Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria

doi: 10.3850/ESREL2026061419_esrel26-p26222-cd

Advanced POD-Based Evaluation of Load-Dependent Performance in Machine Learning Damage Diagnosis

Jonathan Liebeton

Chair of Dynamics and Control, University of Duisburg-Essen, Germany.
E-mail: jonathan.liebeton@uni-due.de

Zahra Rastin

Chair of Dynamics and Control, University of Duisburg-Essen, Germany. E-mail: zahra.rastin@uni-due.de

Dirk Söffker

Chair of Dynamics and Control, University of Duisburg-Essen, Germany. E-mail: soeffker@uni-due.de

The reliable diagnosis of damage in carbon fiber reinforced polymer (CFRP) structures under variable loading conditions remains a key challenge for ensuring structural integrity and extending service life. Acoustic Emission (AE)-based machine learning (ML) methods have shown great potential, their performance often depends on loading conditions, which are rarely quantified systematically. This study aims to quantitatively evaluate how loading conditions affect the performance of ML-based damage diagnosis models. An advanced probability of detection (POD) framework is developed mapping the diagnosis performance onto newly-introduced two-dimensional load-dependent POD surfaces, incorporating both load amplitude and cyclic frequency as process parameters as well as the likelihood of correct diagnosis statements as output. Acoustic Emission signals generated during controlled cyclic loading of CFRP plates were measured, preprocessed, and used to train and test supervised ML models for damage detection and classification. The resulting detection probabilities were computed for varying load configurations, enabling a detailed visualization and analysis of load sensitivity. The results reveal distinct regions of reduced detection reliability at specific combinations of load magnitude and frequency, highlighting the necessity of considering operational loading when validating ML-based diagnostic systems. The proposed two-dimensional POD approach represents a new and systematic quantification of load-dependent performance in AE-driven ML damage diagnosis. This approach enables performance assessment under varying loading conditions and stimulates future research on load-adaptive diagnostic models for complex composite structures.

Keywords: Acoustic Emission, Machine Learning, Probability of Detection, Damage Diagnosis, Reliability.

1. Introduction and literature review

Carbon fiber reinforced plastics (CFRP) have become indispensable in modern engineering applications due to their high specific strength and stiffness, low mass, and superior resistance to environmental degradation. These attributes enable lightweight structural designs that improve energy efficiency and reduce life-cycle emissions, motivating extensive use of CFRP in aerospace, automotive, civil engineering, and energy-related systems. As the structural relevance of CFRP continues to grow, ensuring their long-term safety, reliability, and cost-effective operation has become a critical challenge.

Structural Health Monitoring (SHM) plays a central role in addressing this challenge by enabling continuous or periodic assessment of structural integrity throughout the service life of a component. In contrast to conventional inspection strategies, which are often interval-based and require system downtime, SHM aims to detect damage at an early stage and to support condition-based maintenance concepts. This is particularly important for CFRP structures, where damage such as matrix cracking, debonding, fiber breakage, or delamination may develop internally without producing visible surface indications Baccar and Söffker (2017). Without adequate monitoring strategies, such hidden damage can accumulate, compromise structural performance, and lead to sudden failure.

Among available non-destructive testing (NDT) techniques, Acoustic Emission (AE) monitoring is well suited for SHM of composite structures. Acoustic Emission is a passive sensing approach that captures transient elastic waves generated by active damage processes during loading. These waves propagate through the material and are measured by surface-mounted piezoelectric wafer active sensors (PWASs), allowing damage to be detected in real time without interrupting operation. Beyond simple event detection, AE provide temporal and spectral information to characterize the underlying damage mechanisms, making it a powerful diagnostic tool for in-situ monitoring of CFRP components.

Damage in CFRP manifests through multiple failure modes, including matrix cracking, debonding, delamination, and fiber breakage. These mechanisms are commonly associated with characteristic frequency ranges in the recorded AE signals and can therefore be differentiated using time–frequency or frequency-domain features Azadi et al. (2019); Baccar and Söffker (2017); Chelliah et al. (2019). Lower-frequency AE signals are typically linked to delamination processes, while matrix cracking and debonding occupy intermediate frequency bands. Fiber breakage, representing the most critical damage mode, generally produces high-frequency AE activity. Although the absolute frequency limits depend on material and testing conditions, the relative ordering of these mechanisms is consistently reported.

