Integrating Materials and Manufacturing Innovation (2023) 12:70–81
https://doi.org/10.1007/s40192-023-00293-8

TECHNICAL ARTICLE

Quantitative Benchmarking of Acoustic Emission Machine Learning Frameworks for Damage Mechanism Identification

C. Muir[1] · N. Tulshibagwale[2] · A. Furst[2] · B. Swaminathan[1] · A. S. Almansour[3] · K. Sevener[4] · M. Presby[3] · J. D. Kiser[3] · T. M. Pollock[1] · S. Daly[2] · C. Smith[3]

Received: 9 June 2022 / Accepted: 18 January 2023 / Published online: 13 February 2023

Abstract

A challenging opportunity in structural health monitoring of composite materials is using machine learning (ML) methods to classify acoustic emissions according to the damage mechanism that emitted the signal. A wide variety of ML frameworks have been developed; however, lack of ground truth datasets in addition to limited overlap between experimental configurations has precluded any direct, quantitative benchmarking of their accuracy. Here, we generate a ground truth dataset comprised of pencil lead breaks with known angles of incidence, θ. Each angle generates a unique frequency spectrum that changes continuously with θ, which could be analogous to attributes of acoustic emission signals generated from failure processes, such as those that occur in composites. Five frameworks are then applied to the ground truth dataset and benchmarked according to their ability to discriminate between two sets of signals with a fixed $\Delta\theta$. A discussion of their performance as related to choice of features is given, and a set of guidelines for best-practices for feature selection and standardized practices are proposed.

Keywords Machine learning · Acoustic emission · Composites · Clustering

Introduction

One longstanding hypothesis investigated in the scientific community is that damage mechanisms in multi-phase structural materials, such as composites, can be identified directly from the strain waves, or acoustic emission (AE), they produce [1–4]. Developing this capability has wide-reaching ramifications for lifetime prediction investigations and *in operando* monitoring of advanced structural materials.It would allow researchers to augment damage triangulation [5, 6], lifetime prediction [7], and high-resolution optical studies [8, 9] with complementary mechanism-informed data streams.

However, directly mapping a waveform to its source mechanism is non-trivial. In a single experiment, difficult-to-capture factors such as transducer contact, specimen geometry, and loading configuration all influence the waveform structure. Because of these effects, it is infeasible to implement bottom-up approaches where the measured waveform is directly compared to waveforms generated from computational models [10–14]. While these experimental factors are often difficult to capture in physics-based models, the effect they have on an acoustic signal as it travels from source to sensor is constant. It is therefore more effective to group waveforms according to their *differences* and assign mechanisms to groups. As a consequence, unsupervised machine learning (ML) methods (i.e., pattern recognition techniques) which classify signals based on differences in their signal structure are an effective strategy for damage mechanism identification, with many such frameworks being developed over the last two decades [15–21].

A general inspection of these frameworks yields an important observation: there is no ground truth dataset, wherein mechanisms have been directly assigned to each

✉ C. Smith
craig.e.smith@nasa.gov

[1] Materials Department, University of California Santa Barbara, Santa Barbara, CA, USA

[2] Department of Mechanical Engineering, University of California Santa Barbara, Santa Barbara, CA, USA

[3] NASA Glenn Research Center, Cleveland, OH, USA

[4] Department of Materials Science and Engineering, University of Michigan - Ann Arbor, Ann Arbor, MI, USA

individual signal, which is suitable for benchmarking performance of an AE-ML framework [22, 23]. Previous studies have attempted to create ground truth datasets using a variety of strategies, for example by designing the loading configurations and sample geometry to *promote* 1–2 damage mechanisms [15, 19, 24]. Generally, such strategies produce datasets which are not usable for quantitative benchmarking because the ground truth is still unknown. For example, in the absence of visual confirmation, it is possible that geometries designed to promote fiber failure in composites may contain numerous signals from other mechanisms, such as interfacial damage, as well [25–28].

Because the datasets described above are unsuitable for benchmarking accuracy, indirect measures of framework performance have been employed. Metrics that have been studied include the tendency to fall into a local minimum, compactness of clusters, and how well average cluster characteristics match previous literature [19, 29–31]. However, these cannot be used as a proxy metric for accuracy because they do not measure accuracy or discriminating power [32–34]. Therefore, there is a need for datasets and methodologies that can be used for the standardized, quantitative assessment of AE-ML frameworks.

Toward the goal of generating datasets which can be used to assess discriminating power, pencil lead breaks (PLBs) offer a powerful solution. PLBs emit signals whose frequency content can be controlled by varying the angle of incidence, θ [5, 35, 36]. Incremental increases to the angle of incidence $\Delta\theta$ result in incremental increases to the low-frequency (flexural wave) component of the AE signal.

