Graph neural networks for SHM

exploiting spatial interdependencies of strain data for diagnostics and prognostics

Stamatelatos, Giannis; Galanopoulos, Georgios; Zarouchas, Dimitrios; Loutas, Theodoros

DOI
10.1177/14759217251386802

Publication date
2025

Document Version
Final published version

Published in
Structural Health Monitoring

Citation (APA)
Stamatelatos, G., Galanopoulos, G., Zarouchas, D., & Loutas, T. (2025). Graph neural networks for SHM: exploiting spatial interdependencies of strain data for diagnostics and prognostics. *Structural Health Monitoring*, Article 14759217251386802. https://doi.org/10.1177/14759217251386802

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Original Article

SHM Structural Health Monitoring

Structural Health Monitoring
1–24

DOI: 10.1177/14759217251386802
journals.sagepub.com/home/shm
Sage

Graph neural networks for SHM: exploiting spatial interdependencies of strain data for diagnostics and prognostics

Giannis Stamatelatos[1], Georgios Galanopoulos[2], Dimitrios Zarouchas[2,3] and Theodoros Loutas[1]

Abstract

Structural health monitoring using strain data faces a critical challenge: decoupling subtle structural degradation signatures from the dominant influence of operational loads. This paper introduces a novel methodology to address this by synergistically combining a custom health indicator (HI) with graph neural networks (GNNs). The proposed HI, derived from the cumulative absolute first derivative of strain over time, effectively isolates load-independent features indicative of damage progression. These features serve as input to our proposed GENConv with Edge Attributes (GENEA) model, a GNN that explicitly models the spatially distributed sensors as an interconnected network, leveraging spatial interdependencies and edge attribute information within the strain field to enhance damage assessment. This integrated approach enables accurate structural stiffness reduction estimation (diagnostics) and remaining useful life (RUL) prediction (prognostics). Applied to strain data from fatigue tests on representative aeronautical composite panels, the methodology is rigorously evaluated using Leave-One-Panel-Out cross-validation. The framework shows promising performance on unseen test data, although challenges in generalizing to out-of-distribution specimens were also identified, highlighting the importance of a diverse training set for real-world applicability. Experimental results confirm the framework's superiority. The proposed GENEA model significantly outperforms both a fundamental multi-layer perceptron and a spatially aware convolutional neural network baseline, and successfully generalizes to an unseen panel with a different sensor count. This validates the benefits of using a tailored GNN framework to learn robust, geometrically invariant patterns from load-decoupled spatial strain data.

Keywords

Graph neural networks (GNNs), composite structures, strain-based SHM, remaining useful life (RUL), uncertainty quantification (UQ), health indicator (HI), machine learning (ML)

Introduction

Composite materials, particularly fiber-reinforced polymers (FRPs) such as carbon fiber-reinforced polymer (CFRP), have become widely adopted in numerous high-performance engineering applications. Their adoption across safety-critical sectors, including aerospace, wind energy, automotive, and civil infrastructure, is driven by exceptional mechanical properties, notably high strength-to-weight and stiffness-to-weight ratios. These advantages, however, are accompanied by significant challenges related to structural integrity assessment. Composites exhibit complex, anisotropic behavior and are susceptible to unique damage mechanisms (e.g., delamination, matrix cracking, fiber breakage, and debonding) that are often difficult to detect using conventional methods. A critical characteristic is the potential for internal damage

[1]Applied Mechanics Laboratory (AML), Department of Mechanical Engineering and Aeronautics, University of Patras, Patras, Greece
[2]Structural Integrity and Composites Group, Faculty of Aerospace Engineering, Delft University of Technology, Netherlands
[3]Center of Excellence in Artificial Intelligence for Structures, Faculty of Aerospace Engineering, Delft University of Technology, Delft, Netherlands

Corresponding author:
Giannis Stamatelatos and Theodoros Loutas, Applied Mechanics Laboratory (AML), Department of Mechanical Engineering and Aeronautics, University of Patras, University Campus, Rio, Patras 26504, Greece.
Emails: johnstamly@gmail.com; thloutas@upatras.gr

accumulation, significantly compromising structural integrity long before external signs become apparent.[1] Consequently, ensuring the safety and reliability of safety-critical composite structures necessitates monitoring techniques capable of continuous assessment. Structural health monitoring (SHM) has emerged as a promising technology addressing this need. By integrating sensing systems directly with the structure, SHM aims to enhance operational safety, enable condition-based maintenance (CBM), reduce lifecycle costs, and extend operational lifespans, rendering it essential for effective risk mitigation and asset management.[2]

Modern SHM systems fundamentally rely on networks of sensors to capture data indicative of the structure's health state. Among various quantities, strain is particularly informative, as it directly reflects the internal stress distribution and its redistribution due to damage-induced changes in local stiffness.[3] Monitoring strain patterns thus provides a sensitive means of assessing structural health. Fiber Bragg Grating (FBG) sensors are well-suited for SHM due to their small size, light weight, immunity to electromagnetic interference, and multiplexing capabilities.[4] A network of distributed strain sensors generates spatially correlated data, possessing an intrinsic graph-like structure where sensors are nodes and physical relationships are edges. While rich in information, interpreting the complex relationship between damage (e.g., fatigue-induced stiffness loss) and multi-sensor strain patterns remains challenging, necessitating analysis techniques capable of leveraging the network topology.[5]

Building on the acquisition of spatially distributed strain data, various methodologies have been proposed in the literature to leverage these measurements for SHM diagnostics and prognostics in composite structures.[4,6] Common approaches involve processing raw strain signals to extract damage-sensitive features such as direct characteristics,[7] frequency content,[8] wavelet coefficients,[9] or formulating custom health indicators (HIs) designed to track degradation, such as those based on relative strain changes[10] or sensor array contributions.[11] However, a critical and well-documented limitation of many existing strain-based HI formulations is their inherent sensitivity to operational load. For instance, studies show that HIs comparing current strain to a baseline are only reliable under constant load levels, while other common indicators can become highly load-dependent in different structural regimes, such as post-buckling. This sensitivity poses a significant challenge for creating robust prognostic systems for real-world applications involving variable or spectrum loading.[7,10] These features are then utilized for damage classification or state assessment, employing tools such as machine learning models, statistical methods, or regression techniques.[12] Regarding prognostics, efforts to predict remaining useful life (RUL) typically involve either extrapolating HI trends using time-series or similarity-based models,[13,14] or utilizing probabilistic frameworks, such as Kalman Filter variants[15] and Hidden Markov Models,[16,17] to handle stochasticity and provide predictions with uncertainty quantification. However, these methods often analyze sensor data independently, failing to effectively leverage the valuable information contained within the complex spatial interdependencies between sensors across the network.

The challenges in interpreting complex SHM data, especially under variable loading with evolving damage, have spurred the adoption of data-driven methodologies.[18] Machine learning (ML) and deep learning (DL) methods learn input–output relationships directly from sensor measurements, implicitly capturing the structure's integrated response.[19] However, traditional ML models like multi-layer perceptrons (MLPs), while versatile function approximators, process input features independently and struggle to explicitly model the spatial relationships within a sensor network.[20] Even advanced architectures like convolutional neural networks (CNNs)[21] and recurrent neural networks (RNNs),[22] powerful for grid-like data or time series, respectively, are not inherently designed to process the relational information defined by the sensor network's physical structure. Consequently, they struggle to extract robust patterns reflecting the inter-sensor spatial dependencies essential for understanding global structural behavior, such as strain field redistribution due to damage.[5] Addressing this limitation, the field of geometric deep learning provides a paradigm for analyzing such data with models designed to operate on non-Euclidean structures. This field encompasses several powerful architectures, including spectral graph models[23] and the more recent graph transformers.[24] Among these, graph neural networks (GNNs) that operate via message-passing mechanisms have become particularly prominent for their flexibility and effectiveness.[25,26] This approach allows nodes (sensors) to iteratively aggregate information from their neighbors, enabling the learning of context-aware representations that explicitly capture the relational information and spatial dependencies inherent in the sensor network.[27] This makes GNNs especially suitable for SHM applications, as they can learn the underlying relational patterns within the strain field that are governed by the structure's physical response to damage.

Despite the merits of GNNs, their application to strain-based SHM faces specific challenges and knowledge gaps. While GNNs are increasingly adopted in SHM,[28] their use has often focused on scenarios distinct from the approach explored here. For instance,

some studies employ GNNs for fusing data from multi-modal sensor networks (integrating accelerometers, temperature, strain, etc.), often in non-composite applications,[29,30] while others model temporal dependencies by constructing graphs from single-sensor time-series segments.[31] Considerably less attention has been paid to leveraging GNNs specifically for extracting complex spatial patterns directly from a network of spatially distributed, homogeneous sensors, like the FBG strain network used here, to assess damage progression (e.g., fatigue-induced stiffness loss or RUL) in composite structures by learning inter-sensor relationships. This gap is critical, and the primary reason it remains less explored is the fundamental challenge of working with strain-only data. Without information from other sensor modalities, the strain signals are easily dominated and obscured by operational load variations, which can mask the subtle spatial patterns of damage that a GNN is intended to learn. Therefore, developing input features sensitive to degradation yet robust to load effects is a crucial prerequisite for this application.[10] Furthermore, demonstrating the specific advantage of GNNs over simpler models (like MLPs and grid-based CNNs) in leveraging spatial patterns for predicting global health indicators requires careful experimental validation. A key challenge for any data-driven SHM framework is generalizing to outlier specimens or unexpected damage scenarios that fall outside the training distribution. As noted in the literature, the predictions of standard models are often erroneous in such cases because their performance relies on the test data closely resembling the training data, and creating a comprehensive training set covering all real-world possibilities is impractical.[14] This necessitates the development of robust models capable of learning more fundamental degradation patterns. Consequently, the central challenge addressed in this paper is the development and validation of an effective SHM framework for diagnostics and prognostics in composite structures using multi-sensor strain data. This necessitates synergistically tackling two key difficulties: (1) extracting reliable, damage-sensitive features effectively decoupled from operational load variations inherent in strain signals, and (2) employing methods that explicitly leverage the crucial spatial interdependencies within the homogeneous sensor network to interpret these features. Successfully addressing both aspects motivates the integrated approach presented herein.

We propose and evaluate a novel framework that integrates GNNs with a custom HI derived from strain measurements. The core idea is to leverage the HI to decouple load effects from degradation signatures, while utilizing the GNN to explicitly model and exploit spatial correlations within the sensor network data for improved predictive accuracy. Our key contributions, detailed in the subsequent sections, begin with the development and application of a custom HI based on the cumulative absolute first derivative of strain signals (as detailed in Custom HI construction), specifically designed to effectively decouple load effects from fatigue degradation signatures. This HI is then used to construct the inputs for our proposed modeling framework. The core of our approach is a novel input design for a GENConv-based GNN architecture (termed GENEA), where the HI defines not only the node features but, crucially, also the edge attributes as pre-calculated pairwise differentials. This model is implemented using a point-wise regression approach to explicitly exploit spatial correlations within the sensor network data (represented as a fully connected graph, as described in the dataset construction for GNN) for predicting structural health states. The framework's performance is comprehensively validated for both stiffness estimation (diagnostics) and multi-threshold RUL prediction (prognostics) using experimental fatigue test data from hybrid composite-metal panels instrumented with FBG sensors (see "Results and discussion"), with rigorous assessment via Leave-One-Panel-Out Cross-Validation (LOPO CV). Furthermore, the framework incorporates uncertainty quantification using Monte Carlo (MC) dropout to provide essential confidence bounds for the predictions. Finally, a Framework for Agile, Integrated, and Reproducible (FAIR) model comparison,[32] including systematic evaluations against both a fundamental MLP and a spatially aware CNN baseline, as well as an ablation study of various GNN architectures (see "FAIR model comparison"), quantitatively demonstrates the significant performance advantages conferred by the GNN-based approach and underscores the critical contribution of the proposed HI.

The paper is organized as follows: the section "Experimental setup" describes the experimental setup and the dataset. The section "Proposed methodology" details the proposed methodology, from HI calculation to the GNN architecture and training. The section "Results and discussion" presents the results, including the main stiffness and RUL predictions, the FAIR model comparison, and the analysis of the GNN's generalization to varying geometries. Finally, the section "Conclusions and recommendations" summarizes the findings and discusses future work. Additional insights are offered in Appendices A through E of the Supplemental Material.

Experimental setup

The analysis, GNN modeling framework, and all findings presented in this manuscript are entirely novel.