Accurate classification of damage types based on AE data is a key requirement for meaningful SHM. Recent studies demonstrate that machine learning methods, including deep learning models, can achieve high classification accuracy when distinguishing between isolated damage mechanisms Sai Krishna et al. (2024); Liu et al. (2023); Barile et al. (2022); Xue et al. (2025).

The performance of AE-based ML methods is strongly affected by loading conditions such as load amplitude and cyclic frequency, yet these effects are rarely quantified systematically (Liebeton and Söffker (2025)). Loading conditions influence AE signal generation and propagation, and neglecting their effects can lead to overly optimistic performance estimates. A systematic performance assessment under varying loading conditions is essential to ensure the reliability of AE-driven ML models. In this context, the probability of detection (POD) approach is well suited to assess ML performance (Bakhshande et al. (2022)), as it can account for variations in process parameters, which are task-specific factors like loading conditions, impacting the recognizability of the target and the outcome, independently of training/model-specific hyperparameters.

The POD approach is a statistical method for assessing the reliability of diagnostic systems by estimating the probability that a detection or prediction method will successfully achieve its intended objective. While POD was initially developed to validate nondestructive evaluation techniques, it has been more recently applied in other areas, including SHM (Falcetelli et al. (2022); Leung and Corcoran (2021); Yue and Aliabadi (2021)). The standard POD evaluation produces a so-called POD curve, which depicts the probability of successful target detection or prediction as a function of the considered process parameter.

The use of POD to assess the performance of ML models is still relatively uncommon. For example, Nerlikar et al. (2024) applied the hit/miss POD method to ML-based damage detection in composites, considering delamination size as the process parameter. Also, in Ameyaw et al. (2022) and Rastin and Söffker (2024) POD was employed to validate classifiers developed for predicting driver lane-changing behavior.

In this paper, an advanced POD framework is developed to enable the performance assessment of AE-driven ML classifiers under varying loading conditions. This approach maps diagnostic performance onto newly introduced two-dimensional (2D), load-dependent POD surfaces, incorporating both load amplitude and cyclic frequency as process parameters, with the likelihood of correct diagnosis serving as the output.

A one class support vector machine (1C-SVM) is used for the detection of AE events and multiclass support vector machines (SVMs) are used to differentiate between the different damage types. In addition to POD-based evaluation, the combined detection and classification results, considering all loading conditions together, are also reported to provide a more comprehensive assessment.

The advanced POD-based evaluation results show the specific combinations of load magnitude and frequency at which the ML models can reliably detect and classify different damage types in the CFRP plates. By explicitly linking model performance to load-dependent behavior, this approach enhances the understanding of AE-driven ML reliability, and motivates future research on load-adaptive diagnostic models for complex composite structures.

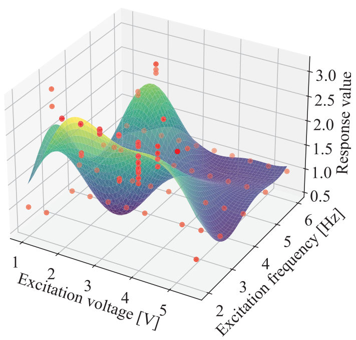
After the introduction, in section 2, the experimental setup is described, followed by a discussion of the signal processing and feature extraction methods, along with the machine learning-based classification approach. The proposed POD approach is also explained in this section. Machine learning evaluation results are presented in section 3, and the findings are discussed in section 4. Finally, the work is concluded with a summary and outlook in section 5.

