

Fair Evaluation of Classical Machine Learning Methods for Vibration-Based SHM

M. Sc. Friedrich Hennig

Hochschulring 1, D-15745 Wildau, Germany
friedrich.hennig@th-wildau.de

M. Sc. Hao Cheng

Hochschulring 1, D-15745 Wildau, Germany
hao.cheng@th-wildau.de

Prof. Dr. Rainer Stollhoff

Hochschulring 1, D-15745 Wildau, Germany
rainer.stollhoff@th-wildau.de

Prof. Dr. Dina Hannebauer

Hochschulring 1, D-15745 Wildau, Germany
dina.hannebauer@th-wildau.de

Abstract: Glass Fiber Reinforced Plastic (GFRP) is widely used in industrial applications, from wind turbine blades to lightweight drones and aircraft structures. Its non-linear, anisotropic behavior makes damage detection harder than for metallic materials, while damage types such as delamination, matrix cracking, and fiber breakage substantially affect structural durability. Structural Health Monitoring (SHM) aims to replace scheduled manual inspections with automated, data-driven damage detection.

This work presents a vibration-based dataset of 32 GFRP plates (8 undamaged and 24 damaged across four damage levels) in two aging states, together with a fair benchmarking framework for classical machine learning. Six models are evaluated across five feature sets using repeated nested cross-validation (4 outer folds $\times$ 30 repeats), Optuna-based hyperparameter optimization, and Friedman-Nemenyi statistical comparison. A trivial baseline that always predicts the majority class is included to expose the optimistic bias of accuracy and F1 under class imbalance.

The benchmark shows that rigorous nested evaluation reduces the reported balanced accuracy from 88% to 69.5% on identical data, that feature choice has a stronger effect than model selection, and that accuracy and F1 are misleading under this class imbalance. The trivial baseline reaches 75% accuracy and 86% F1, while only balanced accuracy reveals actual learning.

1 Introduction

Composite materials are widely used in engineering structures such as aircraft components and wind turbine blades. However, due to their complex internal structure and the variety of possible damage states, they still require frequent manual inspections [1]. Structural Health Monitoring (SHM) aims to address this limitation by using sensors and algorithms to automatically detect damage within such structures [2]. In this paper, we focus on vibration-based measurements of Glass Fiber Reinforced Plastic (GFRP) plates as a model system.

A common challenge in experimental SHM studies is that only small and imbalanced datasets are available [3, 4]. This complicates the implementation of a robust evaluation methodology for distinguishing damaged from undamaged states and requires careful methodological design. Using an inadequate evaluation methodology or an unsuitable metric can lead to misleading conclusions or optimistic bias in the reported results [5].

This paper does not propose a novel algorithm for SHM discrimination of damaged and undamaged states. Instead, it documents the design decisions made when constructing a robust evaluation framework for a small dataset of vibration-based measurements of GFRP plates. The remainder of the paper is structured as follows: we first describe the data acquisition process and the vibration measurements performed on the plates, then present the methodology of the machine learning pipeline, and finally evaluate and discuss the results. This ordering reflects that several methodological choices in the evaluation framework are direct consequences of the dataset's size and class structure.

2 Data Acquisition

2.1 Sample production

The specimens utilized in this study were Glass Fiber Reinforced Plastic (GFRP) plates, representing materials found in wind turbine blades and fuselages of

lightweight drones. To ease the manufacturing process and state-of-the-art analysis, a simplified flat-plate geometry was adopted.

Initial samples were produced using a manual hand lay-up process on a glass base (process see Figure 1). While functional for initial testing, this method proved labor-intensive and resulted in inherent manufacturing defects, including entrapped air (voids), surface delamination, and thickness inconsistencies. Only 32 specimens produced achieved the quality necessary for the following FRF measurements.