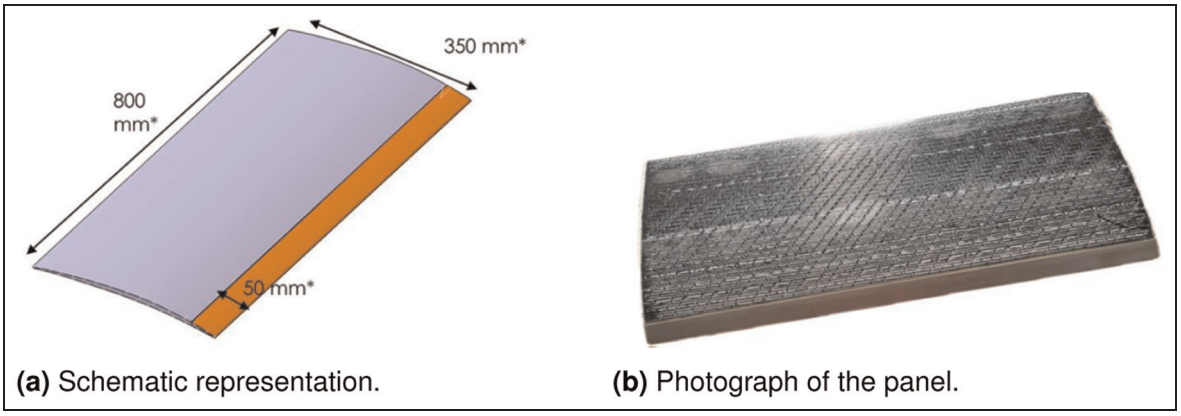

Figure 1. A hybrid composite-metal panel was used in the experimental campaign. (a) Schematic representation and (b) photograph of the panel.

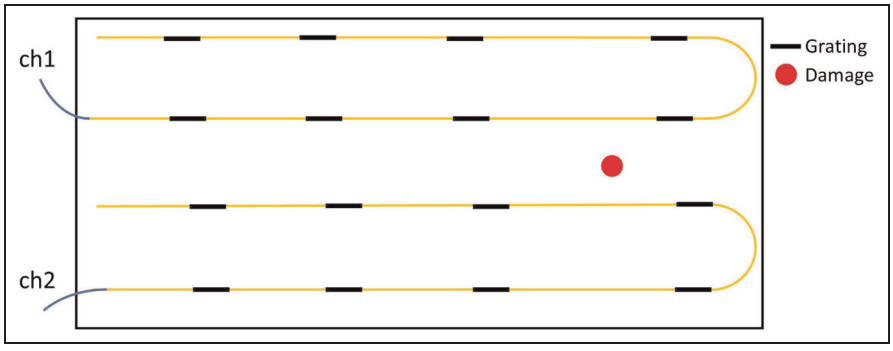

Figure 2. Schematic indicating the placement of FBG sensor arrays (tapes) on the FOD panel surface.
FBG: fiber Bragg grating; FOD: foreign object damage.

The strain data used in this study were acquired during an experimental campaign designed to induce accelerated degradation on hybrid composite-metal panels, representative of aircraft engine components; the details of this campaign are published in Galanopoulos et al.[33] For full transparency and reproducibility, the raw dataset generated during this campaign has been made publicly available in the Zenodo open research repository.[34] The scope of those prior publications was limited to describing the experimental procedure and the dataset, respectively, while the current work focuses on the novel data analysis and prognostic modeling.

Materials and geometry

The specimens, referred to as "FOD panels" (simulating Foreign Object Damage scenarios), consisted of a 3D woven carbon fiber-reinforced polymer (CFRP) composite section bonded to a steel leading edge manufactured by Safran Composites. The specific geometry, layup, and test details can be found in detail in Galanopoulos et al.[33] Figure 1 shows schematic and photographic representations.

Sensor network

Strain was monitored using FBG sensors integrated into optical fibers. The fibers were surface-bonded to the panel's bottom (tensile) side using a room-temperature curing adhesive. Sensors were strategically placed between the loading pins, concentrating on areas of expected high deformation and potential damage initiation/propagation, including near the trailing edge, at the center, and near the steel leading edge (Figure 2). Data were acquired at a sampling rate of 1 Hz under ambient conditions.

Loads and damage introduction

Panels were subjected to four-point bending fatigue loading using an MTS hydraulic test machine. The loading protocol involved blocks of cycles at progressively increasing load amplitudes (ranging from 4 kN to 28 kN peak load), designed to accelerate fatigue damage accumulation. To create a diverse set of test cases, the initial damage state of the panels was varied. Panels FOD4, FOD6, and FOD7 were subjected to

Table 1. Summary of the experimental dataset characteristics for each panel.

Panel ID	Impacted	Load steps (kN)	Total cycles	FBG number
FOD3	Yes	4-24	~3300	6
FOD4	Yes	4-22	~3080	16
FOD5	No	4-24	~3400	16
FOD6	Yes	4-22	~3120	16
FOD7	Yes	4-24	~3250	16

FBG: fiber Bragg grating; FOD: foreign object damage.

controlled impacts at different locations and timesteps, while panel FOD5 was left pristine (see Table 1). As detailed in the primary experimental work,[33] the subsequent damage evolution during fatigue was highly stochastic. Damage typically initiated as matrix cracking and delamination, progressing in unique patterns on each specimen to include fiber-skin separation, debonding between the steel and CFRP components, and ultimately massive fiber failures of the 3D woven CFRP material that caused a significant stiffness loss (>50%).

Dataset overview

The dataset comprises synchronized strain measurements, applied load levels, cycle counts, and global stiffness values (as estimated by the load/displacement test data) for five FOD panels (FOD3–FOD7). As summarized in Table 1, this collection of specimens provides a rich and challenging testbed for evaluating model generalization due to its significant diversity. A key variation is the sensor count: panels FOD4–FOD7 were equipped with 16 FBG sensors each, while panel FOD3 had only 6. The initial damage states also varied; panel FOD5 was left pristine (non-impacted), whereas FOD4, FOD6, and FOD7 were subjected to impacts at different locations. Furthermore, exploratory analysis revealed other panel-specific characteristics, such as intermittent sensor noise on FOD4 and a high degree of similarity in the data distributions of FOD6 and FOD7. A key advantage of the GNN framework, central to this study, is its ability to process graphs of varying sizes, allowing the inclusion of the six-sensor FOD3 panel to create a more robust and realistic test of generalization capabilities.

Proposed methodology

The proposed SHM framework, depicted in Figure 3, integrates data preprocessing, HI construction, GNN-based prediction for diagnostics (stiffness) and prognostics (RUL), by utilizing our proposed model GENEA. Uncertainty quantification is incorporated using MC dropout.

Preprocessing

The raw data acquired from the experiments undergo several preprocessing steps, with strain and stiffness data being treated separately as described below:

Strain data

1. **Initial cleaning:** Removal of any obvious outliers or sensor malfunction periods identified during visual inspection or based on prior knowledge. Handling of missing values (Not a Number [NaNs]) using appropriate methods like interpolation or removal, although minimal NaNs were present in the used dataset segments.
2. **Downsampling:** The raw 1 Hz strain data (Figure 4) contain significant redundancy for tracking fatigue degradation trends. To reduce computational load while preserving essential information, the data are downsampled. First, the time series for each sensor is segmented into non-overlapping windows of 200 s, and the mean strain value within each window is computed. This significantly reduces the number of data points.
3. **Smoothing:** A moving average filter with a window size of 10 (applied to the downsampled data) is used to further smooth the signals and mitigate high-frequency noise, focusing on the lower-frequency trends associated with degradation.

The resulting preprocessed strain data (Figure 5) form the input for the HI calculation.

Stiffness data. The global coupon stiffness was estimated by the load/displacement data from the test machine at regular loading cycle intervals.

1. **Interpolation:** To align the stiffness measurements with the strain data time steps, linear interpolation is applied to the stiffness values.
2. **Normalization:** Stiffness is then converted to a percentage of the initial (maximum recorded) stiffness for each panel: $\text{Stiffness}_{\%} = (S(t)/S_{\max}) \times 100$. This provides a normalized measure of degradation.
3. **Target definition:** The interpolated and normalized stiffness values serve as the target variable for the diagnostic (stiffness reduction estimation) task. For the prognostic (RUL) task, the target is the number of remaining time steps (or cycles) until a specific stiffness reduction threshold is reached.

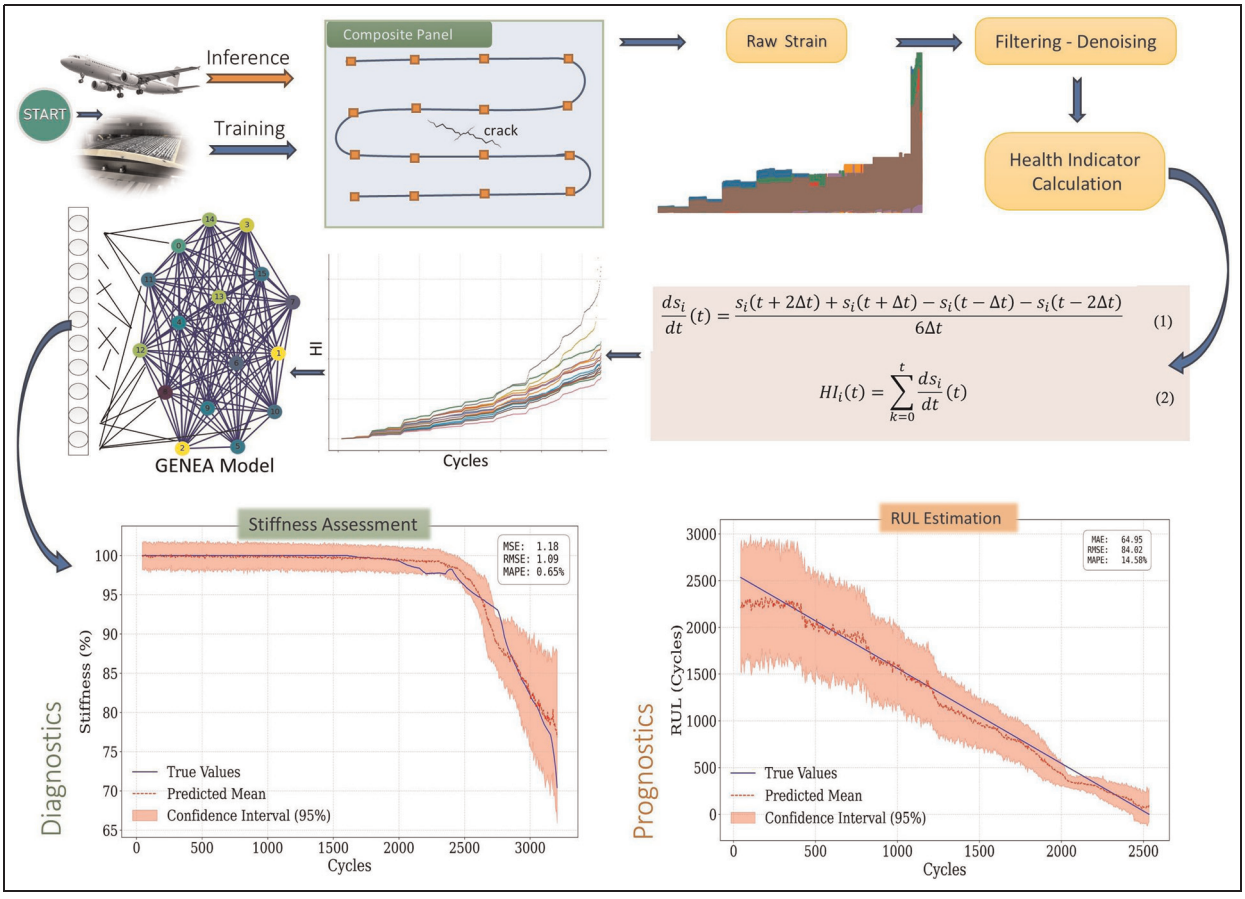

Figure 3. Overview of the proposed GNN-based framework for strain-based diagnostics and prognostics, incorporating preprocessing, custom HI construction, GNN for spatial feature learning, MLP for prediction, and MC dropout for uncertainty quantification.

HI: health indicator; GNN: graph neural network; ML: machine learning; MC: Monte Carlo; MLP: multi-layer perceptron.

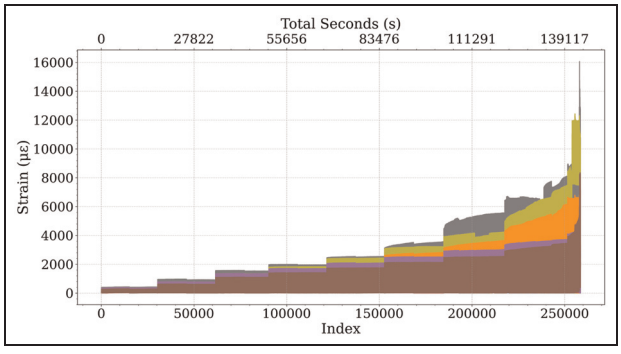

Figure 4. Example of raw strain data (1 Hz) for all 16 sensors of panel FOD7. The *x*-axes show data index and corresponding time in seconds; the *y*-axis shows strain values.

FOD: foreign object damage.

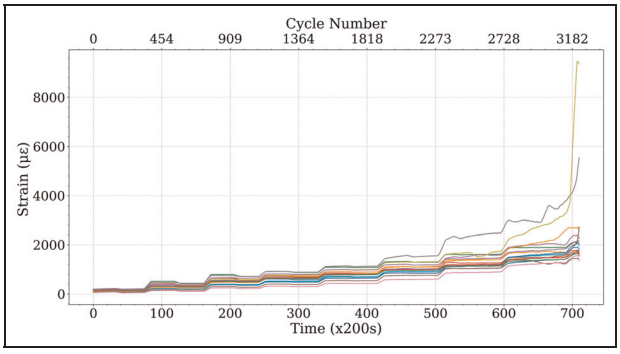

Figure 5. Example of preprocessed strain data for panel FOD7 after downsampling (mean over 200 s windows) and smoothing (moving average window 10). The *x*-axes show the downsampled time index and corresponding fatigue cycles.