2. Methods

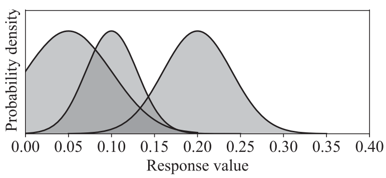
The experimental methodology follows a two-stage concept: controlled damage generation and subsequent fatigue-type loading. Investigations are conducted on CFRP laminates with 90°/0°/90° fiber orientation, where damage is introduced via mechanical indentation. This procedure, adapted from Wirtz et al. (2019), utilizes varying indentor configurations to induce global bending or localized matrix damage Baccar and Söffker (2017). Additional bending tests are conducted to further diversify the damage mechanisms within the dataset. Post-indentation, specimens undergo cyclic bending. Cyclic excitation is generated via a sinusoidal voltage signal, resulting in a nearly constant force Liebeton et al. (2023). Frequencies range from 2 to 6 Hz with voltage amplitudes of 2 to 6 V, inducing displacements of 6 to 10 mm. To account for stiffness reduction during progressive damage Wang and Zhang (2020), excitation is regulated via electrical input rather than displacement control. Acoustic Emission data are recorded from the pristine state to failure using surface-bonded PWASs. Signals are conditioned and digitized by an field programmable gate array (FPGA) at 4 MHz. Recorded signals are manually labeled as noise or damage based on spectral characteristics described in the literature. For processing, measurements are divided into 2^{10}-sample windows, treated with a half Hamming window, and characterized by continuous wavelet transform (CWT)-based features. Classification is performed using SVM-based models. Binary classifiers maximize margins between classes, while soft-margin formulations and kernel functions handle nonlinear boundaries. Multi-class tasks utilize one-versus-all or one-versus-one strategies. Decision values are transformed into posterior probabilities via sigmoid mapping Platt (1999); Chang and Lin (2011) and quadratic optimization Wu et al. (2004). Additionally, a one-class classification approach identifies anomalies based on deviations from a learned boundary. Two workflows are evaluated Liebeton and Söffker (2025): a direct five-class classifier and a two-stage process where a one-class classifier removes noise prior to multi-class classification. Model performance is validated using 10-fold cross-validation. Each damage category includes 390 labeled samples, while the noise class contains 1560 samples, partitioned for training and testing.

To validate ML model performance under varying load amplitude and frequency, a 2D-POD framework is developed, producing POD surfaces that describe the probability of correct classification as a function of excitation voltage and frequency. The ML models' output class probabilities are used to evaluate them within this framework. Test samples are fed into the trained models, which provide the probabilities of belonging to each class. The probability of correct classification for a sample is defined as the probability that the true class probability exceeds all others, that is, greater than the maximum of the remaining class probabilities. For six classes, this probability can be expressed as

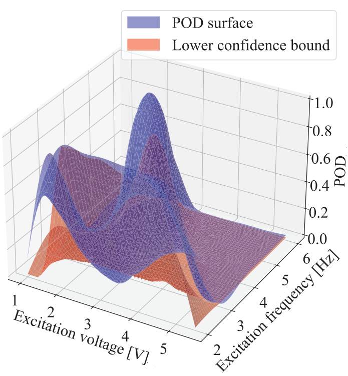
$$\mathrm{P_{cc}} = \mathrm{P}\Big(p_c - \max_{j\in\{1,\ldots,6\},\, j\neq c} p_j > 0\Big), \qquad (1)$$



(a)



(b)



(c)

Fig. 1. POD-based evaluation steps: (a) fitting a surface to the data, (b) estimating the probability that the response exceeds zero at each process parameter pair, and (c) generating the POD surface and its 95 % lower confidence bound.

where p_c denotes the predicted probability of the true class and p_j denotes the predicted probability of class j. The quantity

$$\hat{a} = p_c - \max_{j \neq c} p_j \tag{2}$$

is referred to as the response signal in the POD terminology. Consequently, the probability of correct and reliable classification is defined as the probability that the response signal is positive.