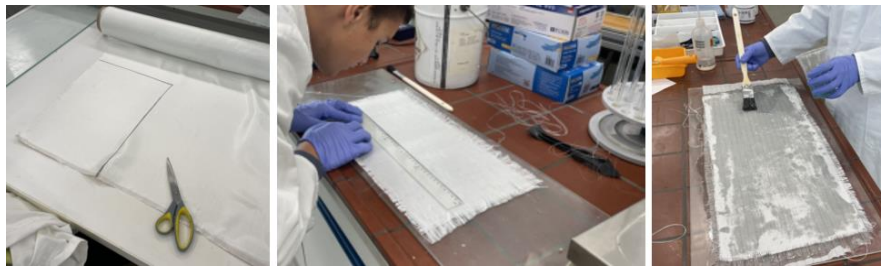


Figure 1
Manual lamination process. Left: cutting the glass fiber fabric.
Middle: Alignment of the fabric layers on a glass base plate.
Right: Application of resin by brushing

To mitigate the defects and labor intensity of manual lamination, a Vacuum Infusion Process (VIP), where the liquid resin is pulled evenly through the fiber layers using vacuum (Figure 2), was adopted for the second production phase. This method utilizes a vacuum-driven pressure differential to ensure a consistent, pore-free resin impregnation, enabling the high-volume production of over 100 specimens. However, due to the extensive volume of data generated, these newer samples have not yet been fully analyzed. Consequently, the Machine Learning analysis presented in the following section pertains exclusively to the initial 32 specimens produced via manual lamination.



Figure 2
Vacuum infusion process

2.2 Variation of the Sample Plates: Weathering and Damage

To create the diverse material states required for this study, specimens were subjected to two distinct degradation methods, applied either independently or in combination:

- Environmental Weathering (Figure 3): Accelerated aging in a climate chamber (high temperature, humidity, and UV) to simulate long-term atmospheric exposure.
- Mechanical Impact (Figure 4): Controlled damage via a drop-weight machine (dropping of steel balls with predefined energy) to induce failures such as delamination, fiber breakage, and matrix cracking.



Figure 3
Aging of the sample plates in the climate chamber

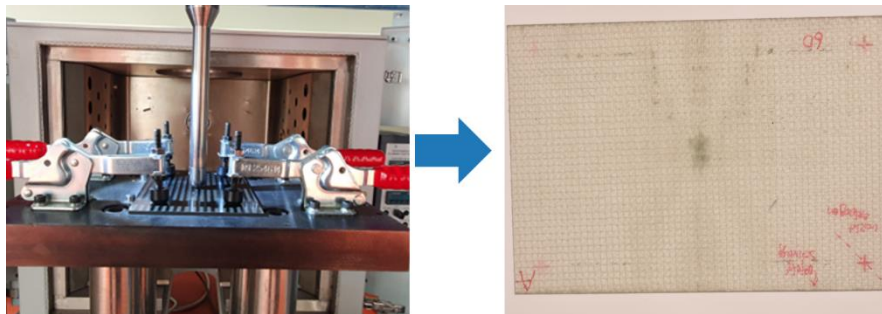


Figure 4
Damaging of the sample plates by a drop-weight test machine

The resulting experimental matrix consists of 8 variations from the 32 samples (4 samples per variation, Figure 5).