FOD: foreign object damage.

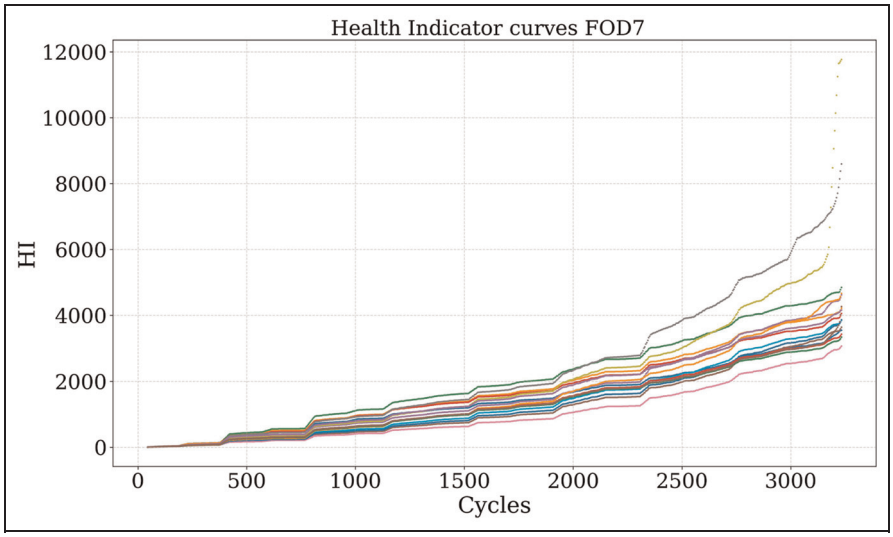

Figure 6. Custom HI curves calculated using Equation (1) for all 16 sensors of panel FOD7. The *x*-axis represents fatigue cycles, and the *y*-axis shows the accumulated HI values.
HI: health indicator; FOD: foreign object damage.

Custom health indicator construction

To decouple the load effects from the degradation effect on strain data, we compute a custom HI for each sensor i based on the cumulative absolute differentiation of its preprocessed strain signal $\epsilon_i(t)$. The calculation is according to Equation (1):

$$\mathrm{HI}_i(t) = \sum_{k=1}^{t} |\Delta\epsilon_i(k)| = \sum_{k=1}^{t} |\epsilon_i(k) - \epsilon_i(k-1)| \qquad (1)$$

where $\mathrm{HI}_i(t)$ is the health indicator for the sensor i at the current discrete time step t, and the summation index k iterates from the first time step up to t.

This formulation is designed to provide a direct physical interpretation of damage progression. In a healthy structure under stable loading, the strain response is repeatable and the rate of change is minimal, causing the HI to accumulate slowly. As physical damage (e.g., matrix cracking, delamination) occurs, the local stiffness is permanently reduced, causing a redistribution of the internal strain field. This results in irreversible shifts in the strain readings at various sensor locations. The absolute first derivative in the HI formulation captures the magnitude of these instantaneous changes, while the cumulative summation ensures that the HI monotonically increases with each successive damage event. Therefore, the HI acts as a proxy for the total history of irreversible changes in the material's load-carrying behavior, effectively tracking progressive degradation rather than just transient signal fluctuations. Furthermore, using the first derivative as the basis of the HI is a deliberate choice to decouple the analysis from the absolute strain magnitude, which varies significantly across the panel. In a healthy structure, the rate of change of strain at different sensor locations should be highly correlated and proportional to the rate of change of the applied load. Damage is a local phenomenon that breaks this synchronous response. A sensor near a developing crack or delamination will exhibit a different strain rate compared to the majority of sensors in unaffected regions. By focusing on the absolute derivative, our HI is specifically designed to be sensitive to these deviations from the expected, synchronous behavior, capturing the emergence of localized, damage-induced changes in the strain field's trend.

This HI quantifies the total amount of strain variation experienced by the sensor over time. The absolute value ensures that both increases and decreases in strain rate contribute positively, reflecting overall mechanical activity and potential micro-damage accumulation. The accumulation captures the progressive nature of degradation. Unlike raw strain, this HI is hypothesized to be less directly dependent on the instantaneous load magnitude but more sensitive to changes in the material's response, which are indicative of damage. The significant improvement in model performance and robustness achieved by employing this HI, compared to directly using processed strain data, will be quantitatively demonstrated through an ablation study presented in "FAIR model comparison." Figure 6 shows the resulting HI curves for an exemplary panel, while Appendix A of the Supplemental Material provides further visual intuition and justification behind the use of the proposed HI by analyzing HI differentials between various sensor pairs.

Table 2. Dimensions of the final HI dataset for each panel before cross-validation splitting.

Panel ID	Shape
FOD3	605 × .6
FOD4	731 × .16
FOD5	762 × .16
FOD6	760 × .16
FOD7	711 × .16

Shape is (number of time steps after preprocessing × .number of sensors).
HI: health indicator; FOD: foreign object damage.

Dataset construction for GNN

The preprocessed data and calculated HIs are structured for input into the GNN model as described below. The final shapes of the input data are shown in Table 2.

- **Input features:** For each time step t, the input to the GNN is a feature matrix $\mathbf{X}_t \in \mathbb{R}^{N \times F}$, where N is the number of sensors (nodes, $N = 16$ for FOD4-7) and F is the number of features per sensor. In this study, we use the HI value as the primary feature, so $F = 1$, and the input is effectively a vector $\mathbf{x}_t \in \mathbb{R}^N$ where $[\mathbf{x}_t]_i = HI_i(t)$.
- **Graph structure:** A graph $G = (V, E)$ is defined where V is the set of $N = 16$ sensors. Given the lack of precise prior knowledge on damage propagation paths and to allow the model maximum flexibility in learning relationships, we assume a "fully connected graph." This means an edge (i, j) exists between every pair of sensors i and j. The graph is therefore represented by a dense adjacency matrix $\mathbf{A}$, excluding self-loops. A node's own information is incorporated into its updated representation through residual connections within our GNN architecture, making explicit self-loops in the graph structure unnecessary.
- **Targets:** For stiffness estimation, the target at time t is the corresponding interpolated stiffness percentage $y_t = \text{Stiffness}_{\%}(t)$. For RUL estimation targeting an EOL threshold S_{EoL}, the target at the time t is $y_t = \text{Cycles}_{\text{EoL}} - \text{Cycles}(t)$, where $\text{Cycles}_{\text{EoL}}$ is the cycle count when stiffness first drops below S_{EoL}.
- **Handling of varying time-series lengths:** It is important to note that the proposed framework employs a point-wise regression approach. The GNN model processes the data for each time step independently to predict for that specific moment. Consequently, the input to the model at any given forward pass is a single graph representing the sensor states at one point in time, not the entire time series. This architectural choice means that the varying number of time steps across different panels (e.g., 731 for FOD4 vs 762 for FOD5) does not affect the training process, and, therefore, no padding or truncation of the time series is necessary.
- **Data splitting (cross-validation):** To rigorously evaluate the model, two cross-validation schemes are employed. (1) For the direct comparison against the MLP and CNN baselines in the section "FAIR model comparison," a fourfold Leave-One-Panel-Out (LOPO) CV is performed on the 16-sensor panels (FOD4–7), as these models require a fixed input size. (2) To specifically test the GNN's ability to generalize across different geometries, a fivefold LOPO CV is conducted on the full dataset (FOD3–7) for the proposed GENEA model, as detailed in section "Generalization to varying geometries: incorporating FOD3."

GNN architecture and training

At the core of the proposed framework is a GNN model that processes spatial sensor information encoded in the HI features. GNNs are a class of deep learning models designed for graph-structured data, operating on a principle known as message passing,[27,35] which is illustrated in Figure 7. In this framework, nodes (sensors) iteratively update their feature representations by aggregating information from their neighbors. A single layer of a message-passing GNN can be described by a general function:

$$\mathbf{h}_i^{(l+1)} = \psi^{(l)}\left(\mathbf{h}_i^{(l)}, \bigoplus_{j \in \mathcal{N}(i)} \phi^{(l)}\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \mathbf{e}_{ji}\right)\right) \quad (2)$$

where $\mathbf{h}_i^{(l)}$ is the feature vector of node i at layer l, $\mathcal{N}(i)$ is the set of its neighbors, and $\mathbf{e}_{ji}$ represents the features of the edge from the node j to i. The learnable functions $\phi^{(l)}$ (message creation) and $\psi^{(l)}$ (update), along with the permutation-invariant aggregation function $\oplus$ (e.g., sum or mean), define the specific behavior of the GNN layer. The initial node representations, $\mathbf{h}_i^{(0)}$, are set to the input features $\mathbf{x}_i$. While many GNN variants exist, our work employs a specific architecture designed to leverage edge information effectively, as detailed below.

Model architecture. The specific GNN architecture developed and utilized in this work, termed GENEA—GENConv with Edge Attributes to highlight its use of

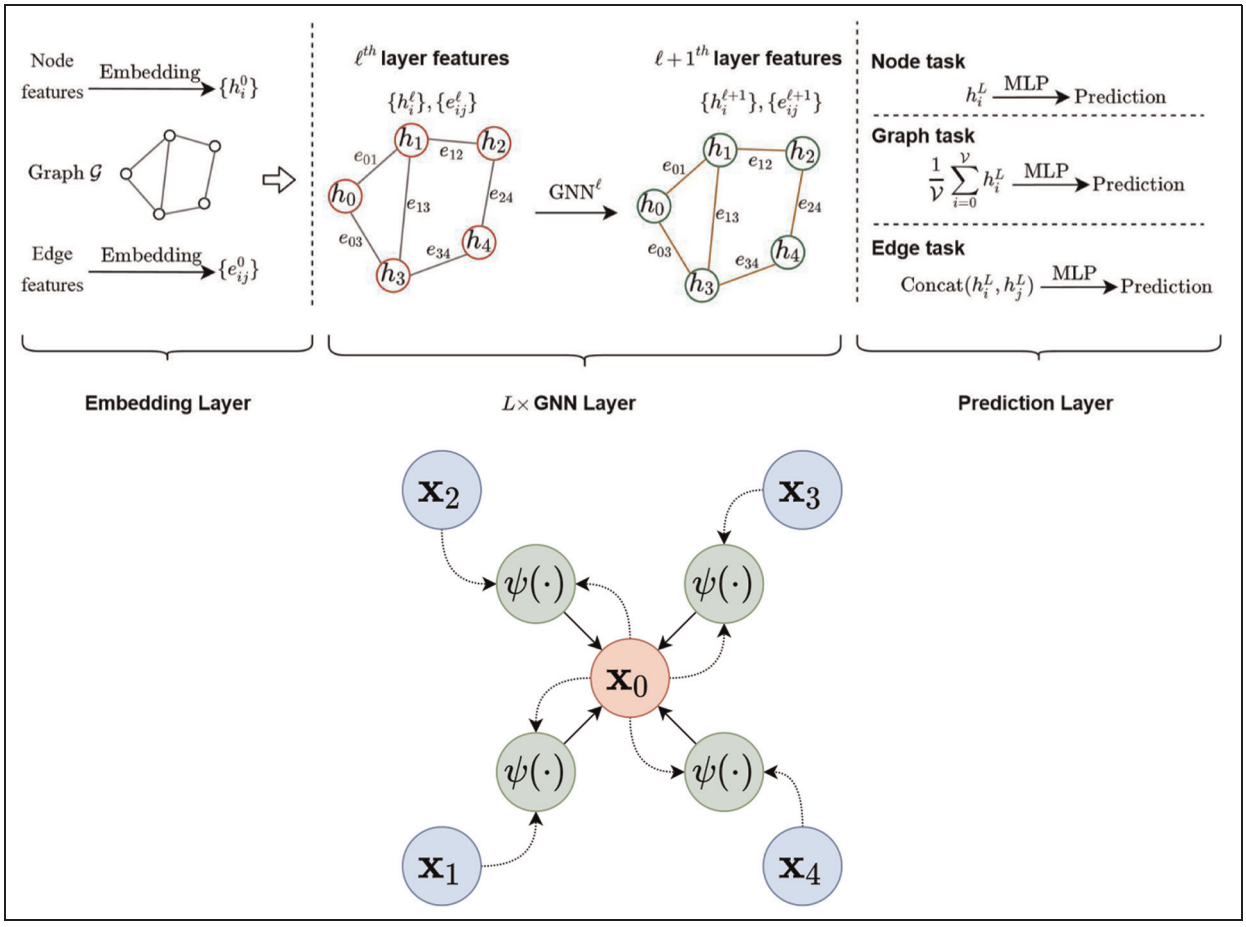

Figure 7. Illustration of the GNN message-passing mechanism. The top panel shows the overall GNN pipeline, while the bottom panel provides a magnified view of a single message-passing step where a target node aggregates information from its neighbors to update its features.
GNN: graph neural network.

edge attributes with `GENConv` graph convolutional layers, is implemented within the PyTorch Geometric library. This architecture explicitly leverages edge information within the powerful `GENConv` framework[36] to better model the sensor interdependencies. The architecture consists of the following components:

1. Input Data:
    - **Node features:** For each time step t, the input includes node features $\mathbf{x}_t \in \mathbb{R}^{16\times 1}$, representing the HI values for each of the $N = 16$ sensors.
    - **Edge attributes:** The graph is defined as fully connected. For every pair of nodes (sensors) (i,j) in this graph, edge attributes $\mathbf{e}_{ij,t}$ are explicitly provided to the model. These are pre-calculated as the HI differential between sensor i and sensor j at time step t: $\mathbf{e}_{ij,t} = HI_i(t) - HI_j(t)$. By incorporating these pre-calculated differentials as edge features, the model is directly informed about the pairwise relational dynamics between sensors. This approach is intended to guide the feature extraction process, offering valuable relational inductive bias, which can be particularly advantageous when working with a limited number of experimental specimens.