The main steps of the POD-based evaluation in this study are illustrated in figure 1. First, a model is needed to capture the relationship between the two process parameters and the response. Gaussian process regression (GPR) is used to approximate the response surface, as the response lacks a clear monotonic trend (figure 1a). Using the original (untransformed) variable scales, the model is formulated as

$$\hat{a} = \mathbb{E}[\hat{a}] = \mathrm{f}(a_1, a_2) + \varepsilon_n, \tag{3}$$

where $\mathrm{f}(\cdot)$ is an unknown latent function modeled as a Gaussian process, and $\varepsilon_n \sim \mathcal{N}(0, \sigma_n^2)$ represents additive Gaussian noise. For any point (a_1, a_2), the Gaussian process provides a predictive distribution for the function value, which can be written as

$$\mathrm{f}(a_1, a_2) \sim \mathcal{N}\Big(\hat{\mathrm{f}}(a_1, a_2),\ \sigma_f^2(a_1, a_2)\Big), \tag{4}$$

where $\mathrm{f}(a_1, a_2)$ is the predicted mean, and $\sigma_f^2(a_1, a_2)$ quantifies the prediction uncertainty due to the limited or sparse training data. This variance grows in regions farther from the training points, reflecting higher uncertainty. The final error for each predicted observation combines the Gaussian process predictive uncertainty and Gaussian noise, resulting in a normal distribution with mean equal to the predicted value and standard deviation equal to

$$\sigma_{\text{total}} = \sqrt{\sigma_f^2(a_1^*, a_2^*) + \sigma_n^2}. \tag{5}$$

Four different combinations of Cartesian and logarithmic scales for the independent variables a_1, a_2 and the dependent variable $\hat{a}$ are tested when fitting the GPR models. The model with the highest in-sample R^2 is selected to best characterize the response and its uncertainty at the observed parameter combinations, rather than to optimize prediction at unseen points. Because of repeated measurements and explicit modeling of measurement noise, the predictive uncertainty does not vanish at the training points.

The probability of correctly detecting the damage class at a given voltage-frequency pair (Eq. 1), is the probability that the predicted response exceeds the decision threshold (here zero). Considering the total GPR prediction error $\varepsilon_{total} \sim \mathcal{N}(0, \sigma^2_{total})$, this probability is equal to the area under the probability density function (PDF) of ε_{total}, centered at the GPR-predicted mean, and above the threshold. This area is illustrated by the shaded regions in figure 1b for response values of 0.05, 0.10, and 0.20. The variables are expressed on Cartesian scales, so the probability of detecting the damage class can be given by

$$\mathrm{POD}(a_1, a_2) = \mathrm{P}(\hat{a} > 0) = 1 - \Phi\left(\frac{0 - (\mathbb{E}[\hat{a}])}{\sigma_{total}}\right), \quad (6)$$

where Φ represents the cumulative standard normal distribution function. Although GPR provides point-wise uncertainty, a 95 $\%$ lower confidence bound of the POD surface is computed using the Wald method to quantify statistical uncertainty. This bound ensures that the true detection probability is at least the reported value with 95 $\%$ confidence. A POD surface and its 95 $\%$ lower confidence bound are shown in figure 1c. The bound can be presented as a heatmap over the voltage–frequency plane, which highlights regions of reliable classification and areas where detection performance declines.

3. Results

The performance of the proposed classification strategies is assessed based on their ability to jointly detect and classify AE events using CWT–based features.

3.1. *Detection and Classification results*

In the direct classification approach, a five class support vector machine (5C-SVM) is trained to distinguish noise from four damage mechanisms; detailed results using CWT-based features are presented in Liebeton and Söffker (2025) and summarized in figure 2. This strategy identifies all noise-related signals and achieves an overall accuracy of 98.8 $\%$, with individual class accuracies reaching up to 98.5 $\%$.

The two-stage framework—utilizing an initial one-class detection step followed by a four-class classification—is evaluated in figure 3. This approach achieves perfect detection of all damage events and an improved combined accuracy of 99.3 $\%$. These findings, further detailed in Liebeton and Söffker (2025), indicate that decoupling noise suppression from damage classification enhances the robustness and performance of CWT-based feature sets.