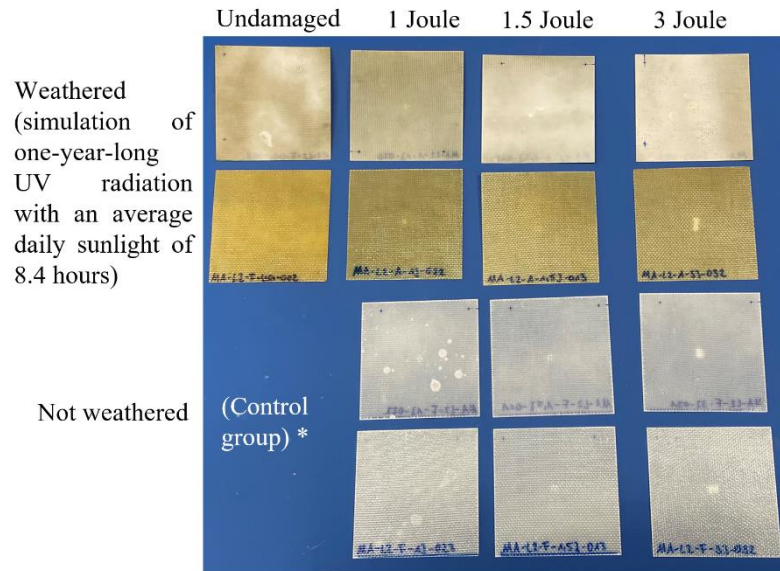


Figure 5
Specimen variations. Undamaged, not weathered control samples consumed during destructive material characterization

2.3 Vibration Measurement

The specimens were measured using a Scanning Laser Doppler Vibrometer (SLDV) and an automated modal hammer. The experimental setup employed a "fixed-excitation, roving-response" configuration. While an automated hammer provided consistent impulse excitation at a single corner point, the laser scanned over various points (measurement grid) across the specimen. Both the impulse force and velocity responses were processed via Fourier Transform to generate Frequency Response Functions (FRFs) for every measurement point (Figure 6 left).

The peaks within these FRFs identify the material's resonance frequencies. By correlating data across all scan points, the mode shapes (physical deformation patterns) were reconstructed (Figure 6 right). These range from simple bending at lower frequencies to complex twisting patterns at higher frequencies. Together, these resonance frequencies and mode shapes form a unique "vibration fingerprint," serving as the primary feature set for the subsequent Machine Learning algorithms.

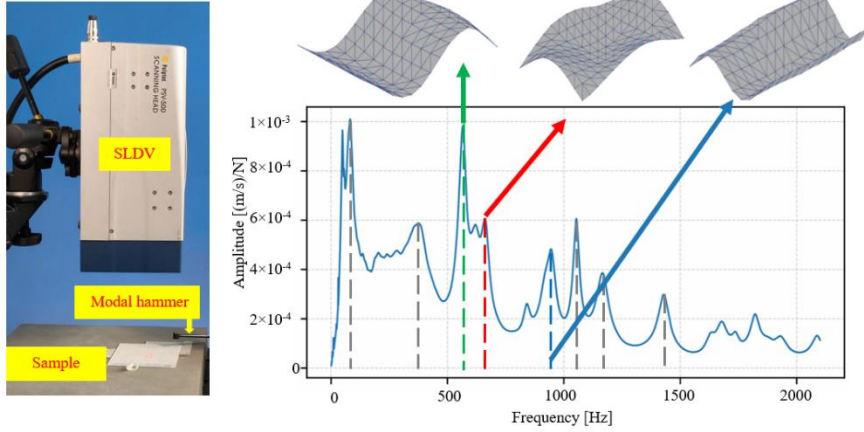


Figure 6
Measurement setup (left) and the output of the measurements (right)

3 Benchmarking Methodology

For the machine learning benchmark, the experimental conditions were collapsed into a binary classification task. Plates without mechanical impact damage were labeled as undamaged, whereas all impact-damaged plates were labeled as damaged, independent of the weathering state. This resulted in an imbalanced dataset of 8 undamaged and 24 damaged specimens. To compare and evaluate predictive performance, a range of classical machine learning models was applied to several feature sets extracted from the measured Frequency Response Functions (FRFs). One limitation is that, while many measurements are available per plate, the small number of plates overall constrains the dimensionality of the feature sets. It should also be noted that only frequencies up to 2000 Hz were used, due to constraints arising from the planned use of simulated plates. Although this discards some measurement information, we observed only a marginal decrease in the reported evaluation.