Since signal structure is uniquely determined by its frequency content, sets comprised of signals generated at angles θ and $\theta + \Delta\theta$ can be used to quantitatively evaluate the ability of a framework to group signals according to their emitting source. Frameworks that can accurately distinguish between signals generated from θ and $\theta + \Delta\theta$, when $\Delta\theta$ is small, will have higher discriminating power than frameworks which cannot. Therefore, a dataset comprised of signals generated from known values of θ can be used to quantitatively assess the discriminating power of AE-ML frameworks, investigate how specific changes to frameworks impacts discriminating power, and guide decisions on improvement.

In this work, an acoustic emissions (AE) dataset comprised of PLB acoustic sources was generated at one reference angle θ_0 and five benchmarking angles θ_b. Five ML frameworks from literature were applied to this dataset, and their performance was assessed. We investigated how changing feature choice impacts framework discriminating power and found that when *only* frequency domain features are used, discriminating power rises. Moreover, it is shown that for discriminating between different PLBs, the choice of ML algorithm was unimportant and a framework's performance could be attributed primarily to the feature set. Finally, we propose a set of guidelines for standardized benchmarking procedures for AE-ML frameworks, strategies for identification of salient features, and future benchmarking procedures.

Materials and Methods

Data Collection

All pencil lead breaks (PLBs) were conducted with Pentel 0.5 mm HB leads and a nominal free lead length of 4 mm [36]. A Pentel GraphGear 500 mechanical drafting pencil was fixed to a custom-built, displacement controlled, load frame (Fig. 1). The load frame was composed of a rotational stage, which allowed for angle adjustments in increments of 2° (corresponding manual angle measurement error is $\frac{1}{2}$ the unit of measurement, or $\pm 1°$), and two precision-adjust linear stages. The aluminum plate on which the PLBs were conducted had an unsupported span of 200.7 mm, width of 51.0 mm, and thickness of 1.2 mm.

PLBs were recorded at $20°, 22°, 26°, 30°, 36°$, and $40°$. PLBs were generated by lowering the pencil via the linear Z stage until the lead fractured on the aluminum plate. For each angular condition, the rotational stage was fixed using a set screw and the set screw was loosened only to

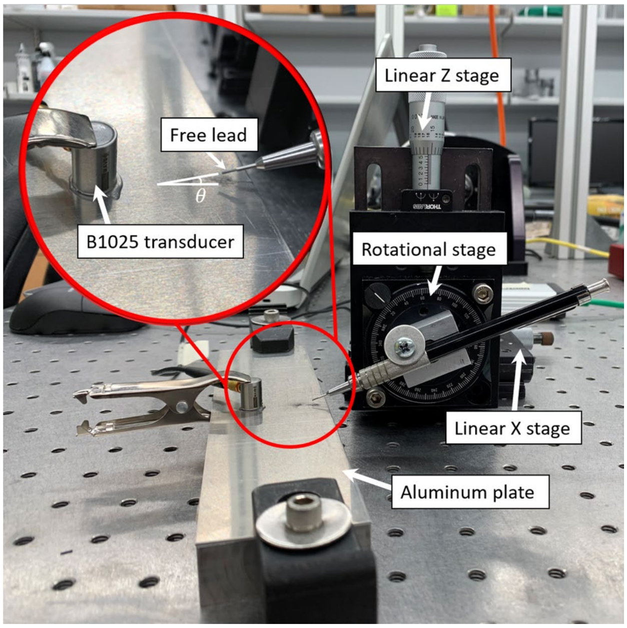

Fig. 1 Photograph of experimental setup. A mechanical pencil is attached to a rotational stage which controls the angle of incidence θ. The linear X stage is used to position the tip of free lead at a consistent location on the aluminum plate. The linear Z stage is used to lower the pencil lead until fracture. The resultant waveform generated is recorded by the piezoelectric B1025 transducer, located approximately 25 mm from the tip of free lead

change angles. During an angle change, the linear X stage was adjusted to maintain a nominal distance of 25 mm from the PLB source to the sensor. Upon inspection of collected AE signals, some were found to be reflections. These reflections presented themselves as a second, low amplitude signal occurring immediately after the initial PLB, and were excluded from the data set. Due to the exclusion of these reflections, each angle has a differing number of signals, between 75-111. For the purposes of this analysis, only the first 75 signals of each angle were clustered by an AE-ML framework.

AE was recorded using a piezoelectric B1025 transducer (Digital Wave Corporation, Centennial, CO) with a broadband response of 50–2000 kHz (Fig. 1). The threshold voltage was set to 0.1 V, the number of pre-trigger points was set to 256, and the total length of signal captured was 1024 points at a rate of 10 MHz. The sensor was fixed to the aluminum plate with an alligator clamp using vacuum grease as a coupling agent. The sensor was not remounted at any point during the experiment, meaning PLBs at all angles were conducted for a fixed sensor coupling. The authors note here that because the coupling is unchanging during all data collection, an absolute calibration of the sensor as described in [37] is not necessary. Additionally, all signals were collected within a single 3 h span in a temperature controlled laboratory where environmental factors which might otherwise affect the absolute sensor calibration, such as temperature, were assumed to be unchanging. Unsupervised frameworks group signals according to differences in signal features, rather than the absolute values of those features. Since the absolute sensor calibration is unchanging, any differences in signal features can be attributed to changes in the angle of incidence.

