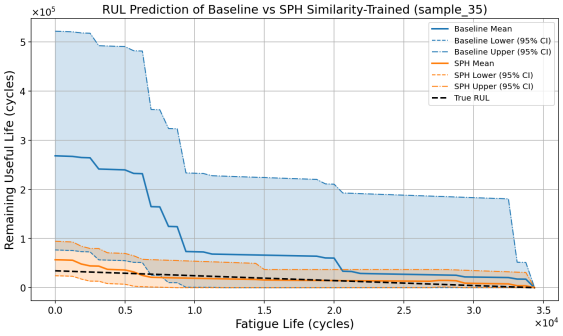


Figure D.7.: RUL comparison baseline vs SPH-method for Sample35

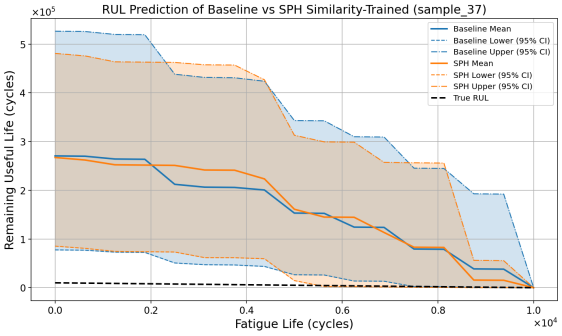


Figure D.8.: RUL comparison baseline vs SPH-method for Sample37

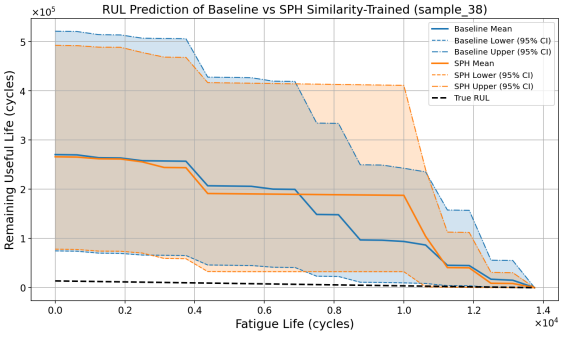


Figure D.9.: RUL comparison baseline vs SPH-method for Sample38

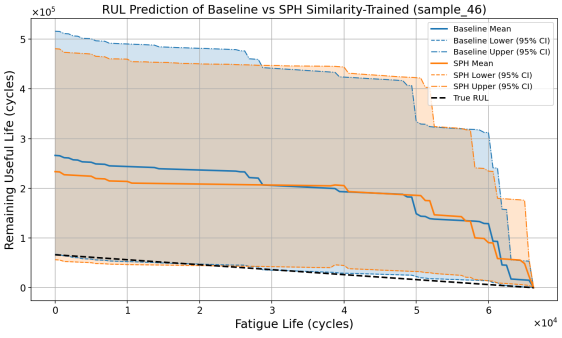


Figure D.10.: RUL comparison baseline vs SPH-method for Sample46

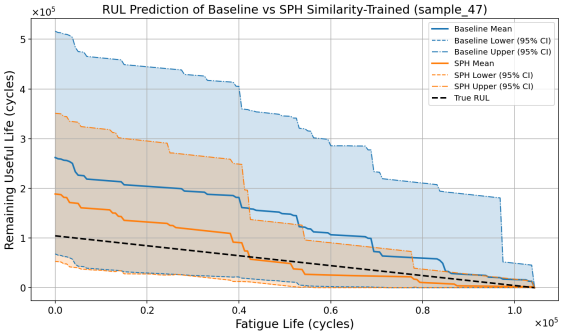


Figure D.11.: RUL comparison baseline vs SPH-method for Sample47

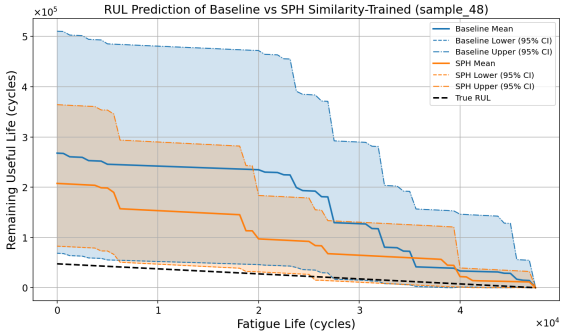


Figure D.12.: RUL comparison baseline vs SPH-method for Sample48

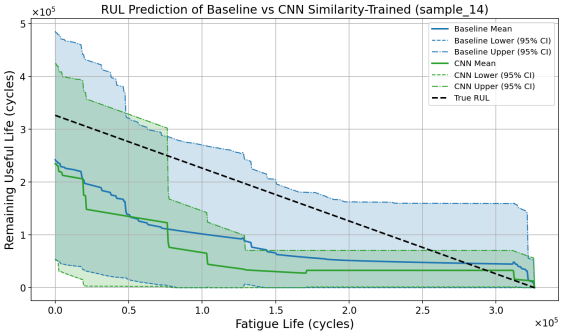


Figure D.13.: RUL comparison baseline vs CNN-method for Sample14

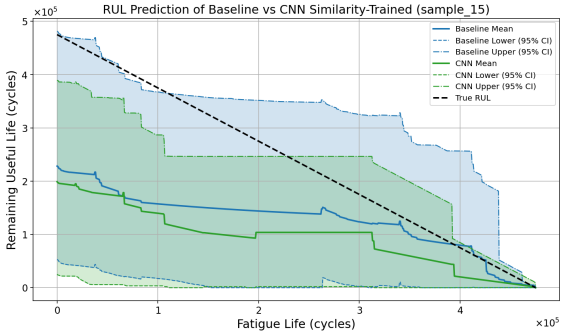


Figure D.14.: RUL comparison baseline vs CNN-method for Sample15

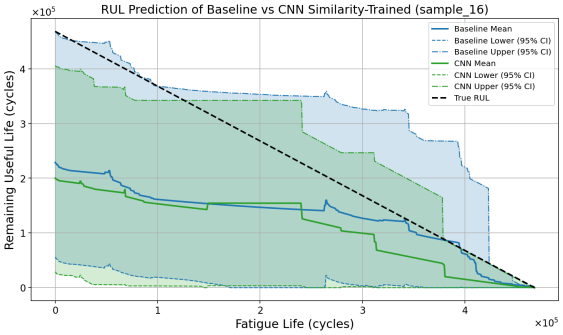


Figure D.15.: RUL comparison baseline vs CNN-method for Sample16

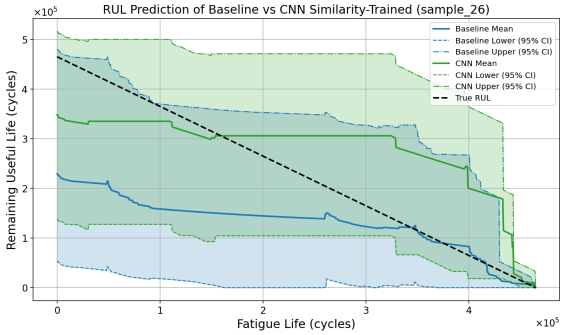


Figure D.16.: RUL comparison baseline vs CNN-method for Sample26

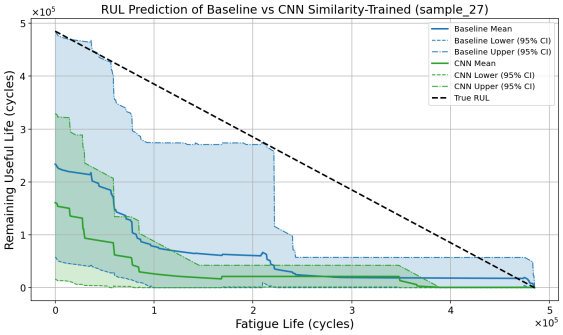


Figure D.17.: RUL comparison baseline vs CNN-method for Sample27