3.1 Feature sets

Five different feature sets were extracted, each from its own domain to capture complementary information. The first feature set was inspired by the standard manual inspection of the plates, in which modal properties are analyzed. For feature set A, we extracted the first six modes and computed their frequency, amplitude,

damping, and Cross-MAC values (mean, standard deviation, and maximum) [6]. We used the SMAC algorithm [7], which considers all 121 measurement positions of the plate. Feature set B describes the spectral domain of the measurements through properties such as spectral centroid, spread, energy, bandwidth, roll-off, and flatness [8]. The third set describes statistical properties of subbands of the spectrum: the primary-channel measurements were subdivided into four bands (0–500 Hz, 500–1000 Hz, 1000–1500 Hz, and 1500–2000 Hz), and from each band we extracted the mean, standard deviation, skewness, and kurtosis; the same statistics, together with the Shannon entropy of the magnitude spectrum, were also computed over the full 0–2000 Hz range. A combined feature set was additionally evaluated to assess potential synergy effects between feature sets A, B, and C. The final set was based on a property typically applied in the time domain in other studies [9]: Wavelet Packet Energy. In our case, the same principle was applied to a frequency-domain wavelet packet decomposition (Daubechies-5, level 5), yielding 32 normalized subband energies complemented by their Shannon entropy, the log total energy, and the maximum subband share.

Set	Description	Dim
A	Modal	21
B	Spectral	6
C	Statistical	21
G	Combined: A + B + C	48
WPE	Wavelet packet energy (db5, level 5) (Frequency-Domain)	35

Figure 7

Overview of the five evaluated feature sets and their dimensionality.

3.2 Model selection

On the model side, a range of classical machine learning methods was evaluated. The methods include robust implementations such as Support Vector Machines with different kernels, Random Forest, and eXtreme Gradient Boosting ensembles, as well as simpler models such as Logistic Regression and k-Nearest Neighbors [10, 11]. We additionally included a Multilayer Perceptron (a small neural network) restricted to a limited number of neurons and at most one or two hidden layers. This restriction was motivated by the small number of observations in the dataset and is also the reason why no deep learning methods were considered.

The aim of the evaluation was to determine whether model choice has a substantial impact on the results, or whether simple models perform comparably given the selected features.

Model	Type
Support Vector Machine	Kernel-based (RBF, linear, poly)
Random Forest	Bagging ensemble
eXtreme Gradient Boosting	Boosting ensemble
k-Nearest-Neighbors	Distance-based
Logistic Regression	Linear baseline
Multilayer Perceptron	Neural network (1–2 layers)

Figure 8
Classical machine learning models evaluated in the benchmark.

3.3 Nested cross-validation

From the start, hyperparameter optimization was planned to adjust model parameters based on observed performance. However, this can introduce an optimistic bias if the same observations are used both to tune the parameters and to score performance. If hyperparameters are tuned on the same data on which the model is evaluated, the optimizer effectively selects the configuration that best fits the noise of that specific split, and the reported score reflects this overfit rather than true performance. It is therefore imperative that optimization is kept separate from evaluation [5].

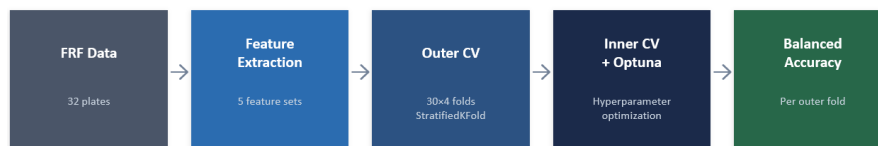


Figure 9
Nested cross-validation pipeline with outer evaluation and inner Optuna-based hyperparameter tuning.

For this reason, we used two cross-validation loops: a 4-fold outer loop and a 6-fold inner loop, both implemented as StratifiedKFold to preserve the class distribution after splitting. The outer loop was used solely for evaluating each model–feature-set combination and was repeated over 30 different random seeds, yielding 120 evaluation results per combination. The repetition was intended to mitigate the

influence of favorable splits. The training portion of each outer split was then passed to the inner loop.