A schematic representation of these HI-derived node and edge features utilized by the GENEA model is presented in Figure 8.

2. **GNN layers:** A stack of GNN layers processes the input graph data, utilizing both the node features and the pre-calculated edge attributes. Each layer in this stack primarily employs a `GENConv` operation, chosen for its effectiveness in integrating edge features into the message-passing mechanism. This operation is typically followed by batch

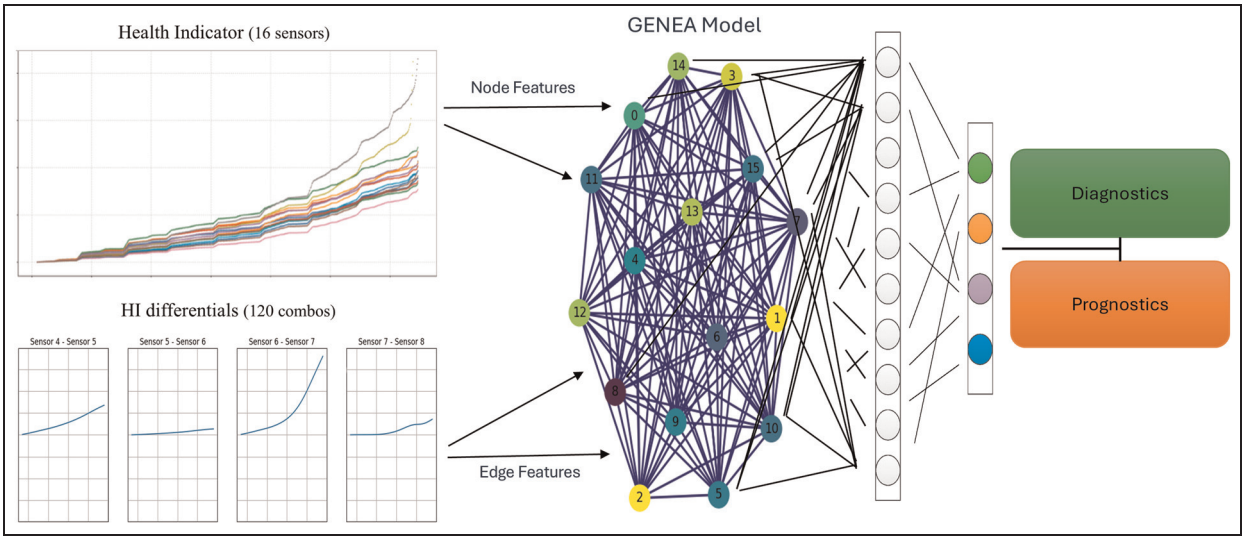

Figure 8. Illustration of the GNN input features for the GENEA model at a given time step t. Node features are the HI values for each sensor (e.g., $\mathrm{HI}_i(t)$). Edge attributes are the pre-calculated HI differentials (e.g., $\mathrm{HI}_i(t) - \mathrm{HI}_j(t)$) between sensors. This figure exemplifies the edge attributes by displaying four such differentials, out of the 120 unique pairwise combinations computed for the fully connected 16-sensor network.
HI: health indicator; GNN: graph neural network; GENEA: GENConv with Edge Attributes.

normalization, a ReLU activation function, and Dropout for regularization. The number of GNN layers and the size of their hidden dimensions are key hyperparameters that were determined through an optimization process (detailed in this subsection, "Hyperparameter optimization").

3. **Graph pooling/Readout:** After the final GNN layer, a global pooling function (e.g., global mean pooling) aggregates the updated node features from across the graph into a single graph-level representation vector, $\mathbf{h}_{\text{graph}}$.
4. **MLP prediction head:** This consolidated graph-level representation $\mathbf{h}_{\text{graph}}$ is then fed into a standard MLP. This MLP head (e.g., comprising two fully connected layers with Rectified Linear Unit (ReLU) activation in the hidden layer and a linear output layer) performs the final regression task, outputting the scalar estimation for stiffness or RUL.

Separate models, sharing this GENEA architecture but trained on different targets, are used for the stiffness estimation task and the RUL prediction.

Hyperparameter optimization. Key hyperparameters for the GNN architecture and training were systematically determined using an automated optimization process. We employed Bayesian optimization (specifically, the Tree-structured Parzen Estimator algorithm) implemented in the Optuna framework.[37] A total of 300 trials were conducted within the LOPOCV scheme, aiming to minimize the mean squared error (MSE) on the validation set for each fold. To accelerate the process, unpromising trials were terminated early using Optuna's "MedianPruner" with a warmup period of 15 steps. The optimization focused on four critical hyperparameters, searching within the following ranges:

- Number of GNN layers: Integer range [2, 5]
- Hidden dimension size: Categorical 16, 32, 64, 128
- Dropout rate (applied within GNN layers): Float range [0.2, 0.6]
- Batch size: Categorical 64, 128, 256

Parameter importance analysis during the study indicated that the dropout rate had the most significant influence on validation performance within the explored space. The configuration yielding the best average validation performance across the CV folds was selected for the final model training and evaluation presented in this paper.

Training procedure and final parameters. The models were trained using the AdamW optimizer[38] and the MSE loss function. A learning rate scheduler `ReduceLROnPlateau` dynamically adjusted the learning rate based on validation loss (initial rate 0.01, reduction factor 0.8, patience 10 epochs). Early stopping, monitoring the validation loss with a patience of 50 epochs, was employed to prevent overfitting and retain the best-performing model weights for each training run. The final optimized hyperparameters, along with other fixed training settings, are summarized in Table 3.

Table 3. GNN model architecture and training hyperparameters resulting from Optuna Bayesian optimization.

Parameter category	Value/setting
Model architecture	
GNN type	GENConv
Num. GNN layers	3
Hidden dimension	16
Graph readout	Global mean pooling
Activation function	ReLU
Regularization	
Dropout rate (p)	0.31
Batch normalization	After each GNN layer
Weight decay (L2)	1×10^{-6}
Training settings	
Task	Regression (stiffness/RUL)
Loss function	Mean squared error (MSE)
Optimizer	AdamW
Learning rate (LR)	0.01
LR scheduler	ReduceLROnPlateau
LR reduction factor	0.8
LR patience	10 epochs
Batch size	64
Max. epochs	2000
Early stopping	Enabled
Early stopping patience	50 epochs

GNN: graph neural network; RUL: remaining useful life.

Uncertainty quantification (MC dropout). To obtain uncertainty estimates, MC dropout is applied during inference. The dropout layers (with probability $p = 0.31$, used during training for regularization) are kept active during prediction. The model performs T multiple stochastic forward passes (e.g., $T = 200$) for the same input time step t. The mean of the T predictions is taken as the final point estimate (stiffness or RUL), and the standard deviation of the T predictions serve as a measure of epistemic uncertainty. This provides confidence intervals around the predictions with low computational overhead compared to fully Bayesian methods.[39]

Results and discussion

This section presents the performance evaluation of the proposed GENEA framework for both stiffness estimation (diagnostics) and RUL prediction (prognostics), using the LOPO CV strategy described in "Dataset construction for GNN."

Stiffness estimation

Target stiffness profiles. The raw stiffness degradation curves for the panels (Figure 9) exhibit significant variability in degradation rates and final failure points. Some panels degrade gradually, while others show sharper drops, highlighting the challenge of developing a generalizable model.

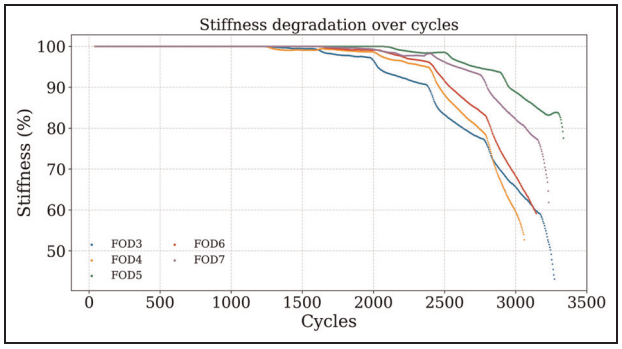

Figure 9. Measured stiffness reduction curves (percentage of initial stiffness) for all five FOD panels over fatigue cycles. FOD: foreign object damage.

Table 4. Stiffness estimation performance (mean over LOPO CV folds) for different truncation thresholds.

Truncation threshold	Mean RMSE	Mean MAPE (%)
85%	1.68	1.12
80%	1.60	1.06
70%	**1.41**	**1.01**
60%	2.06	1.53

LOPO CV: Leave-One-Panel-Out Cross-Validation; RMSE: root mean squared error; MAPE: mean absolute percentage error.
Bolded row is the selected threshold which gives the best model performance.

Stiffness truncation analysis. To determine the optimal training data depth, we systematically evaluated model performance within the operationally critical range of 100% down to 85% remaining stiffness, which is a primary focus for aerospace and manufacturing applications. We trained our model with data truncated at different end-of-life thresholds (85%, 80%, 70%, and 60%) to see how including more late-stage damage information would improve predictions in this critical early-to-mid-life phase. The results, summarized in Table 4, show a clear optimum. Model accuracy improves as more degradation history is included down to the 70% threshold, as this exposes the model to more robust damage patterns. However, including data from the extreme end-of-life (down to 60%) degrades performance, likely because it introduces a bias that is not representative of the critical operational range. Therefore, the 70% threshold was selected as it provides the best predictive accuracy for the most relevant stiffness range.

Prediction performance

Figure 10 shows the predicted stiffness versus the true (truncated) stiffness for each panel when it was used as

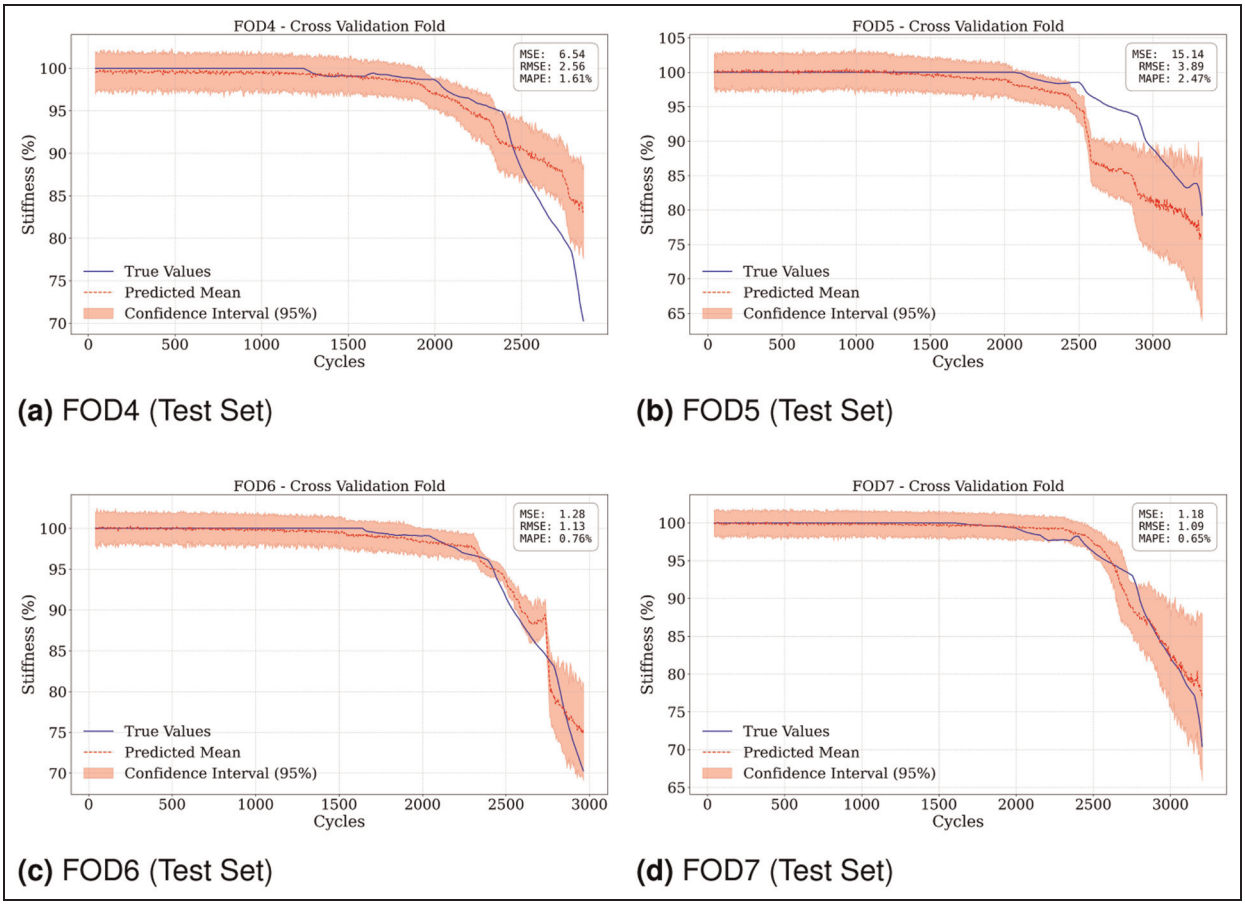

Figure 10. Stiffness estimation results from LOPO CV. Each plot shows the true truncated stiffness (blue) versus the GNN model's prediction (orange mean) with uncertainty bounds (shaded area, from MC dropout) for the held-out test panel: (a) FOD4 (Test Set), (b) FOD5 (Test Set), (c) FOD6 (Test Set), and (d) FOD7 (Test Set).
GNN: graph neural network; LOPO CV: Leave-One-Panel-Out Cross-Validation; MC: Monte Carlo; FOD: foreign object damage.

the test set in the LOPO CV. The plots include the mean prediction and the uncertainty bounds derived from MC dropout with a 95% confidence interval. The model generally tracks the stiffness degradation accurately across all four panels, capturing both gradual declines and sharper drops. The uncertainty bounds tend to widen in regions of rapid change or toward the end of life, reflecting increased model uncertainty as expected.