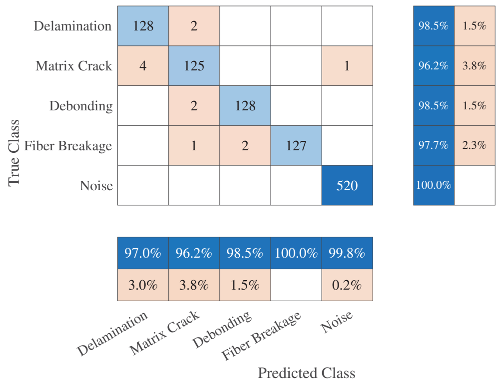


Fig. 2. 5C-SVM direct classification results using CWT-based features Liebeton and Söffker (2025)

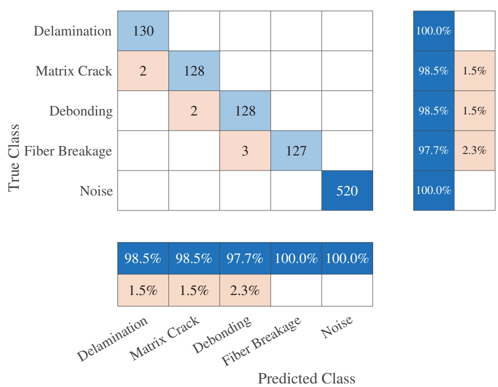


Fig. 3. Combined 1C-SVM detection and 4C-SVM classification results using CWT-based features Liebeton and Söffker (2025)

Overall, both evaluated strategies successfully distinguish damage-related AE signals from noise when using CWT-derived features. In the direct classification approach, only a single damage event remains undetected, whereas the two-stage framework achieves perfect detection. The improved performance of the latter approach highlights the benefit of decoupling noise suppression from damage classification. The strong results obtained with CWT-based features confirm their suitability for characterizing highly transient AE signals and support their use in both direct and staged classification architectures.

3.2. *POD-based evaluation results*

The POD approach described in section 2 is applied to evaluate the ability of the ML models to correctly detect and classify each damage type under varying loading conditions. Due to the limited number of available test samples for each damage class, the GPR model does not generalize well to unseen data. Nevertheless, by incorporating the noise level inferred from repeated measurements at identical process parameters together with the model uncertainty, the predicted means and standard deviations are expected to approximate the true values at the training points without overfitting the training data. Only the total GPR prediction error at the training points is used for the POD calculation. A cubic scattered data interpolation is applied to the sparse POD values to obtain smooth POD surfaces.

The resulting lower confidence bound POD surfaces for the 5C-SVM classification model and the combined 1C-SVM detection with 4C-SVM classification models are shown in figures 4 and 5 as heatmaps over the voltage–frequency plane. For brevity, only the results obtained using CWT-based features are presented, as these features yield better detection and classification performance, as shown in section 3.1. The POD values are illustrated in figures 4 and 5 throughout the parameter space, enabling direct identification of excitation voltage–excitation frequency combinations that yield reliable correct classification, as well as regions where detection performance deteriorates. In standard POD practice, a probability of detection above 90 % is typically considered indicative of reliable performance. The regions with POD values exceeding 90 % are shown in yellow in the presented heatmaps. Similar POD patterns can be observed when using the 5C-SVM direct classification model and the combined 1C-SVM detection with 4C-SVM classification models. In both cases, reliable detection and classification of delamination and fiber breakage are achieved over a wider range of excitation voltage–frequency combinations. The region of reliable damage identification is slightly smaller for matrix crack damage, while for debonding it is limited to a narrow region around excitation voltages of 2-4 V and an excitation frequency of approximately 2 Hz.