Signals collected at the reference angle $\theta_0 = 20°$ and signals at a single benchmarking angle $\theta_b \in [22°, 26°, 30°, 36°, 40°]$ were clustered using each of the frameworks described in Sect. Framework Descriptions and Accuracy Metrics, and relative discriminating power was assessed quantitatively using the procedure described below.

Quantitative Benchmarking

The permutation model of the adjusted Rand index (ARI) was used to benchmark frameworks. The ARI, which ranges from 0 to 1, measures accuracy of ML-calculated clusters as compared to the ground truth in a label-agnostic way. It compares the membership similarity of objects in the ML-calculated clustering, A, to the membership similarity of objects in the ground truth clustering, B, and assigns a higher number if similarities are high [38]. In the context of this work, signals from θ_0 and θ_b are fed to an AE-ML framework. These signals are then assigned a label by the framework, either 0 or 1, depending on if the framework believes the signal should be grouped with θ_0 or θ_b. The ML-assigned label of each signal is then compared to the ground truth, the known angle at which the signal was collected. If the membership similarity of the ML-assigned labels and the true angles are similar, then the ARI will take on higher values.

The ARI is an adjusted-for-chance version of the Rand index (RI) and is calculated as [38]:

$$RI(A, B) = \frac{N_{11} + N_{00}}{\binom{N}{2}} \tag{1}$$

where N is the number of signals, N_{11} is the number of signal pairs which are grouped into the same cluster in A and B, and N_{00} is the number of signal pairs that are grouped into different clusters in both A and B. The ARI can then be calculated as [39, 40]:

$$ARI(A, B) = \frac{RI(A, B) - \mathbb{E}[RI(A, B)]}{\max[RI(A, B)] - \mathbb{E}[RI(A, B)]} \tag{2}$$

where $\mathbb{E}[RI(A, B)]$ is the expected value of the RI under a random model. The ARI is bound between 0 and 1, with 0 corresponding to random label assignments and 1 corresponding to perfectly matching labels. While many cluster similarity metrics exist, the ARI was chosen to compare partitions because it can be calculated for any number of clusters (provided the number of clusters in each partition is equal), and it accommodates unbalanced cluster sizes [40, 41]. Signals from θ_0 and θ_b were clustered by each framework. The value of $\Delta\theta = \theta_b - \theta_0$ at which the ARI vanished represents the point at which the framework has lost all discriminating power, and is unable to identify differences between two groups of different signals.

Framework Descriptions and Accuracy Metrics

A general AE-ML framework can be described by workflow in Fig. 2. Following data collection (Fig. 2a), the most important step in the framework is the selection of the feature set (Fig. 2b). Waveforms can only be sorted according to their source mechanism *if* the feature set captures something fundamental about the waveform-mechanism relationship. Features can be classified as belonging to the time domain, frequency domain, or time-frequency domain. However, there is little consensus as to which category is best suited for damage mechanism identification. In fact, even when two frameworks leverage features within the same domain, their feature sets differ. Consequently, each framework uses a unique feature set, where d pertinent features are identified, extracted, and stored as