Hyperparameter optimization took place in the inner loop, separately from the outer loop. Optimization was performed using Optuna with its built-in Tree-structured Parzen Estimator [12, 13]. The data passed to the inner loop was again split via cross-validation, with the training set used to fit the model and the inner test set used to tune its hyperparameters. After tuning, the selected hyperparameters were returned to the outer loop, where the model was trained on the outer training portion and evaluated on the outer test set. For the fair evaluation, only the outer test set results are reported. The gap between inner and outer test scores nevertheless serves as an indicator of overfitting.

Within each model–feature-set evaluation, some pipeline steps were fixed while others were tuned by Optuna. The pipeline was structured as follows: Imputation → StandardScaler → Feature Selection (none / ANOVA / PCA) → optional SMOTE [14] → Classifier. Imputation and scaling were fixed, whereas Optuna decided whether feature selection or oversampling improves performance, and additionally tuned the parameters and the algorithm of the feature selection step itself, only within the inner loop.

3.4 Evaluation metrics

The primary evaluation metric was balanced accuracy. Accuracy, F1-score, and MCC were tracked only for additional examination. Balanced accuracy was chosen because of the imbalanced class distribution: it weights performance according to class membership rather than overall sample frequency [15]. In addition, we report a trivial baseline that always predicts the majority class, which in our case is damaged. This trivial baseline reaches 75% accuracy and 86% F1, but only 50% balanced accuracy.

3.5 Statistical comparison

Rather than relying solely on the mean balanced accuracy, we additionally applied rank-based statistical tests for comparison. We performed two Friedman tests, one across the selected models and one across the feature sets. For the model comparison, scores were averaged across all feature sets, and vice versa for the feature-set comparison. If statistical significance was found, a Nemenyi post-hoc test was performed [16]. The Friedman test should, however, not be treated as a definitive answer, since the results are not independent across repeats. It is therefore used only as an indicator of the relative importance of model and feature-set choice, while the Nemenyi post-hoc indicates whether a clear winner exists within each category. The reported critical differences should therefore be interpreted as indicative rather than as strict inferential bounds.

4 Results

4.1 Main benchmark

The evaluation reveals three distinct findings. First, the best overall model–feature combination, measured by mean balanced accuracy across the 120 outer folds, is k-Nearest Neighbors with feature set G, closely followed by the same model with the statistical feature set C. Second, on the model side there is no substantial difference between the six classifiers, while on the feature side, sets C and G (which includes C as a subset) perform consistently better than the remaining ones. Third, feature set WPE performs noticeably worse than the other feature sets.

The best combination reaches 69.5% balanced accuracy, indicating that the models do learn information about the plate condition, although performance remains far from perfect. Clustered bootstrap ($n = 5000$, clustered by plate) yields 95% confidence intervals of $[0.58, 0.80]$ for the best combination (kNN + G) and $[0.58, 0.78]$ for the second-best (kNN + C). The intervals overlap substantially and reach close to chance level at the lower bound, indicating that the reported point estimates carry considerable variance under the small dataset size.