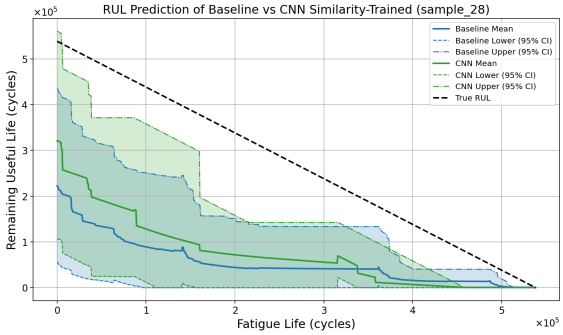


Figure D.18.: RUL comparison baseline vs CNN-method for Sample28

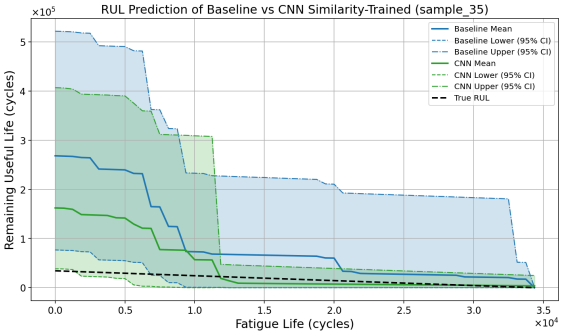


Figure D.19.: RUL comparison baseline vs CNN-method for Sample35

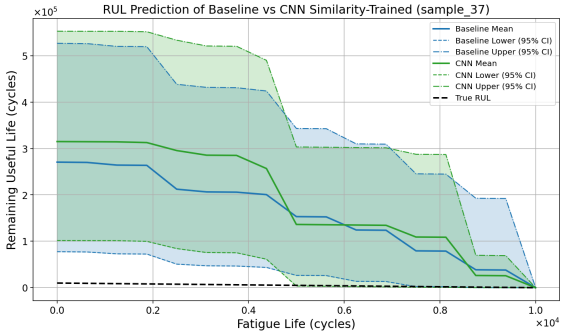


Figure D.20.: RUL comparison baseline vs CNN-method for Sample37

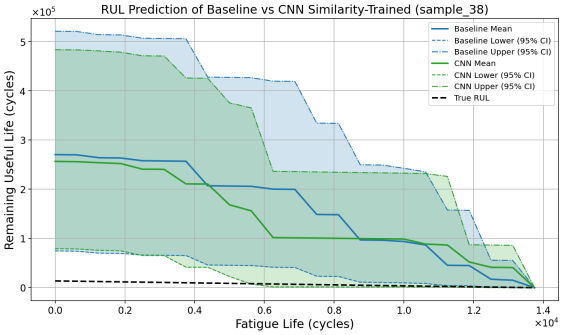


Figure D.21.: RUL comparison baseline vs CNN-method for Sample38

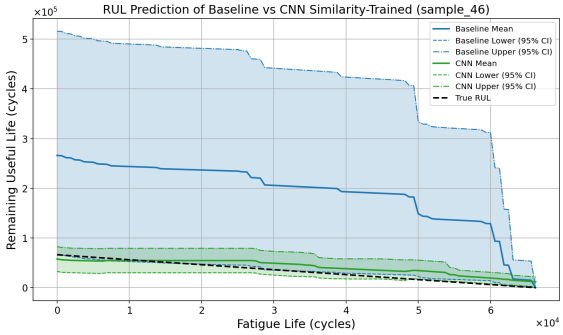


Figure D.22.: RUL comparison baseline vs CNN-method for Sample46

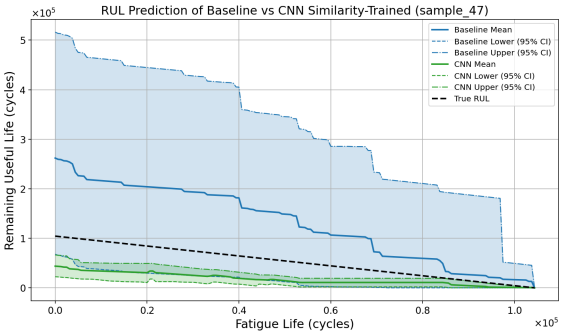


Figure D.23.: RUL comparison baseline vs CNN-method for Sample47

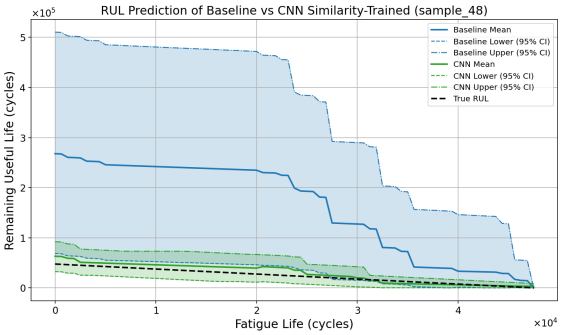


Figure D.24.: RUL comparison baseline vs CNN-method for Sample48

Bibliography

[1] J. Aizpurua, V. Catterson, Y. Papadopoulos, F. Chiacchio, and D. D'Urso, "Supporting group maintenance through prognostics-enhanced dynamic dependability prediction," *Reliability Engineering and System Safety*, vol. 168, pp. 171–188, 2017. Maintenance Modelling.

[2] "Process Material Reference — explorecomposites.com." https://explorecomposites.com/materials-library/process-material-reference/. [Accessed 29-02-2025].