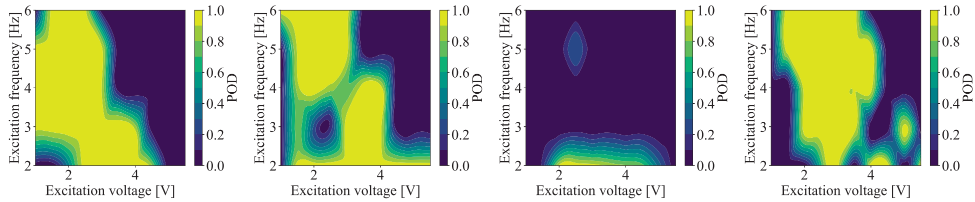
4. Discussion

From the presented results (figures 2 and 3), it can be observed that validating the performance of the ML models solely using the confusion matrix and the associated performance metrics, without accounting for the loading conditions, suggests an overall very good model performance. However, such an evaluation can be misleading. The proposed POD-based evaluation clearly demonstrates that this good performance is achieved only under specific loading conditions (figures 4 and 5).

Unlike conventional classification metrics, the POD framework provides a probabilistic assessment of detection capability while explicitly accounting for uncertainties arising from loading variability, measurement noise, and model predictions. This enables a more realistic and reliability-oriented evaluation of the ML models. The POD-based approach allows the identification of operational regions in which the models perform reliably, as well as conditions under which their reliability significantly degrades. Consequently, POD-based evaluation offers a more comprehensive and physically meaningful validation framework, particularly for safety-critical applications, where understanding the limits of model applicability and quantifying detection confidence are essential.

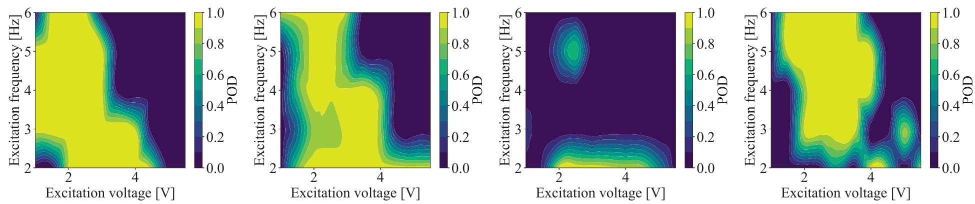
5. Summary and Outlook

The investigation into AE signal processing confirms that CWT-based features provide a highly effective foundation for both direct and two-stage classification strategies. While a direct 5-class SVM approach performs admirably, the decoupled two-stage framework, separating noise suppression from damage classification, reaches a superior detection rate and an overall accuracy



(a) Delamination (b) Matrix crack (c) Debonding (d) Fiber breakage

Fig. 4. Heatmaps of the lower confidence bound POD surfaces when using the 5C-SVM direct classification model with CWT-based features.



(a) Delamination (b) Matrix crack (c) Debonding (d) Fiber breakage

Fig. 5. Heatmaps of the lower confidence bound POD surfaces when using 1C-SVM detection and 4C-SVM classification models with CWT-based features.

of 99.3 %. To move beyond static performance metrics, a POD evaluation was implemented using Gaussian Process Regression to map model reliability across a range of excitation voltages and frequencies. This analysis reveals that reliable classification (POD $>$ 90 %) for delamination and fiber breakage occurs over a broad operational envelope, whereas the window for identifying debonding is significantly more constrained. The resulting POD heatmaps serve as a critical diagnostic tool, pinpointing the specific loading conditions where the machine learning models maintain their integrity and where they begin to falter. This probabilistic validation framework demonstrates that traditional accuracy scores can be misleading, as they mask performance sensitivities to physical loading variability. Consequently, incorporating POD into the validation pipeline offers a more rigorous and physically meaningful assessment of model reliability for safety-critical structural health monitoring.

Future work will focus on further improving diagnostic reliability using decision fusion considering POD performance of the individual approaches.

References

Ameyaw, D. A., Q. Deng, and D. Söffker (2022). How to evaluate classifier performance in the presence of additional effects: A new pod-based approach allowing certification of machine learning approaches. *Machine Learning with Applications 7*, 100220.

Azadi, M., H. Sayar, A. Ghasemi-Ghalebahman, and S. M. Jafari (2019). Tensile loading rate effect on mechanical properties and failure mechanisms in open-hole carbon fiber reinforced polymer composites by acoustic emission approach. *Composites Part B: Engineering 158*, 448–458.