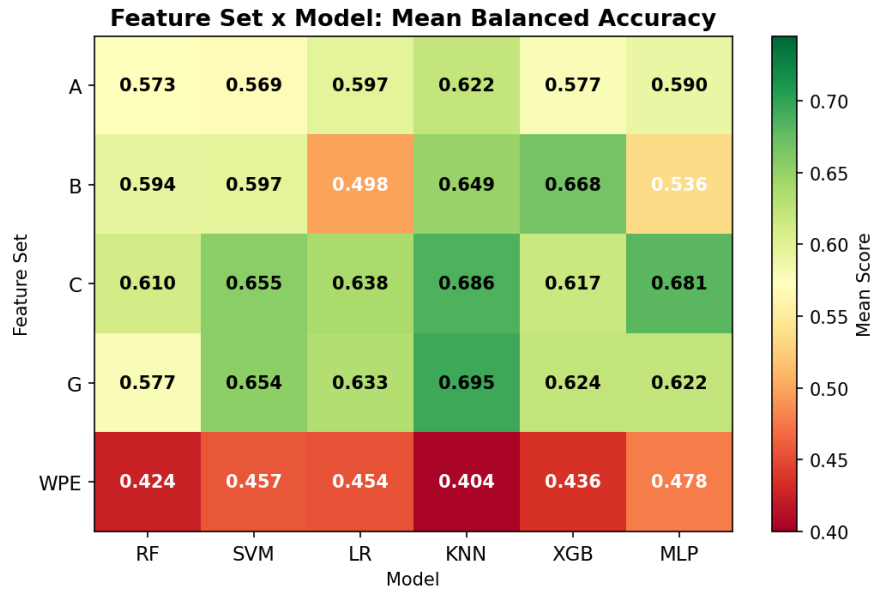


Figure 10
Mean balanced accuracy across all model–feature-set combinations over 120 outer folds.

4.2 Comparison of evaluation metrics

The comparison between our results and the trivial baseline illustrates why the choice of metric is critical. Any trained model that fails to outperform this baseline on balanced accuracy has not learned anything beyond the class prior. When the same result is reported in terms of accuracy, the best combination reaches 68%, compared to 75% for the baseline. A similar discrepancy appears for the F1-score, with the baseline reaching 86% and the trained classifier only 76%. Based on these two metrics alone, one might conclude that the trained classifier performs worse than the baseline, yet the baseline, by construction, never detects any undamaged plate. Under the imbalanced class distribution, the trained model is therefore the more informative predictor: it does learn the damage-related information in the plates, even though accuracy and F1-score fail to reflect this. Balanced accuracy makes the difference visible, with the baseline reaching 50% and the trained classifier 69.5%.

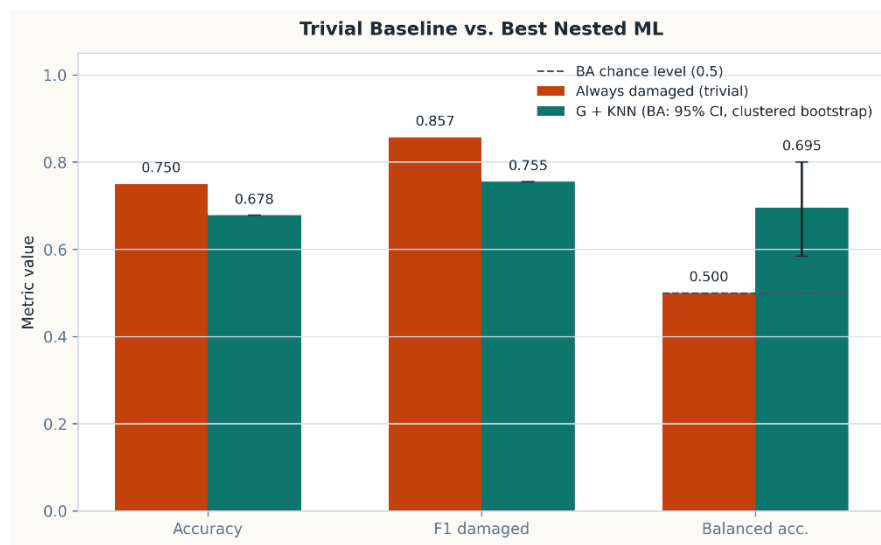


Figure 11

Trivial baseline compared to the best nested model (kNN + feature set G).