[3] M. A. Matties, "Vector embeddings with subvector permutation invariance using a triplet enhanced autoencoder," *CoRR*, vol. abs/2011.09550, 2020.

[4] M. Norkhairunnisa, T. Chai Hua, S. M. Sapuan, and R. A. Ilyas, *Evolution of Aerospace Composite Materials*, pp. 367–385. Cham: Springer International Publishing, 2022.

[5] R. S. Pierce and B. G. Falzon, "Simulating resin infusion through textile reinforcement materials for the manufacture of complex composite structures," *Engineering*, vol. 3, no. 5, pp. 596–607, 2017.

[6] N. R. Council, *Going to Extremes: Meeting the Emerging Demand for Durable Polymer Matrix Composites*. Washington, DC: The National Academies Press, 2005.

[7] A. Brasington, C. Sacco, J. Halbritter, R. Wehbe, and R. Harik, "Automated fiber placement: A review of history, current technologies, and future paths forward," *Composites Part C: Open Access*, vol. 6, p. 100182, 2021.

[8] B. Parveez, M. I. Kittur, I. Badruddin, S. Kamangar, M. Hussien, and M. A. Umarfarooq, "Scientific advancements in composite materials for aircraft applications: A review," *Polymers*, vol. 14, p. 5007, 11 2022.