Quantitative metrics. Standard regression root mean squared error (RMSE) and mean absolute percentage error (MAPE) were calculated for each test fold. The results are summarized in Table 5 and visualized in Figure 11.

The metrics indicate good performance overall, with an average RMSE of 2.17% and MAPE of 1.37% for stiffness estimation, suggesting high accuracy. Performance was notably better on FOD6 and FOD7 compared to FOD4 and FOD5, which exhibited higher prediction errors (Table 5). Performance on panel FOD4 was particularly informative. The raw strain signals for this panel show evidence of intermittent sensor malfunctions, leading to anomalous, high-amplitude strain readings. While this noisy, real-world data

Table 5. Stiffness percentage estimation metrics from LOPO CV (tested on the indicated panel) using the proposed GENEA model.

Test panel	RMSE	MAPE (%)
FOD4	2.56	1.61
FOD5	3.89	2.47
FOD6	1.13	0.76
FOD7	1.09	0.65
Mean	2.17	1.37
Std. Dev.	1.33	0.87

LOPO CV: Leave-One-Panel-Out Cross-Validation; RMSE: root mean squared error; MAPE: mean absolute percentage error; FOD: foreign object damage; GENEA: GENConv with Edge Attributes.

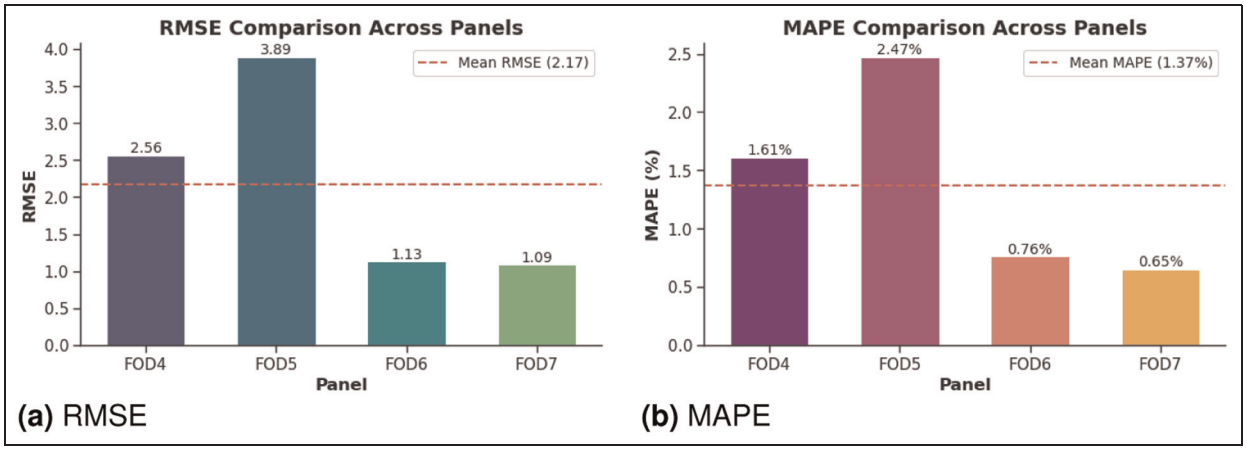

Figure 11. Stiffness estimation evaluation metrics: (a) RMSE and (b) MAPE for each LOPO CV fold (test panel indicated).
LOPO CV: Leave-One-Panel-Out Cross-Validation; RMSE: root mean squared error; MAPE: mean absolute percentage error.

caused the simpler MLP baseline to struggle (as shown in Supplemental Material Appendix B), the GENEA framework still achieved high accuracy. This demonstrates the GNN's robustness and its ability to learn meaningful degradation patterns even in the presence of imperfect or faulty sensor data. Regarding FOD5, the only non-impacted panel in the set, the predictions consistently underestimated the true remaining stiffness (Figure 10(b)). Since the model was trained exclusively on impacted specimens known to degrade faster initially, its predictions for the non-impacted FOD5 panel were likely biased. The model may have interpreted FOD5's distinct response signatures as indicating more damage than was actually present, relative to the patterns learned during training. From a practical SHM perspective, this tendency toward conservative predictions (i.e., estimating slightly worse health than actual) for scenarios outside the training distribution can be considered a desirable characteristic, potentially leading to earlier cautionary alerts. However, it also highlights that incorporating a more diverse training set, including non-impacted specimens, would likely improve the model's generalization capabilities and predictive accuracy for such initially undamaged structures.

Discussion of prediction characteristics. Two characteristics of the model's predictions warrant discussion. First, the tendency to slightly underpredict stiffness at high cycles (e.g., in Figure 10(b)) is likely due to the model learning an "average" degradation trajectory from the varied training set. If a test specimen degrades more slowly than this learned average, the model produces a conservative "fail-safe" prediction. Second, the sudden drops observed in some predictions (e.g., FOD5 and FOD6) are characteristic of deep neural networks learning highly nonlinear decision boundaries. The model learns to associate specific patterns of HI values across the sensor network with significant damage events. When the input features cross a learned threshold corresponding to such an event, the output can shift abruptly to a new, lower-stiffness regime.

Remaining useful life prediction

RUL target definition and multi-threshold approach. RUL prediction involves the estimation of the remaining cycles until a specific End-of-Life (EoL) criterion is met. In the SHM context, EoL can be defined by a critical stiffness degradation threshold. Since different applications may have different tolerance levels, evaluating the prediction performance across various degradation limits is important. Furthermore, investigating multiple EoL definitions allows us to explicitly demonstrate how prediction accuracy and the associated uncertainty levels vary depending on the chosen failure threshold. Predicting RUL to an early, subtle degradation point (e.g., 1% stiffness loss) often presents different challenges, prediction horizons, and confidence levels compared to predicting toward a more advanced failure state (e.g., 15% stiffness loss). Therefore, to accommodate diverse application needs and to thoroughly assess these performance and uncertainty variations across different degradation stages, we adopt a multi-threshold approach.

Specifically, we train separate RUL prediction models for four distinct EoL thresholds, corresponding to common industry practices as summarized in Table 6. This strategy of using dedicated models allows for optimized training and evaluation tailored to the specific characteristics and challenges associated with each EoL definition. For each threshold, the target RUL at

Table 6. EOL stiffness thresholds used for RUL prediction.

EOL threshold	Remaining stiffness (%)	Stiffness degradation (%)
Level 1	99	1
Level 2	95	5
Level 3	90	10
Level 4	85	15

EOL: end-of-life; RUL: remaining useful life.

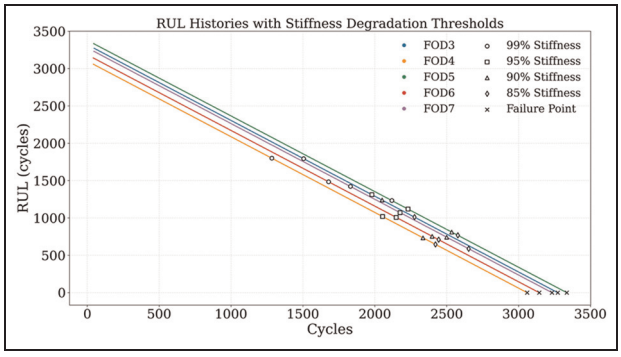

Figure 12. Ground truth RUL histories (in cycles) for panels FOD4–FOD7, calculated for the four different EOL stiffness thresholds (99%, 95%, 90%, and 85%) indicated by distinct markers on the graph.
RUL: remaining useful life; FOD: foreign object damage.

a given time t is calculated as the number of cycles remaining until that stiffness level is first crossed. Figure 12 shows the ground truth RUL curves for all panels across these four thresholds, illustrating the decreasing RUL value as the panel degrades.

RUL prediction performance. Using the same GNN architecture and LOPO CV strategy, separate models were trained for each EoL threshold. Figure 13 (showing results for the 85% and 90% EoL stiffness thresholds) and Figure 14 (showing results for the 95% and 99% EoL stiffness thresholds) present the RUL predictions for each test panel across these four EoL thresholds.

While the predicted RUL generally follows the ground truth, visual analysis highlights some perturbations in the prediction uncertainty. As expected, predictions made earlier in the component's life yield wider uncertainty bounds due to the long prediction horizon. However, the choice of EoL threshold also significantly impacts uncertainty. For earlier thresholds (e.g., 99%, 95% stiffness), where inter-specimen EoL time variation is minimal, the model often produces narrower uncertainty bounds. By contrast, for later thresholds (e.g., 90%, 85%), corresponding to stages where damage evolution diverges more substantially between specimens, the model reflects this increased intrinsic variability by generating noticeably wider uncertainty bounds, in addition to the effect of the prediction horizon.

Quantitative metrics and discussion. To quantify prediction accuracy, standard evaluation metrics (RMSE, MAPE) were calculated for each panel and EOL threshold, with the results visualized in Figure 15 for comparative analysis across all test cases. Analyzing the RUL prediction metrics reveals a clear trend: prediction accuracy is highest (i.e., lower RMSE and MAPE) for the most stringent EOL threshold (99% remaining stiffness) and tends to decrease as the threshold becomes less stringent (moving toward 85%). This trend can be attributed to the level of inter-specimen variability at different stages of degradation. In the early damage phase (e.g., 1% degradation), the degradation behavior is highly consistent across all specimens, providing the model with a clear, low-variance learning target. This allows for more precise RUL predictions. Conversely, at later stages of life (e.g., 15% degradation), the stochastic nature of damage accumulation causes the degradation paths of the individual panels to diverge significantly. This high variability in the training data makes the prognostic task inherently more challenging, resulting in larger prediction errors for an unseen test specimen. Performance variation across panels mirrors the stiffness estimation results. FOD5, the non-impacted panel, often shows slightly higher errors, suggesting its degradation pattern might differ from the impacted panels used predominantly for training. FOD4 also shows variability, likely due to the aforementioned signal quality issues. FOD6 and FOD7 generally exhibit the lowest prediction errors across thresholds. The magnitude of the errors (RMSE typically ranging from 100 to 400 cycles, depending on the panel and threshold) needs context. Given the total life of 3300 cycles, these errors represent approximately 3%–12% of the total lifespan. While lower errors are always desirable, these levels, particularly for the later EoL thresholds (90%, 85%), often used in practice, could be considered acceptable for informing maintenance planning, especially when coupled with the uncertainty bounds. The RUL predictions exhibit similar characteristics to the stiffness estimations. The tendency toward conservative predictions and the presence of sudden drops in the estimated RUL are attributed to the same underlying model behaviors, learning an "average" degradation trajectory and nonlinear decision boundaries, as detailed in section "Stiffness estimation." The consistent performance across different panels and EoL thresholds suggests that the GNN + HI approach effectively captures