Baccar, D. and D. Söffker (2017). Identification and classification of failure modes in laminated composites by using a multivariate statistical analysis of wavelet coefficients. *Mechanical Systems and Signal Processing 96*, 77–87.

Bakhshande, F., D. A. Ameyaw, N. Madan, and D. Söffker (2022). New metric for evaluation of deep neural network applied in vision-based systems. *Applied Sciences 12*(7).

Barile, C., C. Casavola, G. Pappalettera, and V. Paramsamy Kannan (2022). Damage monitoring of carbon fibre reinforced polymer composites using acoustic emission technique and deep learning. *Composite Structures 292*, 115629.

Chang, C.-C. and C.-J. Lin (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*(3), 1–27.

Chelliah, S. K., P. Parameswaran, S. Ramasamy, A. Vellayaraj, and S. Subramanian (2019). Optimization of acoustic emission parameters to discriminate failure modes in glass–epoxy composite laminates using pattern recognition. *Structural Health Monitoring 18*(4), 1253–1267.

Falcetelli, F., N. Yue, R. D. Sante, and D. Zarouchas (2022). Probability of detection, localization, and sizing: The evolution of reliability metrics in structural health monitoring. *Structural Health Monitoring 21*(6), 2990–3017.

Leung, M. S. H. and J. Corcoran (2021). Evaluating the probability of detection capability of permanently installed sensors using a structural integrity informed approach. *Journal of Nondestructive Evaluation 40*(3), 82.

Liebeton, J. and D. Söffker (2025). Damage detection in cfrp using acoustic emission. *Proceedings of the XX International Symposium on Dynamic Problems of Mechanics*.

Liebeton, J., S. M. S. Soon, and D. Söffker (2023). Improving the diagnostic reliability of ae-based failure mode detection and distinction in cfrp. In M. P. Brito, T. Aven, P. Baraldi, M. Čepin, and E. Zio (Eds.), *Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)*, pp. 1298–1304. Research Publishing, Singapore.

Liebeton, J. and D. Söffker (2025). Detectability of damages in carbon fiber reinforced plastics using acoustic emission. *Annual Conference of the PHM Society 17*(1).

Liu, Y., K. Huang, Z.-x. Wang, Z. Li, L. Chen, Q. Shi, S. Yu, Z. Li, L. Zhang, and L. Guo (2023). Cross-scale data-based damage identification of cfrp laminates using acoustic emission and deep learning. *Engineering Fracture Mechanics 294*, 109724.

Nerlikar, V., O. Mesnil, R. Miorelli, and O. D'Almeida (2024). Damage detection with ultrasonic guided waves using machine learning and aggregated baselines. *Structural Health Monitoring 23*(1), 443–462.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers 10*(3), 61–74.

Rastin, Z. and D. Söffker (2024). Multi-level feature extraction and classification for lane changing behavior prediction and pod-based evaluation. *Automation 5*(3), 310–323.

Sai Krishna, Y., G. Raju, and M. S. Desarkar (2024). Damage mode classification in cfrp laminates using convolutional autoencoder and convolutional neural network on acoustic emission waveforms. *Structural Health Monitoring*.

Wang, C. and J. Zhang (2020). Experimental and analytical study on residual stiffness/strength of cfrp tendons under cyclic loading. *Materials 13*(24), 5653.

Wirtz, S. F., N. Beganovic, and D. Söffker (2019). Investigation of damage detectability in composites using frequency-based classification of acoustic emission measurements. *Structural Health Monitoring 18*(4), 1207–1218.

Wu, T.-f., C.-J. Lin, and R. Weng (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research 5*, 975–1005.

Xue, S.-N., J. Wang, Y.-Z. Liang, L.-H. Ma, and W. Zhou (2025). Indentation damage identification of carbon fiber composite laminates based on modal acoustic emission and machine learning. *Polymer Composites 46*(8), 6944–6955.

Yue, N. and M. Aliabadi (2021). Hierarchical approach for uncertainty quantification and reliability assessment of guided wave-based structural health monitoring. *Structural Health Monitoring 20*(5), 2274–2299.