[9] M. Inês Silva, E. Malitckii, T. G. Santos, and P. Vilaça, "Review of conventional and advanced non-destructive testing techniques for detection and characterization of small-scale defects," *Progress in Materials Science*, vol. 138, p. 101155, 2023.

[10] R. Li, W. J. Verhagen, and R. Curran, "Toward a methodology of requirements definition for prognostics and health management system to support aircraft predictive maintenance," *Aerospace Science and Technology*, vol. 102, p. 105877, 2020.

[11] M. Salinas-Camus and N. Eleftheroglou, "Uncertainty in aircraft turbofan engine prognostics on the c-mapss dataset," *PHM Society European Conference*, vol. 8, p. 10, jun 2024.

[12] K.-L. Tsui, N. Chen, Q. Zhou, Y. Hai, and W. Wang, "Prognostics and health management: A review on data driven approaches," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–17, 05 2015.

[13] L. Zhao, Y. Li, J. Zhang, L. Zhou, and N. Hu, "A novel material degradation model for unidirectional cfrp composites," *Composites Part B: Engineering*, vol. 135, pp. 84–94, 2018.

[14] T. Loutas, N. Eleftheroglou, and D. Zarouchas, "A data-driven probabilistic framework towards the in-situ prognostics of fatigue life of composites based on acoustic emission data," *Composite Structures*, vol. 161, pp. 522–529, 2017.