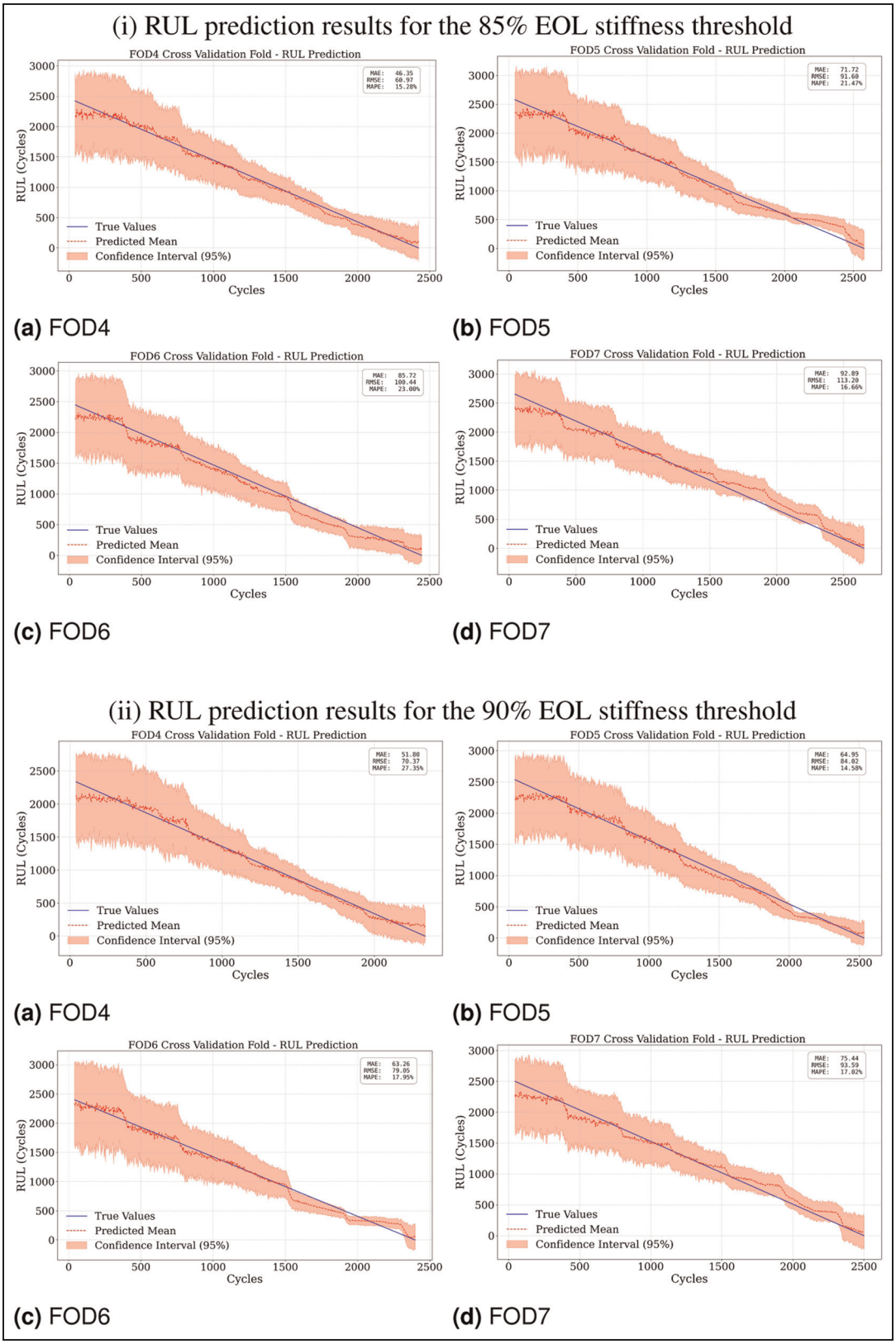

Figure 13. RUL prediction results (LOPO CV) for 85% (i) and 90% (ii) EOL stiffness thresholds. Plots compare true RUL (blue) versus predicted mean (orange) with MC dropout uncertainty (shaded). (i) RUL prediction results for the 85% EOL stiffness threshold. (a) FOD4. (b) FOD5. (c) FOD6. (d) FOD7. (ii) RUL prediction results for the 90% EOL stiffness threshold. (a) FOD4. (b) FOD5. (c) FOD6. (d) FOD7.

EOL: end-of-life; LOPO CV: Leave-One-Panel-Out Cross-Validation; FOD: foreign object damage; RUL: remaining useful life.

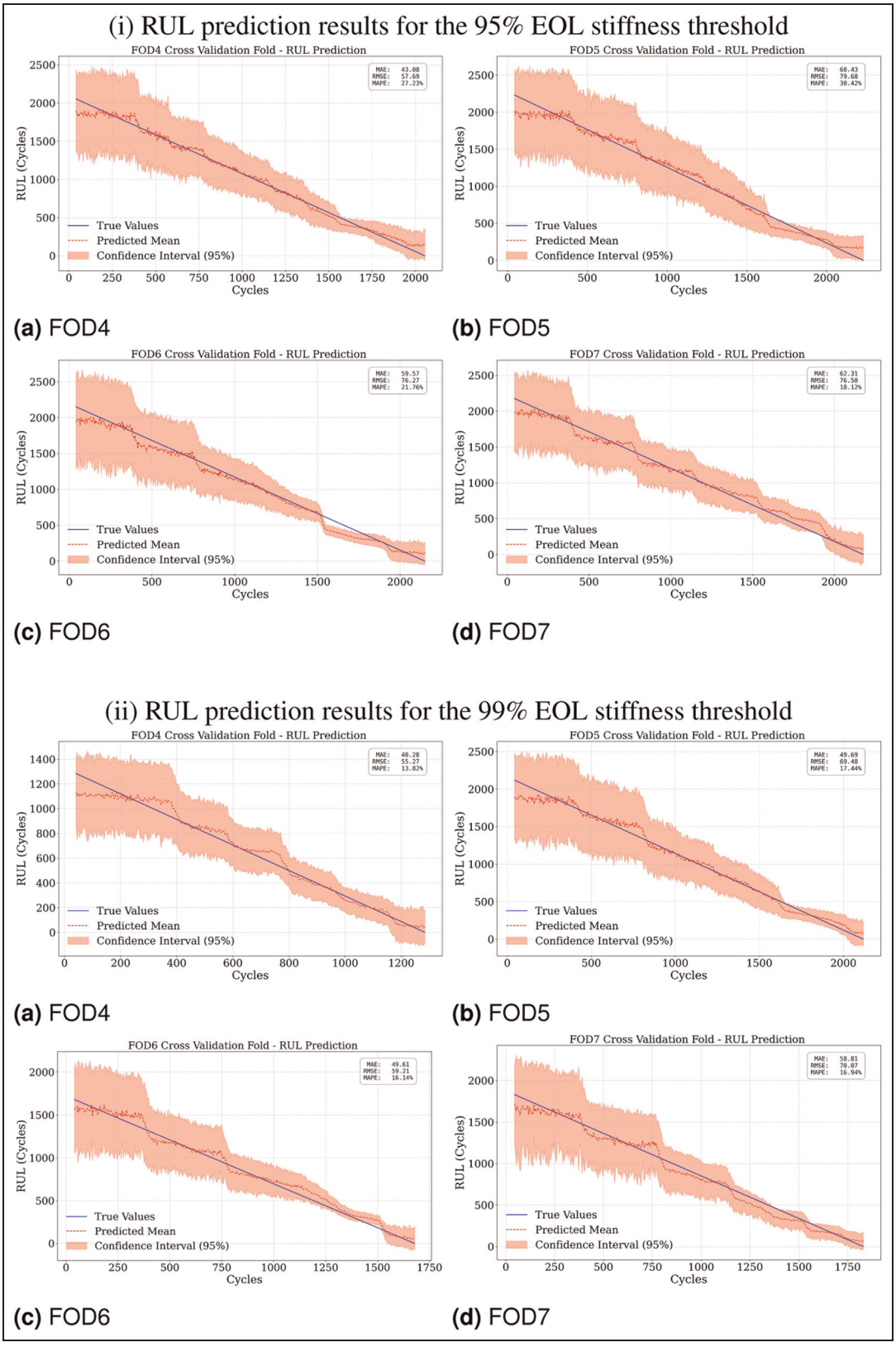

Figure 14. RUL prediction results (LOPO CV) for 95% (i) and 99% (ii) EOL stiffness thresholds. Plots compare true RUL (blue) versus predicted mean (orange) with MC dropout uncertainty (shaded). (i) RUL prediction results for the 95% EOL stiffness threshold. (a) FOD4. (b) FOD5. (c) FOD6. (d) FOD7. (ii) RUL prediction results for the 99% EOL stiffness threshold. (a) FOD4. (b) FOD5. (c) FOD6. (d) FOD7.

EOL: end-of-life; LOPO CV: Leave-One-Panel-Out Cross-Validation; FOD: foreign object damage; RUL: remaining useful life.

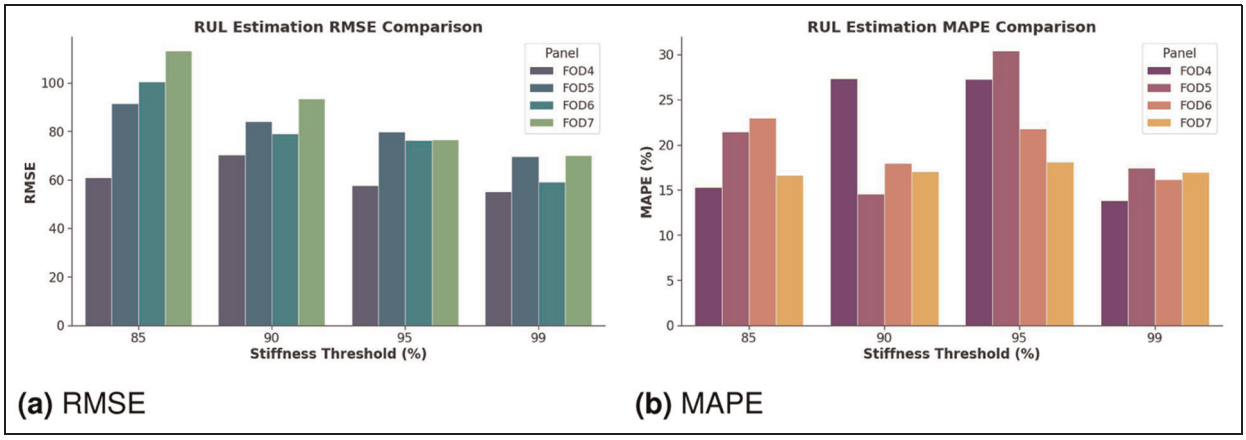

Figure 15. RUL prediction evaluation metrics (a) RMSE and (b) MAPE grouped by test panel (LOPO CV fold) and plotted against each stiffness EoL threshold.
EOL: end-of-life; LOPO CV: Leave-One-Panel-Out Cross-Validation; RUL: remaining useful life; RMSE: root mean squared error; MAPE: mean absolute percentage error.

degradation-related information relevant for RUL prediction. The MC dropout uncertainty bounds appear reasonable, generally being wider for predictions made early in life or for panels/thresholds with higher variability, appropriately reflecting the model's confidence. Overall, the results suggest that the proposed GNN framework, utilizing the load-decoupled HI and spatial sensor information, provides robust and accurate RUL predictions across various EoL criteria and specimens with different damage histories. The ability to generate reliable predictions with uncertainty estimates enhances the practical applicability of the method for CBM decision-making.

FAIR model comparison

To rigorously evaluate the effectiveness of the proposed GNN framework, this section presents a multi-stage comparative analysis following the FAIR protocol.[32] First, we benchmark our `GENConv`-based model against several other GNN architectures to validate our specific design choices. Next, we compare the performance of our winning GENEA model against two key baselines: a fundamental MLP and a more sophisticated, spatially aware 2D CNN. To ensure a direct and fair comparison against these fixed-input models, this head-to-head analysis is conducted on the consistent 16-sensor dataset (FOD4-7). An ablation study is also presented to quantify the critical contribution of our proposed health indicator. This comprehensive comparison on a fixed geometry sets the stage for the subsequent section, which tests the GNN's key architectural advantage: its ability to generalize across varying sensor geometries.

Benchmarking GNNs. We evaluated several prominent GNN models employing different graph convolution mechanisms, focusing on the impact of utilizing edge attributes within the fully connected sensor graph structure. All tested architectures, detailed in Table 7, were optimized and trained under comparable and FAIR conditions using the HI features to predict the truncated stiffness (70% threshold). Tables 8 and 9 summarize the average performance metrics across the LOPO CV folds for GNN models grouped by whether their core layers utilize edge attributes. The results identified *GENConv* as the best-performing core layer type among those using edge attributes (Table 8). This aligns with our proposed GENEA architecture detailed in "GNN architecture and training." Among models not using edge attributes, *GraphSAGE* performed best (Table 9). Comparing the best from each category, the GENEA model showed a significant benefit (22% RMSE improvement based on those specific runs).

Comparison with MLP and CNN baselines. Crucially, all models in this comparison (GENEA, CNN, and MLP) were trained using the same point-wise approach, taking the HI features from a single time step as input to predict that same step. This deliberate experimental design ensures the comparison focuses purely on the architectural benefit of processing spatial information, avoiding confounding the analysis with the temporal processing capabilities inherent in sequence-based models (like LSTMs or Transformers), which were intentionally excluded from this specific benchmark. To rigorously evaluate the performance of our proposed GENEA framework, we compared it against two key baselines on the consistent four-panel dataset (FOD4-

Table 7. Overview of GNN architectures evaluated.