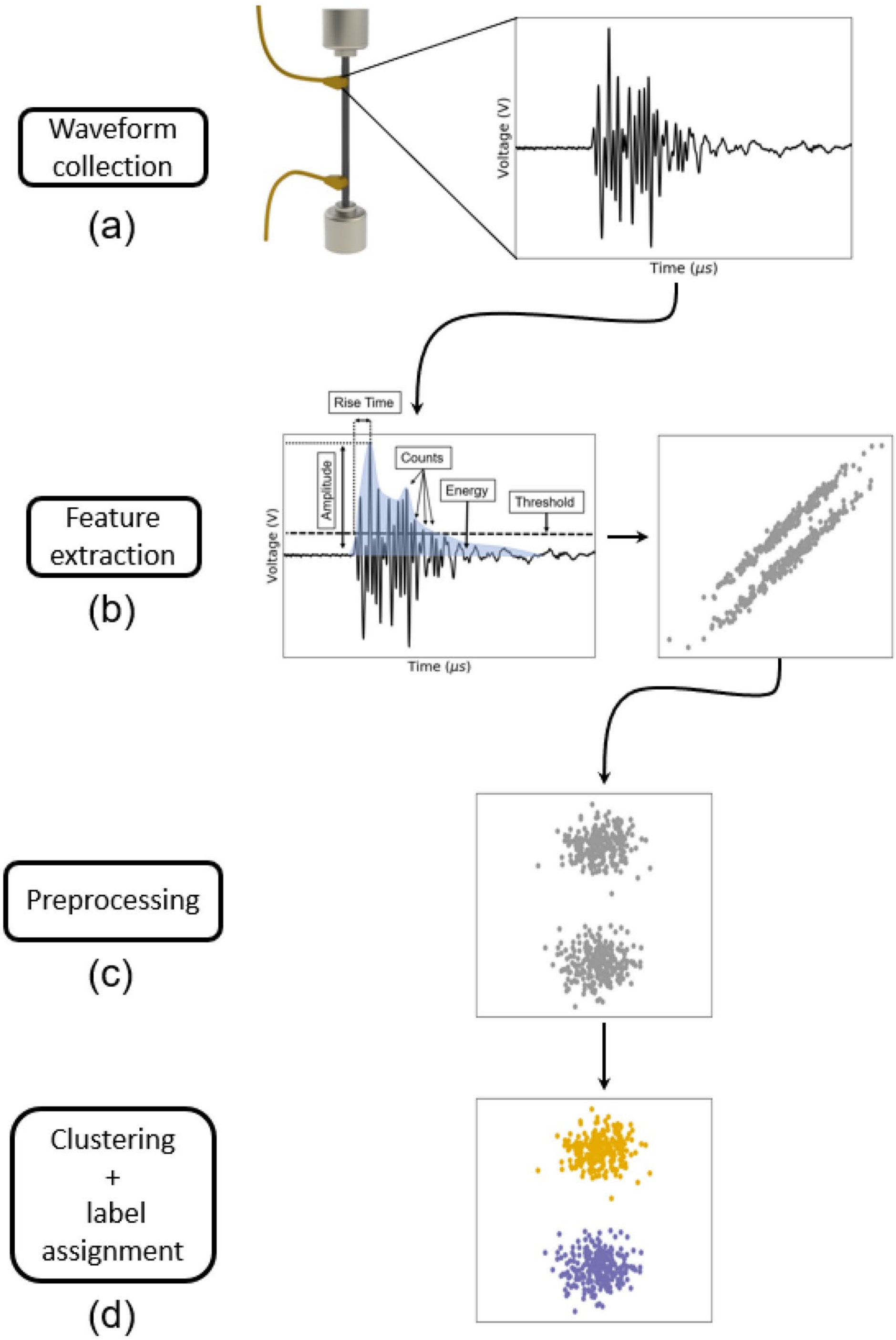

Fig. 2 Workflow diagram of an AE-ML framework. **a** Waveforms are collected and **b** pertinent features are extracted from the waveforms, which are then represented as vectors in feature space. **c** Feature vectors can then be re-scaled and/or re-mapped before **d** the clustering algorithm is applied and feature vectors are labeled. Every AE-ML framework follows this procedure

a feature vector $\mathbf{v} \in \mathbb{R}^d$ (Fig. 2b). The reader is referred to the original investigations for discussions on why particular feature sets were chosen [4, 17, 20, 21, 29].

Next, individual features of a feature vector may be re-scaled or re-mapped with a transformation (Fig. 2c). Similar to the variations in feature sets, each framework utilizes a different set of pre-processing steps. Finally, the ML algorithm is applied to partition feature vectors by assigning them to clusters, where feature vectors in the same cluster are proximal under a chosen distance metric (Fig. 2d).

The frameworks described in the following sub-sections follow this workflow and were adopted *directly* from literature. They were chosen to span the current space of diverse feature set types and ML algorithms [22]. The key differences between frameworks are the choice of feature sets, pre-processing steps, and ML algorithm. The specific choice of feature set, pre-processing steps, and ML algorithm are further summarized in Table 1. In Sect. Results and Discussion, we provide key findings and discuss the impact of feature selection.

Base Framework

We define a Base Framework relative to subsequent frameworks, which are variations on this base (either by swapping out the feature set, ML algorithm or both). This framework employs a time-domain feature set as investigated in [17]:

1. average frequency (number of counts/signal length)
2. rise frequency (average frequency from signal start to maximum amplitude)
3. *ln*(energy)
4. *ln*(rise time/duration)
5. *ln*(amplitude/rise)
6. *ln*(amplitude/decay time)
7. *ln*(amplitude/average frequency)

The start and end time of an experimental signal was determined by the first and last crossing of a floating 10% voltage threshold.

Each feature was scaled by the maximum observed value of that feature, using the MaxAbsScaler method in [42], such that they fell in the range [-1,1]. A principal component analysis (PCA) transformation was performed, and principal components containing ≥ 95% of the variance were retained. Distances, d, between any two feature vectors, $\mathbf{x}, \mathbf{y}$ were calculated using a modified Euclidean metric:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i \lambda_i (x_i - y_i)^2} \tag{3}$$

where x_i and y_i are the ith vector components of the feature vectors in the PCA basis, and λ_i is the eigenvalue of the i^{th} PCA axis. As the Scikit-learn implementation of k-means enforced the standard Euclidean metric, a rescaling of feature vectors was required to accommodate the modified Euclidean metric (proof is given in 6):

$$x_i' = \sqrt{\lambda_i} x_i \tag{4}$$

It should be noted that the distance metric in Eq. 3 differs from the standard PCA whitening approach, where distances along axes with large eigenvalues are contracted, rather than elongated [33].