[15] S. Sankararaman and K. Goebel, "Uncertainty in prognostics and systems health management," *International Journal of Prognostics and Health Management*, vol. 6, 07 2015.

[16] W. Li, L.-C. Zhang, C.-H. Wu, Y. Wang, Z.-X. Cui, and C. Niu, "A data-driven approach to rul prediction of tools," *Advances in Manufacturing*, vol. 12, no. 1, pp. 6–18, 2024.

[17] S. Sankararaman, "Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction," *Mechanical Systems and Signal Processing*, vol. 52-53, pp. 228–247, 2015.

[18] N. Eleftheroglou, *Adaptive prognostics for remaining useful life of composite structures*. PhD thesis, Delft University of Technology, 2020.

[19] B. Wang, S. Zhong, T.-L. Lee, K. S. Fancey, and J. Mi, "Non-destructive testing and evaluation of composite materials/structures: A state-of-the-art review," *Advances in Mechanical Engineering*, vol. 12, p. 168781402091376, Apr. 2020.

[20] "NDT Ultrasonic Testing — NDT Technology — dolphitech.com." https://dolphitech.com/. [Accessed 19-11-2024].

[21] J. O. Endrerud, "Ndt for non-experts," *Reinforced Plastics*, vol. 58, no. 4, pp. 44–46, 2014.

[22] "Aircraft Inspection — NDT Services for Aviation — dolphitech.com." https://dolphitech.com/industries/aerospace/. [Accessed 19-11-2024].

[23] F. Calabrese, A. Regattieri, M. Bortolini, M. Gamberi, and F. Pilati, "Predictive maintenance: A novel framework for a data-driven, semi-supervised, and partially online prognostic health management application in industries," *Applied Sciences*, vol. 11, no. 8, 2021.

[24] K. Goebel, M. Daigle, A. Saxena, S. Sankararaman, I. Roychoudhury, and J. Celaya, *Prognostics: The Science of Making Predictions*. CreateSpace Independent Publishing Platform, 2017.

[25] N. Eleftheroglou, D. Zarouchas, and R. Benedictus, "An adaptive probabilistic data-driven methodology for prognosis of the fatigue life of composite structures," *Composite Structures*, vol. 245, p. 112386, 2020.

[26] W. J. C. Verhagen, B. F. Santos, F. Freeman, P. van Kessel, D. Zarouchas, T. Loutas, R. C. K. Yeun, and I. Heiets, "Condition-based maintenance in aviation: Challenges and opportunities," *Aerospace*, vol. 10, no. 9, 2023.

[27] M. Schwabacher and K. Goebel, "A survey of artificial intelligence for prognostics," *AAAI Fall Symposium - Technical Report*, 01 2007.

[28] K. Bajarunas, "Generic hybrid models for prognostics of complex systems," *Annual Conference of the PHM Society*, vol. 15, Oct. 2023.

[29] O. Ditlevsen, "Aleatory or epistemic? does it matter?," *Structural Safety*, vol. 31, pp. 105–112, 03 2009.

[30] L. Tang, G. J. Kacprzynski, K. Goebel, and G. Vachtsevanos, "Methodologies for uncertainty management in prognostics," in *2009 IEEE Aerospace conference*, pp. 1–12, 2009.

[31] S. Khalid, J. Song, M. Azad, M. Elahi, J. Lee, S.-H. Jo, and H. S. Kim, "A comprehensive review of emerging trends in aircraft structural prognostics and health management," *Mathematics*, vol. 11, p. 3837, 09 2023.

[32] S. Fu and N. Avdelidis, "Prognostic and health management of critical aircraft systems and components: An overview," *Sensors*, vol. 23, p. 8124, 09 2023.

[33] J. J. Montero Jimenez, S. Schwartz, R. Vingerhoeds, B. Grabot, and M. Salaün, "Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics," *Journal of Manufacturing Systems*, vol. 56, pp. 539–557, 2020.

[34] F.-Y. Xie, Y.-M. Hu, B. Wu, and Y. Wang, "A generalized hidden markov model and its applications in recognition of cutting states," *Int. J. Precis. Eng. Manuf.*, vol. 17, pp. 1471–1482, nov 2016.