Model name	Core layer type/description	Edge Attr.
GENConv	Generalized Powerful Graph Convolution	Yes
GCN	Standard Graph Convolutional Network	Yes
EdgeConv	Dynamic Edge Feature Computation	Yes
SAGPool	Self-Attention Graph Pooling Architecture	Yes
GATv2	Graph Attention Network v2	Yes
GraphSAGE	Graph Sample and Aggregate	No
GIN	Graph Isomorphism Network	No
ChebConv	Chebyshev Spectral Graph Convolution	No
GCN_NoEdges	Standard Graph Convolutional Network	No
SGConv	Simplified Graph Convolution	No

GNN: graph neural network.

Table 8. Stiffness estimation performance comparison for GNN models with edge attributes (average over LOPO CV folds).

Model	Mean RMSE	Mean MAPE (%)
GENEA (proposed)	**2.17**	**1.37**
GCN	2.585	1.844
EdgeConv	2.862	1.788
SAGPool-based	2.573	1.817
GATv2	3.171	2.068

GENEA: GENConv with Edge Attributes; GNN: graph neural network; LOPO CV: Leave-One-Panel-Out Cross-Validation; RMSE: root mean squared error; MAPE: mean absolute percentage error.
Bolded row cooresponds to the best perfomring model.

Table 9. Stiffness estimation performance comparison for GNN models without edge attributes (average over LOPO CV folds).

Model	Mean RMSE	Mean MAPE (%)
GraphSAGE	2.884	1.664
GIN	2.768	1.986
ChebConv	2.925	2.180
GCN (No Edge Attr.)	3.185	2.415
SGConv	3.671	2.628

GNN: graph neural network; LOPO CV: Leave-One-Panel-Out Cross-Validation; RMSE: root mean squared error; MAPE: mean absolute percentage error.

7): a fundamental MLP baseline with no spatial awareness, and a more sophisticated, spatially aware 2D CNN baseline. For the CNN, the 16-sensor HI input vector was reshaped into a 4×4 grid. Both baselines were optimized using the same FAIR protocol as our GNN models. The comparative performance is presented in Table 10. The results show that our proposed GENEA model is demonstrably superior, achieving a significantly lower Mean RMSE (2.17) compared to both the MLP (3.27) and the CNN (3.41). Notably, the elementary MLP slightly outperforms the CNN, suggesting that imposing a rigid grid structure on the sensor data may be less effective than a simple vectorized approach if the grid does not perfectly capture the underlying physical relationships. This result highlights the challenge of using grid-based models for non-Euclidean sensor layouts and motivates the use of a more flexible graph-based approach. The challenge of incorporating the geometrically distinct FOD3 panel is explored in detail in the following section.

Furthermore, to specifically isolate and quantify the contribution of the proposed HI, an additional ablation study was conducted. The same optimized MLP baseline architecture was trained and evaluated under the LOPO CV scheme, but this time using the preprocessed (downsampled and smoothed) strain data directly as input, *without* the HI transformation (Equation (1)).

The performance degraded considerably across all metrics compared to the MLP using the HI features, as summarized in Table 11. The averaged Root Mean Squared Error (RMSE) increased by 70%, and the Mean Absolute Percentage Error (MAPE) by 78%.

Notably, the performance drop was most pronounced for the outlier panels FOD4 (MAPE: 7.95%) and FOD5 (MAPE: 5.58%), which exhibited noisier signals or different initial damage states. While the degradation was less severe for the more typical panels FOD6 (MAPE: 2.51%) and FOD7 (MAPE: 0.96%), performance still worsened compared to using the HI. These results strongly underscore the effectiveness of the proposed HI in decoupling confounding effects and extracting robust degradation-sensitive features from raw strain signals. The HI significantly enhances model accuracy and generalization, proving particularly crucial for handling challenging outlier specimens often encountered in real-world scenarios. For visual comparison, the individual prediction plots for the various baseline models are provided in the appendices. The results for the MLP baseline (fourfold CV) are shown in Supplemental Material Appendix B, while the results for the CNN baseline (fourfold CV) are in Supplemental Material Appendix D. Finally, the plots for the ablation study, showing the MLP performance without the proposed HI, are presented in Supplemental Material Appendix C.

Generalization to varying geometries: incorporating FOD3

This section explicitly tests the ability of the spatially aware models to handle varying graph structures by

Table 10. Detailed stiffness estimation performance from fourfold LOPO CV: GENEA versus baselines.

Test panel	RMSE			MAPE (%)		
	GENEA	CNN	MLP	GENEA	CNN	MLP
FOD4	2.56	4.63	4.75	1.61	2.52	3.71
FOD5	3.89	7.10	5.41	2.47	4.47	3.85
FOD6	1.13	2.52	1.76	0.76	1.59	1.09
FOD7	1.09	1.31	1.17	0.65	0.90	0.89
Mean	2.17	**3.89**	**3.27**	1.37	**2.37**	**2.38**
Std. Dev.	1.33	**2.25**	**2.05**	0.87	**1.42**	**1.43**

GENEA: GENConv with Edge Attributes; CNN: convolutional neural network; LOPO CV: Leave-One-Panel-Out Cross-Validation; RMSE: root mean squared error; MAPE: mean absolute percentage error; FOD: foreign object damage; MLP: multi-layer perceptron.
The significance of these corrected bolds is that they represent the best performing model.

Table 11. Performance comparison: MLP baseline with versus without proposed HI.

MLP input features	Mean RMSE	Mean MAPE (%)
With the proposed HI	3.14	2.39
Without HI (strain)	5.34	4.25
Degradation	−70%	−78%

RMSE: root mean squared error; MAPE: mean absolute percentage error; HI: health indicator; MLP: multi-layer perceptron.

Table 12. Detailed comparison of fourfold and fivefold CV performance for GENEA and CNN models.

	RMSE				MAPE (%)			
	GENEA		CNN		GENEA		CNN	
Test panel	Fourfold	Fivefold	Fourfold	Fivefold	Fourfold	Fivefold	Fourfold	Fivefold
FOD3 (6s)	—	2.08	—	7.78	—	1.43	—	5.27
FOD4 (16s)	2.56	2.41	4.63	3.57	1.61	1.39	2.52	1.88
FOD5 (16s)	3.89	3.25	7.10	4.75	2.47	1.67	4.47	2.90
FOD6 (16s)	1.13	1.01	2.52	4.99	0.76	0.70	1.59	2.97
FOD7 (16s)	1.09	0.97	1.31	3.64	0.65	0.63	0.90	2.33
Mean (FOD4-7)	2.17	1.91	3.89	4.23	1.37	1.10	2.37	2.52

CNN: convolutional neural network; CV: cross-validation; FOD: foreign object damage; GENEA: GENConv with Edge Attributes; RMSE: root mean squared error; MAPE: mean absolute percentage error.

incorporating the six-sensor FOD3 panel in a fivefold LOPO CV. The GENEA model processes this heterogeneous data natively, while the CNN requires the six-sensor data to be padded with zeros into a 4×4 grid. Table 12 presents a comprehensive comparison of the fourfold and fivefold cross-validation results for both models. The results in Table 12 highlight several critical findings. First, comparing the fivefold columns, the GENEA model's ability to handle the six-sensor panel natively is far superior to the CNN's padding approach. The GNN makes a strong prediction on the unseen FOD3 geometry (RMSE 2.08), while the CNN's performance on the padded data remains very poor (RMSE 7.78), demonstrating the fragility of the artificial padding workaround. This successful prediction by the GENEA model is visualized in Figure 16, which shows the model's output closely tracking the true stiffness degradation. Second, and more importantly, the two models show opposite reactions to being trained on a more diverse dataset. For the GENEA model, incorporating the geometrically distinct FOD3 panel significantly enhances its generalizability, with the mean RMSE on the 16-sensor panels dropping from 2.17 to 1.91. In stark contrast, for the CNN, including the padded FOD3 data in its training set degrades its performance on the original panels, with the mean RMSE increasing from 3.89 to 4.23. This opposing behavior is a key finding. It confirms that GNNs are not only capable of handling varied geometries natively but actually become more robust

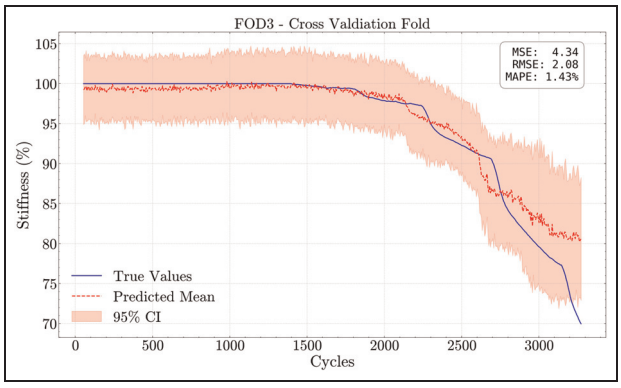

Figure 16. Stiffness estimation by the GENEA model on the unseen six-sensor FOD3 panel. The plot demonstrates the model's ability to successfully generalize to a new sensor geometry.
GENEA: GENConv with edge attributes; FOD: foreign object damage.

and reliable when trained on more diverse data. Conversely, it suggests that for a CNN, forcing it to learn from padded, out-of-distribution data can confuse the model and harm its ability to generalize, even on data it is designed for. This strongly reinforces that a graph-based architecture is fundamentally more suitable for heterogeneous SHM datasets. The full set of prediction plots for this fivefold CNN experiment, visually demonstrating its poor performance on the padded FOD3 data, can be found in Supplemental Material Appendix E.

Conclusions and recommendations

This study introduced a novel framework for SHM of composite structures based on GNNs and strain sensor data. The key innovations include the use of a GNN to explicitly model and leverage spatial interdependencies within a sensor network and the development of a custom HI based on the cumulative absolute strain derivative to effectively decouple the confounding effects of applied load from true structural degradation signatures. The key findings of the present work are summarized as follows:

- The proposed HI successfully captured progressive damage trends in strain data while mitigating the dominant influence of load variations, providing a more robust input feature for diagnostic and prognostic models.
- Our proposed input design for the GENEA model, using HI-derived node and edge attributes, effectively learned meaningful spatial patterns from the multi-sensor HI data, representing the complex strain field redistributions associated with damage accumulation in the anisotropic composite panels.
- The integrated framework incorporating GENEA and the HI demonstrated high accuracy in stiffness estimation (diagnostics), achieving an average MAPE of 1.37% across different test panels in a LOPO CV setting.
- The framework also yielded promising results for RUL prediction (prognostics), providing accurate RUL estimates for various industry-relevant EOL stiffness thresholds (1%–15% degradation).
- The GENEA model demonstrated promising generalization capabilities across diverse and challenging damage scenarios. It performed well when tested on panels with varying initial conditions, including a pristine panel and panels with impact damage introduced at different locations. Crucially, its ability to predict health states on test panels with unique, stochastic damage propagation paths indicates that the framework learns a generalized mapping from strain-field patterns to damage, rather than overfitting to specific damage locations.
- The incorporation of MC dropout in the GENEA model provided computationally efficient uncertainty quantification, generating informative confidence bounds for both stiffness and RUL predictions, which enhances the reliability and trustworthiness of the SHM system.
- Systematic comparative evaluations against spatially aware baselines confirmed the framework's superiority. The GENEA model significantly outperformed not only a fundamental MLP baseline but also a more sophisticated 2D CNN designed to capture local spatial patterns. Notably, the CNN's performance was slightly worse than the MLP's, suggesting that imposing a rigid grid structure on the non-Euclidean sensor data is an ineffective strategy and reinforcing the suitability of a flexible, graph-based approach.
- The specific choice of the GENEA model, which uses `GENConv` layers with pre-calculated HI differentials as edge attributes, was identified as a key factor in its strong performance. The selection of `GENConv` is justified empirically by the FAIR model comparison (see Table 8), where it outperformed other edge-aware architectures like GATv2 for this task. The use of HI differentials as edge features provides a strong relational inductive bias, which is particularly effective given the limited training data. By explicitly feeding the model the differential degradation state between sensor pairs, we inject domain

knowledge that simplifies the learning task. The model is no longer required to learn this fundamental operation of subtraction; instead, it can dedicate its capacity to learning more complex, nonlinear relationships from the data. This acts as a powerful regularizer, guiding the model toward a robust solution and preventing it from learning spurious correlations, which is a significant risk with small datasets.

- An ablation study further confirmed the critical contribution of the proposed HI, showing significantly degraded performance (especially for outlier panels) when the MLP model was trained directly on processed strain data without the HI transformation.
- The GNN's inherent ability to handle varying graph geometries was successfully demonstrated and contrasted with the CNN. In a fivefold cross-validation, the GENEA model effectively generalized to the unseen six-sensor geometry. In stark contrast, the CNN required an artificial padding strategy that proved fragile, resulting in poor performance. Furthermore, including the geometrically diverse data in the training set improved the GNN's generalization on the original panels, while it degraded the CNN's, highlighting the GNN's superior ability to learn from heterogeneous SHM datasets.