K-means was then applied to the feature vectors. For a detailed description of the k-means algorithm, the reader is referred to [33]. Because k-means is not guaranteed to converge to an optimum solution, it is typically run many times and the initialization with the lowest value of loss function is taken [43]. To determine the number of re-initializations needed, convergence checks were performed by increasing the number of re-initializations until the loss remained unchanged. The minimum objective function did not change after 2×10^3 re-initializations. To conservatively ensure a global optimum of the objective function had been reached, the number of re-initializations was set to 2×10^4. Similarly, an error tolerance of 0.0001 and 300 iterations were sufficient to ensure a local optimum was reached within a single initialization of k-means.

Agglomerative Framework

The agglomerative framework [29] used the feature set:

1. amplitude
2. peak frequency

Rather than partitioning feature vectors by k-means as in the Base framework, the Agglomerative framework uses a hierarchical agglomerative approach. In this approach, each data point is initially defined as a cluster. Clusters are then iteratively merged such that the chosen objective function (usually the sum of squared distances) is extremized. For a discussion of this algorithm, the reader is referred to [29, 42].

The linkage type, the parameter defining pairwise distances between points, was not reported in the original work. Here, each linkage type was tested, and no linkage type was found to outperform another.

Spectral Framework

The Spectral framework [21] used the partial power feature set. The i^{th} component of the feature vector is as follows:

$$x_i = \frac{\int_{k_{i-1}}^{k_i} F[s(t)]dk}{\int_{k_0}^{k_d} F[s(t)]dk} \tag{5}$$

where $F[*]$ is the Fourier transform operator, $s(t)$ is the recorded signal, k_i and k_{i-1} are the frequency bounds over which integration is performed, and d is the number of entries in the feature vector. We set $k_0 = 200$ kHz, $k_d = 800$ kHz, and $d = 23$. The width of integration bounds, $k_i - k_{i-1}$, was set to be equal for all i as in [21].

The sklearn implementation of spectral clustering was used to cluster the feature vectors [42]. A detailed explanation of the algorithm can be found in [44]. The ARPACK eigensolver was used and the number of nearest neighbors was set to $NN = 5$. To ensure cluster membership did not depend on initialization parameters, convergence checks were performed for error tolerance and maximum number of iterations. The cluster membership was found to stabilize after 10 re-initializations. To conservatively ensure stability, the number of re-initializations was set to 100.

Frequency Framework

The frequency framework used a feature set in the *frequency* domain [4]:

1. average frequency
2. reverberation frequency
3. rise frequency
4. peak frequency
5. frequency centroid
6. weighted peak frequency
7. partial powers from 0–150 kHz, 150–300 kHz, 450–600 kHz, 600–900 kHz, and 900–1200 kHz

Features were independently normalized with the variance scaler, which centers features to have zero mean and scales them to unit variance. Feature vectors were then clustered with k-means. The same convergence checks as the Base framework were conducted, and identical parameters were found to be sufficient for convergence.

WPT Framework

The wavelet packet transform (WPT) framework extracted features through application of a WPT [20]. Waveforms were subjected to a WPT on three levels using the Daubechies wavelet of order 2 as the mother wavelet. Fractional energies carried in each node were calculated, and the five least correlated values were retained. These, in addition to the waveform energies read by the AE software, were used as features. Feature vectors were then normalized with the maximum value scaler and subjected to PCA. Principal components containing $\geq$ 95% of the variance were retained. The feature vectors were then partitioned via k-means, using the modified Euclidean metric (Eqn. 3).

Convergence checks were conducted and parameters identical to the Base framework were found to be sufficient for convergence. It should be noted that the algorithm described by [30] and used by [20] is k-means *optimized* by a genetic algorithm. Thus, a fully converged k-means solution will not differ from a fully converged genetic solution.

Results and Discussion

The frequency content of PLB signals from our experimental configuration was able to be precisely controlled by varying the angle of incidence. Signals generated were found to follow expectations from plate-wave theory [35, 36]; as the angle of incidence increased, the low-frequency components of the signal were observed to increase in power Fig. 3. Moreover, little variation in PLB signals was observed within a each angular condition. The mean signal and its standard deviation envelope are shown in Fig. 4. As a result of the small variation between signals for a single angular condition, when signals are represented in feature space (Fig. 2b), the standard deviation of those features is smaller than if signals had a large variability. If a benchmark set were constructed from more repeatable acoustic signals, it would be expected that signals at different angles would form tighter clusters in feature space, and subsequently the ARI of each framework at each value of $\Delta\theta$ would increase.