A second source of optimistic bias arises from the evaluation protocol itself. If we report the score that hyperparameter optimization itself produced, i.e. the inner-loop best score, as one would naively do without nesting for an outer evaluation loop, we obtain approximately 88% balanced accuracy on the same data. With the outer test set result at 69.5%, this corresponds to a gap of 18 percentage points. Without nesting for an outer evaluation loop, the reported performance is therefore overly optimistic and likely not achievable on unseen data.

4.3 Statistical significance

The statistical tests confirm the patterns observed in the heatmap. The Friedman test indicates clearer rank differences on the feature side, with $\chi^2 = 177.8$, than on the model side, with $\chi^2 = 33.7$. The pairwise post-hoc comparison shows that, from a model perspective, no single classifier is clearly superior to the others: kNN ranks consistently above Logistic Regression and Random Forest, exceeding the critical difference of 0.688. On the feature side, sets C and G consistently outrank sets A, B, and WPE, exceeding the critical difference of 0.557.

Overall Model Comparison		
Friedman $\chi^2 = 33.7$, $p = 2.7 \times 10^{-6}$		
Rank	Model	Mean Rank
1	KNN	2.85
2-3	SVM	3.38
2-3	XGB	3.38
4	MLP	3.42
5	LR	3.87
6	RF	4.10

CD = 0.688 | Significant: KNN > {LR, RF}, SVM/XGB > RF. Most pairs NOT significant.

Overall Feature Set Comparison		
Friedman $\chi^2 = 177.8$, $p = 2.2 \times 10^{-37}$		
Rank	Feature Set	Mean Rank
1	C (Statistical)	2.20
2	G (Combined)	2.31
3	B (Spectral)	2.94
4	A (Modal)	2.96
5	WPE (Wavelet)	4.59

CD = 0.557 | C = G (not significant).
C/G significantly better than A, B, WPE.
WPE significantly worst across all models.

Figure 12

Friedman ranks and Nemenyi post-hoc results for models (left) and feature sets (right).

5 Discussion

The aim of this study is not to present a novel approach to structural health monitoring, but to demonstrate the methodology behind a fair model and feature comparison and how the choices it involves, such as the choice of metrics, have wide-ranging effects on the results. Omitting nested cross-validation for an outer evaluation loop in combination with hyperparameter optimization yielded an inflated balanced accuracy of 88%, while applying the nested protocol produced lower but fairer results [5]. It should be noted that omitting this step does not necessarily imply that the result is wrong; rather, a sound evaluation framework is required to enable a fair comparison between feature sets and models and to provide a more reliable estimate of how the same combination would perform on unseen data. This is particularly relevant for structural health monitoring, where the response to damage can vary widely and meaningful gaps between models may exist.

The metrics in particular reveal differences and should be interpreted accordingly. Although our trained model achieved less than the trivial baseline in terms of accuracy and F1-score, balanced accuracy demonstrated that the model did in fact

learn damage-related information from the plates. This can be traced back to the imbalance in our class distribution, which influences both key design decisions in implementing the framework and the interpretability of the reported results [17]. A study could plausibly report an F1-score of 88%, but without a description of the underlying methodology and an accompanying baseline or metric explanation, the interpretation would be drastically different. The statistical tests also indicate that, in our dataset, feature choice matters more than model choice. One possible explanation is that there are simply too few observations, and that the models would require more (or more advanced) engineered features to better capture the spatial distribution of the measurements.

The most limiting factor of this study is the small dataset size, which does not permit higher-dimensional feature sets. With only 32 plates measured and a class imbalance of 1:3, and with the framework having to set aside samples for independent testing, the number of undamaged plates available for hyperparameter optimization is reduced even further. In addition, the manufactured plates likely contained errors introduced during the manual production process, which may have influenced the observed results. A further limitation is that environmental data such as humidity and temperature were not tracked during the measurements, even though these are known to affect frequency response functions [18]. These properties could also act as confounders within the dataset, which we were unable to either confirm or rule out. Given these limitations, the results obtained should be regarded as within-dataset findings, while the methodology behind the evaluation remains relevant, particularly for small, imbalanced datasets. While the present results are dataset-bound, the evaluation framework itself transfers directly to larger datasets.