Regarding the limitations as well as our plans for future work, we highlight the following points:

- The study was conducted on a small experimental dataset of five panels. While the fivefold LOPO CV provides a robust estimate of performance on this data, the limited number of specimens inherently restricts the ability to make broad claims about universal generalization. The observed performance drop on the non-impacted outlier panel (FOD5) underscores the challenge of generalizing to samples outside the training distribution and highlights the need for even more diverse data in future work.
- The FAIR model comparison was performed under the constraint of data scarcity. It is possible that more complex architectures (e.g., deeper GNNs) were penalized by overfitting and might outperform the proposed GENEA model if a significantly larger dataset were available. The conclusions on model performance should be interpreted within this low-data context.
- While data augmentation using FE simulations is a standard approach, it was deemed infeasible for this study due to the extreme complexity of the specimens (3D woven composite, hybrid material interfaces, curved geometry). Creating a validated high-fidelity model would constitute a separate, significant research project. Our work, therefore, focused on what can be achieved with a purely data-driven approach.
- The study was based on data from a specific type of hybrid composite-metal panel under four-point bending fatigue. Further validation is needed on different composite materials, geometries, sensor types (e.g., acoustic emission, temperature), and loading conditions (e.g., tension-tension fatigue, variable amplitude loading, impact events).
- The framework utilized a fully connected graph to maintain a location-agnostic approach, which proved effective for handling the geometric diversity in the dataset. Future work could explore incorporating physics-based priors (e.g., connecting nodes based on physical proximity) to potentially improve performance, while carefully considering the trade-off with the model's ability to generalize to new sensor layouts.
- While MC dropout provides efficient uncertainty estimation, it does not inherently model aleatoric uncertainty, as it focuses on variability in model predictions due to parameter uncertainty rather than data noise.
- The current study deliberately employed a point-wise GNN approach (with the GENEA model) to specifically isolate and evaluate the benefits of spatial feature extraction using the proposed HI. Consequently, while compared against point-wise MLP and alternative GNNs, a direct quantitative comparison with state-of-the-art *sequence-based* deep learning architectures (e.g., RNNs, CNNs, Transformers), potentially alongside sequence-aware GNN models capable of handling spatiotemporal dependencies, remains an important area for future investigation to fully contextualize performance.
- Investigate the trade-offs between explicitly providing edge features (as in the GENEA model) and allowing the GNN to learn edge relationships implicitly, particularly as the size and diversity of the training dataset increase. This could clarify architectural choices for different data availability scenarios and determine if relying solely on the GNN for feature extraction becomes more advantageous with more extensive data.
- Investigating the interpretability of the GENEA model (e.g., using GNN explanation techniques to identify critical sensors or spatial patterns related to specific damage states) would enhance understanding and trust.

- Future work could build upon the model's demonstrated ability to generalize to a new sensor geometry (FOD3) by further enhancing its robustness for dynamic or adaptive sensor networks. Investigating training strategies like sensor dropout and node shuffling could explicitly improve the model's resilience to real-world challenges such as sensor failure or deployment on entirely new sensor layouts, increasing its real-world readiness.
- The proposed HI was intentionally designed for quasi-static fatigue, where our preprocessing isolates the low-frequency, irreversible strain shifts characteristic of cumulative damage. Future work could generalize this derivative-based HI for dynamic loading by reformulating it as a multi-dimensional vector that accumulates the element-wise differences of per-cycle feature vectors.

Computational cost and scalability

All models were trained on an NVIDIA GTX 1080 GPU. The training for a single cross-validation fold was highly efficient, taking approximately 1 min. This efficiency is due to the model's small size (around 900 trainable parameters), a deliberate choice to prevent overfitting on the limited experimental dataset. The use of a fully connected graph was a deliberate design choice to create a location-agnostic framework. This approach is particularly powerful in scenarios where sensor coordinates are unknown or, as in this study, vary between specimens. By connecting every sensor to every other sensor, we make no prior assumptions about the physical layout and force the model to learn directly from the relational patterns in the HI data. To clarify, this "location-agnostic" approach does not discard spatial information; rather, it redefines it. The model does not learn from fixed geometric coordinates (the physical "location"). Instead, it learns from the topology of the signal values themselves, the complex pattern of HI differentials across the entire sensor network at a given moment in time. The GNN's task is to recognize how this abstract spatial pattern of signals changes as damage progresses. The robustness of this approach is validated by our experimental setup; the inclusion of the FOD3 panel, with its completely different six-sensor layout, provides a challenging test of the model's ability to handle geometric diversity. However, we acknowledge that the computational and memory complexity of this fully connected approach grows quadratically ($O(N^2)$) with the number of sensors (N), posing a scalability challenge for larger networks. For real-world, large-scale SHM systems, several strategies could be adopted as important directions for future work:

- **Graph sparsification:** A common strategy is to construct a more efficient sparse graph by connecting nodes based on physical proximity (e.g., k-Nearest Neighbors). While computationally efficient, this risks introducing a geometric bias, potentially causing the model to overfit to the specific sensor layout and fail to generalize to new components with different arrangements.
- Neighbor sampling: Another well-established strategy focuses on the training process itself. Scalable GNN methods, most notably GraphSAGE,[40] do not aggregate information from all neighbors. Instead, they sample a fixed-size neighborhood for each node at every layer. This keeps the computational cost per batch constant regardless of the graph's size, making it highly effective for large-scale applications.

In conclusion, this work demonstrates the significant potential of combining GNNs with a carefully designed, load-decoupled HI for robust and accurate strain-based SHM of composite structures. By effectively capturing spatial sensor interdependencies and isolating degradation signals, the proposed approach offers a promising path toward more reliable diagnostics, prognostics, and condition-based maintenance. Ultimately, the synergistic combination of GNNs for spatial reasoning and a tailored HI for load decoupling provides a powerful framework for advancing SHM capabilities in aerospace and other engineering applications utilizing composite materials.

Acknowledgments

The authors would like to acknowledge Safran Composites for the procurement of the FOD panels, Fraunhofer IFAM for the printed PZT sensors, and FiSens GmbH for the optical fibers and optical fiber interrogation systems.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101006854.

ORCID iDs

Giannis Stamatelatos https://orcid.org/0009-0009-3560-6639
Georgios Galanopoulos https://orcid.org/0000-0003-4998-1308

Data and code availability statement

The source code developed for the GNN models and analysis is publicly available on GitHub at: https://github.com/johnstamly/GNN-Spatial-Strain-SHM. The raw experimental dataset analyzed during the current study has been deposited in the Zenodo open research repository.

Supplemental material

Supplemental material for this article is available online.

References

1. Soutis C. Fibre reinforced composites in aircraft construction. *Progress Aerospace Sci* 2005; 41(2): 143–151.
2. Boller C, Chang F and Fujino Y. *Encyclopedia of structural health monitoring*, vol. 4. Hoboken: Wiley, 2009.
3. Wu R and Jahanshahi M. Data fusion approaches for structural health monitoring and system identification: past, present, and future. *Struct Health Monitor* 2020; 19(2): 552–586.
4. Güemes A, Fernández-López A, Díaz-Maroto P, et al. Structural health monitoring in composite structures by fiber-optic sensors. *Sensors* 2018; 18(4): 1094.
5. Bronstein M, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond euclidean data. *IEEE Signal Process Magaz* 2017; 34(4): 18–42.
6. Güemes A, Fernandez-Lopez A, Pozo A, et al. Structural health monitoring for advanced composite structures: a review. *J Compos Sci* 2020; 4(1): 13.
7. Galanopoulos G, Eleftheroglou N, Milanoski D, et al. A novel strain-based health indicator for the remaining useful life estimation of degrading composite structures. *Compos Struct* 2023; 306: 116579.
8. Moradi M, Broer A, Chiachío J, et al. Intelligent health indicator construction for prognostics of composite structures utilizing a semi-supervised deep neural network and SHM data. *Eng Appl Artif Intell* 2023; 117: 105502.
9. Cristiani D, Falcetelli F, Yue N, et al. Strain-based delamination prediction in fatigue loaded CFRP coupon specimens by deep learning and static loading data. *Compos Part B: Eng* 2022; 241: 110020.
10. Galanopoulos G, Milanoski D, Broer A, et al. Health monitoring of aerospace structures utilizing novel health indicators extracted from complex strain and acoustic emission data. *Sensors* 2021; 21(17): 5701.
11. Milanoski D and Loutas T. Strain-based health indicators for the structural health monitoring of stiffened composite panels. *J Intell Mater Syst Struct* 2021; 32(3): 255–266.
12. Wen P, Zhao S, Chen S, et al. A generalized remaining useful life prediction method for complex systems based on composite health indicator. *Reliab Eng Syst Safety* 2021; 205: 107241.
13. Wu F, Wu Q, Tan Y, et al. Remaining useful life prediction based on deep learning: a survey. *Sensors* 2024; 24(11): 3454.
14. Eleftheroglou N, Galanopoulos G and Loutas T. Similarity learning hidden semi-markov model for adaptive prognostics of composite structures. *Reliab Eng Syst Safety* 2024; 243: 109808.
15. Guo J, Zhang Y and Wang J. Real-time prediction of remaining useful life for composite laminates with unknown inputs and varying threshold. *Machines* 2022; 10(12): 1190.
16. Eleftheroglou N and Loutas T. Fatigue damage diagnostics and prognostics of composites utilizing structural health monitoring data and stochastic processes. *Struct Health Monit* 2016; 15: 473–488.
17. Eleftheroglou N, Zarouchas D, Loutas T, et al. Structural health monitoring data fusion for in-situ life prognosis of composite structures. *Reliab Eng Syst Safety* 2018; 178: 40–54.
18. Farrar C and Worden K. *Structural health monitoring: A machine learning perspective*. Chichester, UK: Wiley, 2012.
19. Bishop C. Pattern recognition and machine learning. 1 ed. In: Michael IJ, John DL, and Christopher MB (eds.) *Information science and statistics*, New York, NY: Springer, 2006.
20. Kubat M. Neural networks: a comprehensive foundation. *Knowledge Eng Rev* 1999; 13(4): 409–412.
21. LeCun Y and Bengio Y. Convolutional networks for images, speech, and time series. In: Arbib M (ed.) *Handbook of brain theory and neural networks*. Cambridge, MA: MIT Press, 1995, pp. 255–258.
22. Hochreiter S. Recurrent neural net learning and vanishing gradient. *Int J Uncertainity Fuzziness Knowledge Based Syst* 1998; 6(2): 107–116.
23. Defferrard M, Bresson X and Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: *The 29th Advances in Neural Information Processing Systems conference (NeurIPS 2016)*. Barcelona, Spain, December 2016.
24. Dwivedi V and Bresson X. A generalization of transformers to graphs. In: *Proceedings of the AAAI conference on artificial intelligence*. Virtual due to covid-19. 2021, 35, pp. 3985–3992.
25. Scarselli F, Gori M, Tsoi A, et al. The graph neural network model. *IEEE Trans Neural Netw* 2009; 20(1): 61–80.
26. Kanatsoulis C and Ribeiro A. Representation power of graph neural networks: improved expressivity via algebraic analysis. arXiv:2205.09801, 2022.
27. Zhou J, Cui G, Hu S, et al. Graph neural networks: a review of methods and applications. *AI Open* 2020; 1: 57–81.
28. Bloemheuvel S, van den Hoogen J and Atzmueller M. A computational framework for modeling complex sensor network data using graph signal processing and graph neural networks in structural health monitoring. *Appl Netw Sci* 2021; 6(1): 97.

29. Wang Y, Wu M, Jin R, et al. Local-global correlation fusion-based graph neural network for remaining useful life prediction. *IEEE Trans Neural Netw Learn Syst* 2025; 36(1): 753–766.
30. Wang H, Zhang Z, Li X, et al. Comprehensive dynamic structure graph neural network for aero-engine remaining useful life prediction. *IEEE Trans Instrumen Meas* 2023; 72: 1–16.
31. Kong Z, Jin X, Xu Z, et al. Spatio-temporal fusion attention: a novel approach for remaining useful life prediction based on graph neural network. *IEEE Trans Instrumen Meas* 2022; 71: 1–12.
32. Shukla K, Toscano J, Wang Z, et al. A comprehensive and FAIR comparison between MLP and KAN representations for differential equations and operator networks. *Comput Methods Appl Mech Eng* 2024; 431: 117290.
33. Galanopoulos G, Paunikar S, Stamatelatos G, et al. SHM for complex composite aerospace structures: a case study on engine fan blades. *Aerospace* 2025; Under review.
34. Paunikar S, Galanopoulos G and Rébillat M. An experimental data set for the shm of a substructure of an engine fan blade from the morpho project, 2025. DOI:10.5281/zenodo.14627730.
35. Dwivedi V, Joshi C, Luu A, et al. Benchmarking graph neural networks. arXiv.2003.00982, 2022.
36. Li G, Xiong C, Thabet A, et al. DeeperGCN: all you need to train deeper GCNs, arXiv.2006.07739, 2020.
37. Akiba T, Sano S, Yanase T, et al. Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, Anchorage, Alaska, USA, pp. 2623–2631. ACM.
38. Loshchilov I and Hutter F. Decoupled weight decay regularization. arXiv.1711.05101, 2017.
39. Folgoc L, Baltatzis V, Desai S, et al. Is MC dropout bayesian?. arXiv.2110.04286, 2021.
40. Hamilton W, Ying Z and Leskovec J. Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems* 30. Long Beach, California, USA , 2017, Curran Associates, Inc.