[35] N. Eleftheroglou, G. Galanopoulos, and T. Loutas, "Similarity learning hidden semi-markov model for adaptive prognostics of composite structures," *Reliability Engineering and System Safety*, vol. 243, p. 109808, 2024.

[36] V. Nemani, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran, Y. Wang, X. Zhang, and C. Hu, "Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial," *Mechanical Systems and Signal Processing*, vol. 205, p. 110796, 2023.

[37] T. Xiahou, F. Wang, Y. Liu, and Q. Zhang, "Bayesian dual-input-channel lstm-based prognostics: Toward uncertainty quantification under varying future operations," *IEEE Transactions on Reliability*, vol. 73, no. 1, pp. 328–343, 2024.

[38] L. L. Folgoc, V. Baltatzis, S. Desai, A. Devaraj, S. Ellis, O. E. M. Manzanera, A. Nair, H. Qiu, J. Schnabel, and B. Glocker, "Is mc dropout bayesian?," 2021.

[39] Z. Tian, L. Wong, and N. Safaei, "A neural network approach for remaining useful life prediction utilizing both failure and suspension histories," *Mechanical Systems and Signal Processing*, vol. 24, no. 5, pp. 1542–1555, 2010. Special Issue: Operational Modal Analysis.

[40] T. Benkedjouh, K. Medjaher, N. Zerhouni, and S. Rechak, "Remaining useful life estimation based on nonlinear feature reduction and support vector regression," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1751–1760, 2013.

[41] L. Ren, L. Zhao, S. Hong, S. Zhao, H. Wang, and L. Zhang, "Remaining useful life prediction for lithium-ion battery: A deep learning approach," *IEEE Access*, vol. 6, pp. 50587–50598, 2018.

[42] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Reliability Engineering and System Safety*, vol. 172, pp. 1–11, 2018.

[43] S. B. S. Rosenlund and F. Vrolijk, "Utilization of supervised learning for automatic detection of damages in materials from non-destructive testing based ultrasound data." https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3123941?show=full&locale-attribute=en, May 2020. Accessed 19-11-2024.

[44] "Impact Damage To An Aircraft - Dolphitech — dolphitech.com." https://dolphitech.com/case-study/impact-damage-to-an-aircraft/. [Accessed 21-11-2024].

[45] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[46] M. A.-A. Bhuiyan and A. R. Khan, "Image quality assessment employing rms contrast and histogram similarity.," *Int. Arab J. Inf. Technol.*, vol. 15, no. 6, pp. 983–989, 2018.

[47] J. Nilsson and T. Akenine-Möller, "Understanding ssim," 2020.

[48] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (ssim) revisited: A data-driven approach," *Expert Systems with Applications*, vol. 189, p. 116087, 2022.

[49] A. A. Elngar, M. Arafa, A. Fathy, B. Moustafa, O. Mahmoud, M. Shaban, and N. Fawzy, "Image classification based on cnn: a survey," *Journal of Cybersecurity and Information Management*, vol. 6, no. 1, pp. 18–50, 2021.

[50] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, vol. 9, no. 2, pp. 187–212, 2022.

[51] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, Aug. 2018.

[52] M. G. Forero, C. Arias-Rubio, and B. T. González, "Analytical comparison of histogram distance measures," in *Iberoamerican Congress on Pattern Recognition*, pp. 81–90, Springer, 2018.

[53] M. Ionescu, A. D. Glodeanu, I. R. Marinescu, A. G. Ionescu, and C. C. Vere, "Similarity analysis for medical images using color and texture histograms," *Current Health Sciences Journal*, p. 196–202, July 2022.

[54] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[55] N. Eleftheroglou, "Adaptive prognostics: A reliable rul approach," in *Annual Conference of the PHM Society*, vol. 15, 2023.

[56] T. Kontogiannis, M. Salinas-Camus, and N. Eleftheroglou, *Stochastic Modeling and Statistical Methods*. Elsevier, 2025.