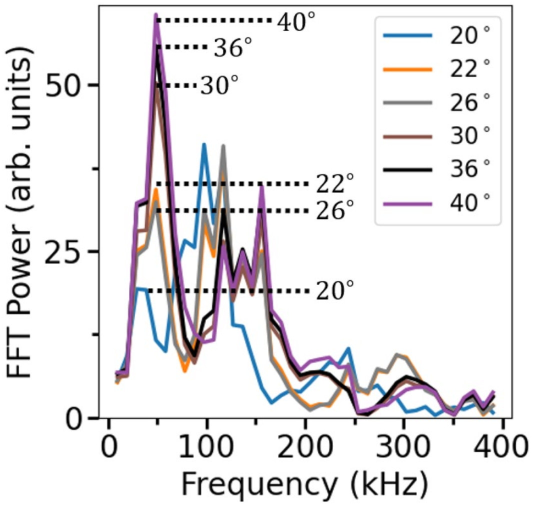

Fig. 3 Fourier transform (FFT) of average signals at each angular condition (Fig. 4). As the angle of incidence increases, the low-frequency components increase in power following the findings of [35] and [36]

By comparing ARI values at a fixed value of $\Delta\theta$, it is possible to quantitatively evaluate the relative discriminating power of various frameworks. The frameworks listed in Table 1 were applied to group signals according to the procedure described in Sect. Quantitative Benchmarking. The accuracy of each framework was plotted as a function of $\Delta\theta$ (Fig. 5). The discriminating power of each framework increased with $\Delta\theta$ which can be attributed to increasing differences in the signal structure as a function of $\Delta\theta$. Frameworks exhibiting higher ARI values at lower values of $\Delta\theta$, such as the Spectral and Frequency frameworks, have higher overall discriminating power and will likely be able to distinguish between damage mechanisms that emit similar signals.

Saliency of Features

Between the five frameworks, no single feature set was used. Although this is common in the broader context of