Conclusion

In summary, we firstly created a dataset of glass fiber reinforced plastic plates that were damaged, weathered, and characterized through their vibration response to impulse excitation. Secondly, we used this dataset to compare different machine learning models for structural health monitoring and implemented a rigorous evaluation framework for comparing different model and feature set combinations. The results show that the methodology behind the evaluation has a substantial impact on the conclusions that can be drawn. In our case, a direct evaluation of the trained models would have inflated the reported result by approximately 18 percentage points. The choice of metric is also particularly important when working with small, imbalanced datasets, where the trivial baseline outperformed our best combination in both accuracy and F1-score, while being clearly outperformed in terms of balanced accuracy.

References

- [1] Gholizadeh S. A review of non-destructive testing methods of composite materials. *Procedia Structural Integrity* 2016; 1: 50–7 [https://doi.org/10.1016/j.prostr.2016.02.008]
- [2] Farrar CR, Worden K. An introduction to structural health monitoring. *Philos Trans A Math Phys Eng Sci* 2007; 365(1851): 303–15 [https://doi.org/10.1098/rsta.2006.1928][PMID: 17255041]
- [3] Avci O, Abdeljaber O, Kiranyaz S, Hussein M, Gabbouj M, Inman DJ. A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. *Mechanical Systems and Signal Processing* 2021; 147: 107077 [https://doi.org/10.1016/j.ymssp.2020.107077]
- [4] Wickramarachchi CT, Gardner P, Poole J, Hübler C, Jonscher C, Rolfes R. Damage localisation using disparate damage states via domain adaptation. *DCE* 2024; 5 [https://doi.org/10.1017/dce.2023.29]
- [5] G.C. Cawley NT. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 2010; (Vol. 11): 2079–107.
- [6] R.J. Allemang. The Modal Assurance Criterion - Twenty Years of Use and Abuse. *Sound and Vibration* 2003; (Vol. 37, No. 8): 14–21.
- [7] D.P. Hensley RM. Extending SMAC to Multiple Reference FRFs. In: *Extending SMAC to Multiple Reference FRFs*; 2006.
- [8] Lerch A. *An Introduction to Audio Content Analysis*. Wiley 2012.
- [9] Staszewski WJ, Robertson AN. Time-frequency and time-scale analyses for structural health monitoring. *Philos Trans A Math Phys Eng Sci* 2007; 365(1851): 449–77 [https://doi.org/10.1098/rsta.2006.1936][PMID: 17255047]
- [10] Chen T, Guestrin C. XGBoost. In: Krishnapuram B, Shah M, Smola A, Aggarwal C, Shen D, Rastogi R, editors. *XGBoost*; 2016. New York, NY, USA: ACM; 785–94.
- [11] Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; (Vol. 12): 2825–30.
- [12] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G, editors. *Optuna*; 2019. New York, NY, USA: ACM; 2623–31.
- [13] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, Balázs Kégl. Algorithms for Hyper-Parameter Optimization. In: *Algorithms for Hyper-Parameter Optimization*; 2011. *NeurIPS*; 2546–54.

- [14] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *jair* 2002; 16: 321–57 [https://doi.org/10.1613/jair.953]
- [15] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. In: *The Balanced Accuracy and Its Posterior Distribution*; 2010. IEEE; 3121–4.
- [16] J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* 2006; (Vol. 7): 1–30.
- [17] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10(3): e0118432 [https://doi.org/10.1371/journal.pone.0118432][PMID: 25738806]
- [18] Sohn H. Effects of environmental and operational variability on structural health monitoring. *Philos Trans A Math Phys Eng Sci* 2007; 365(1851): 539–60 [https://doi.org/10.1098/rsta.2006.1935][PMID: 17255051]