[57] A. Baldi, "Robust algorithms for digital image correlation in the presence of displacement discontinuities," *Optics and Lasers in Engineering*, vol. 133, p. 106113, 2020.

[58] N. Eleftheroglou, D. Zarouchas, T. Loutas, R. Alderliesten, and R. Benedictus, "Structural health monitoring data fusion for in-situ life prognosis of composite structures," *Reliability Engineering and System Safety*, vol. 178, pp. 40–54, 2018.

[59] H. Zhang, W. Gao, X. Chen, and D. Zhao, "Object detection using spatial histogram features," *Image and Vision Computing*, vol. 24, no. 4, pp. 327–341, 2006.

[60] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2169–2178, 2006.

[61] S. M. Lee, J. H. Xin, and S. Westland, "Evaluation of image similarity by histogram intersection," *Color Research and Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, vol. 30, no. 4, pp. 265–274, 2005.

[62] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 815–823, IEEE, June 2015.

[63] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Sampling matters in deep embedding learning," 2018.

[64] X. Yuan, Q. Liu, J. Long, L. Hu, and Y. Wang, "Deep image similarity measurement based on the improved triplet network with spatial pyramid pooling," *Information*, vol. 10, no. 4, p. 129, 2019.

[65] L. Fu, H. Yu, X. Li, C. P. Przybyla, and S. Wang, "Deep learning for object detection in materials-science images: A tutorial," *IEEE Signal Processing Magazine*, vol. 39, no. 1, pp. 78–88, 2022.

[66] M. P. S. Amirshahi and S. Yu, "NTNU Open: Image Quality Assessment by Comparing CNN Features between Images — ntnuopen.ntnu.no." https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2449966. [Accessed 21-11-2024].

[67] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546 vol. 1, 2005.

[68] F. Warburg, M. Jorgensen, J. Civera, and S. Hauberg, "Bayesian triplet loss: Uncertainty quantification in image retrieval," 10 2021.

[69] R. Wang, X. Zhou, Y. Liu, D. Liu, Y. Lu, and M. Su, "Identification of the surface cracks of concrete based on resnet-18 depth residual network," *Applied Sciences*, vol. 14, no. 8, 2024.

[70] M. Prunella, R. M. Scardigno, D. Buongiorno, A. Brunetti, N. Longo, R. Carli, M. Dotoli, and V. Bevilacqua, "Deep learning for automatic vision-based recognition of industrial surface defects: A survey," *IEEE Access*, vol. 11, pp. 43370–43423, 2023.

[71] M. Odusami, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Analysis of features of alzheimer's disease: Detection of early stage from functional brain changes in magnetic resonance images using a finetuned resnet18 network," *Diagnostics*, vol. 11, no. 6, 2021.

[72] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: a large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 06 2009.

[75] S.-Z. Yu, "Hidden semi-markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010. Special Review Issue.

[76] N. Eleftheroglou and T. Loutas, "Fatigue damage diagnostics and prognostics of composites utilizing structural health monitoring data and stochastic processes," *Structural Health Monitoring*, vol. 15, pp. 473–488, 05 2016.

[77] A. Saxena, J. Celaya, E. Balaban, K. Goebel, B. Saha, S. Saha, and M. Schwabacher, "Metrics for evaluating performance of prognostic techniques," in *2008 International Conference on Prognostics and Health Management*, pp. 1–17, 2008.

[78] W. Yu, I. Y. Kim, and C. Mechefske, "An improved similarity-based prognostic algorithm for rul estimation using an rnn autoencoder scheme," *Reliability Engineering and System Safety*, vol. 199, p. 106926, 2020.

[79] J. Wu, L. Kong, Z. Cheng, Y. Yang, and H. Zuo, "Rul prediction for lithium batteries using a novel ensemble learning method," *Energy Reports*, vol. 8, pp. 313–326, 2022. 2022 International Conference on the Energy Internet and Energy Interactive Technology.

[80] I. De Pater and M. Mitici, "Novel metrics to evaluate probabilistic remaining useful life prognostics with applications to turbofan engines," *PHM Society European Conference*, vol. 7, p. 96–109, June 2022.