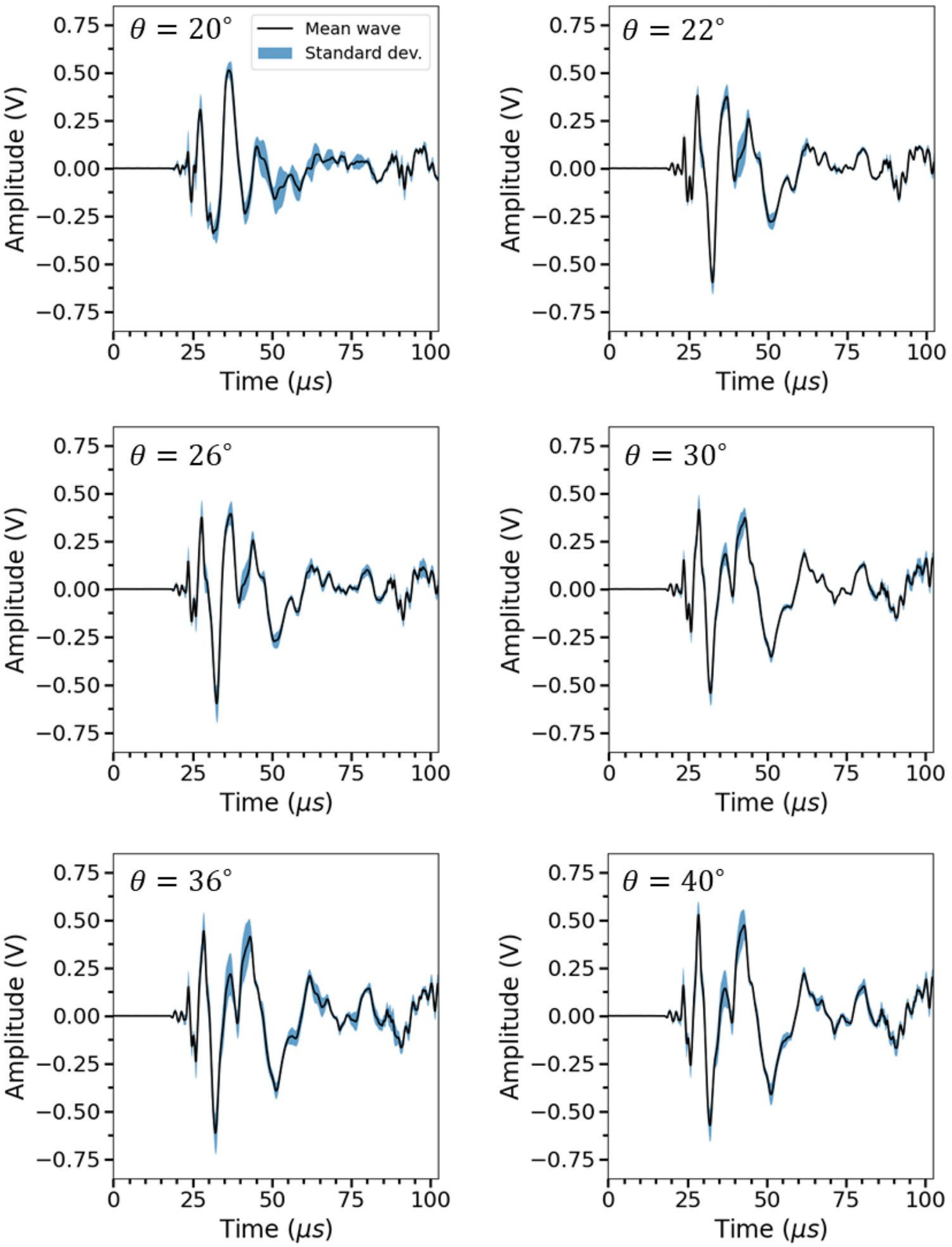

Fig. 4 The mean PLB signal and point-wise standard deviation at each angular condition. Signals generated using the experimental fixture shown in Fig. 1 were found to be repeatable, while still containing variation that might be expected from signals collected during *in operando* health monitoring

AE-ML frameworks [22, 23], the lack of consensus raises an important question: ”What features *should* be used for the purposes of AE signal discrimination?”. Addressing this question is of utmost importance, since the discriminating power of a framework hinges on how signals are represented [22, 45].

For signal discrimination, both exclusion of noisy features and inclusion of useful features is necessary: a principle known as the Ugly Duckling theorem [45]. To highlight the degree to which this principle impacts discriminating power, features were parametrically excluded from the Frequency framework and Base framework. In the Frequency framework, ARI was maximized when clustering was performed using average frequency, rise frequency, and partial power from 150–300 kHz. When these three optimal features were used to encode signals, the ARI at $\Delta\theta = 2°$ increased from 0.681 to 0.973, representing a change from modest to high discriminating power. A similar procedure was conducted for the Base framework, and when only the average frequency and log(amplitude/average frequency) were included, the ARI increased from 0.325 to 0.82.

While such parametric studies can yield insight into which features are useful for a specific dataset, they are less effective in identifying universally salient features. For this purpose, it is necessary to consider the physics of the emitting source on a case-by-case basis and when possible,

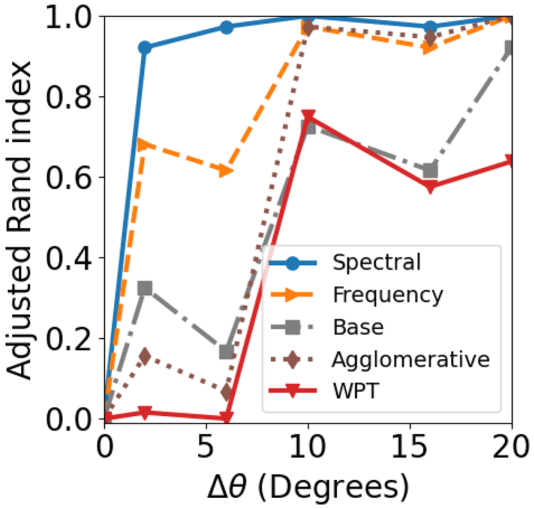

Fig. 5 The ARI of each framework as a function of $\Delta\theta$. ARI values exceeding 0.4 are correlated with good discriminating power, whereas values near 0 correspond to no discriminating power. The discriminating power of each framework increases with $\Delta\theta$ and high ARIs at low values of $\Delta\theta$ are better suited for clustering signals whose differences are minor. The ability to directly compare accuracy between frameworks allows researchers to choose an appropriate framework for their specific needs

exclude features that are dependent on external and uncontrollable factors unrelated to the source mechanism. For example, although amplitude is correlated to the angle in this dataset (Fig. 6a), it should not be used for sorting signals from multi-phase materials because it is convolved with factors, such as crack area formed and the source-to-sensor distance, which are unrelated to individual mechanisms [46]. Similarly, even though rise time is commonly used in AE-ML frameworks, [18, 19, 30, 47–49], it is more strongly related to the source-to-sensor distance, due to the different velocities of the flexural and extensional wave components [5]. Consequently, two signals emitted from similar locations will have similar rise times, even if the emitting mechanisms are different (Fig. 6b).

Limitations of the PLB Dataset

Intuitively, a framework with higher ARI values is a promising candidate for damage mechanism identification, when signals from different mechanisms are expected to be similar. However, the degree to which performance on the PLB dataset translates to performance under more realistic conditions and material systems is unknown. Specifically, the PLB signals in this dataset are collected under the strictest possible conditions; signals are from a single source-to-sensor distance, sensor coupling, and source type (e.g., pencil lead), removing the effect of factors that influence a signal such as dispersion, attenuation, and absolute frequency response. ML approaches for mechanism identification must ultimately be robust to these effects. Although this dataset represents a first step toward quantitative benchmarking, a full characterization of framework performance under realistic conditions is still required.

Another limitation of the dataset we have collected is the angular resolution; the ±1° tolerance of the rotational stage has implications on the measured ARI. For example, if the true $\Delta\theta$ between two angular conditions was less than the reported value, due to the ±1° tolerance, signals generated at these angles would be more similar than expected. Consequently, the ARI measured would be lower than if the signals had been collected from a true angular condition with a larger value of $\Delta\theta$. The exact degree to which the ARI would change is highly dependent on how each feature varies with θ, and also subject to any data dependent

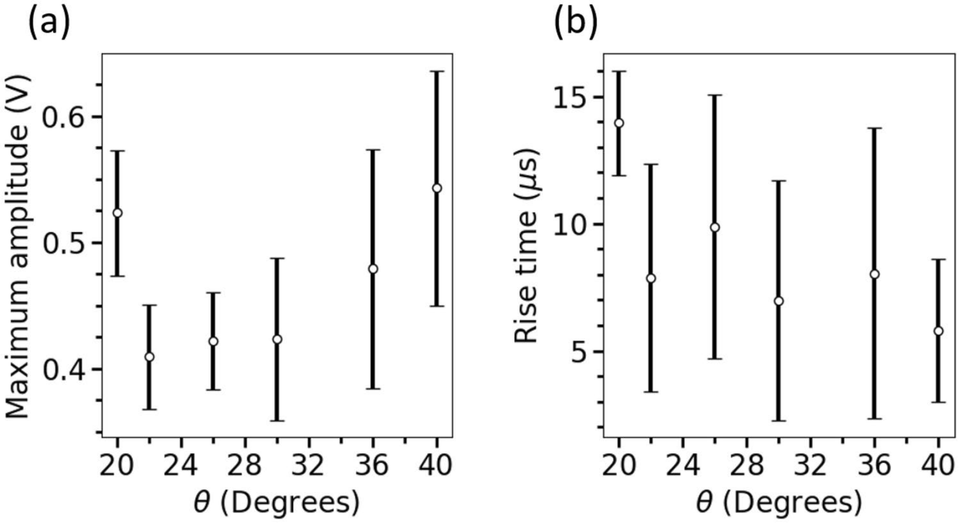

Fig. 6 The **a** average maximum amplitude and **b** average rise time of signals generated at each angle of incidence θ. Error bars correspond to 1 standard deviation. There is no consistent difference between values in either feature. Because it is possible to construct many sets of unique signals with indistinguishable amplitudes and rise times, they should not be considered salient features and their use should be taken with caution

pre-processing, such as PCA, which would further impact framework performance.

Finally, also due to the angular resolution, the current experimental setup prevents collecting signals from values of $\Delta\theta < 2°$. In order to properly benchmark frameworks, there must be at least one value of $\Delta\theta$ where ARIs are not saturated at 1. For example, for $\Delta\theta = 20$, the Spectral, Frequency, and Base frameworks all perform equally well, but $\Delta\theta = 6$ allows for comparison of discriminating power (Fig. 5). As the community continues to improve the discriminating power of frameworks, ARI values will increase. Consequently, it will become necessary to collect signals from values of $\Delta\theta < 2°$, below what we have allowed for in this study, to ensure frameworks' performances can be separated.

Guidelines and Conclusions

While many AE-ML frameworks have been developed and implemented, the lack of ground truth datasets has restricted discussions of their strengths and limitations, particularly with respect to feature choice, and has prevented development of standardized quantitative benchmarking procedures. In this section considerations for the quantity of data in benchmarking sets, types of features that should be included in a framework, and transparent benchmarking practices are discussed.

The performance of an unsupervised framework is intrinsically tied to how well the sampled data represents its population distribution. In the context of AE-ML benchmarking datasets, it is critical to ensure enough signals have been collected to capture statistical variations. If too few waveforms are collected at any angle, it is unlikely that the sampling distribution will represent the population distribution of waveform features (Fig. 2b). Consequently, the addition on new waveforms will lead to spurious performance of an AE-ML framework. To ensure enough signals are in a benchmarking set, a framework's performance must be shown to be independent of the number of signals collected. For this benchmarking set, it is demonstrated that 75 signals per angle are sufficient to ensure the ARI values we calculate are independent of the amount of data (Fig. 7).

Feature selection is of critical importance with respect to maximizing the discriminating power. As demonstrated in Sect. Saliency of Features, the inclusion of non-salient features was directly correlated with poor framework performance. Despite the importance of feature selection, there is little discussion within the literature as to why certain features were chosen [22]. As a result, many modern frameworks continue to include non-salient features (e.g., rise time) which negatively impact framework performance.

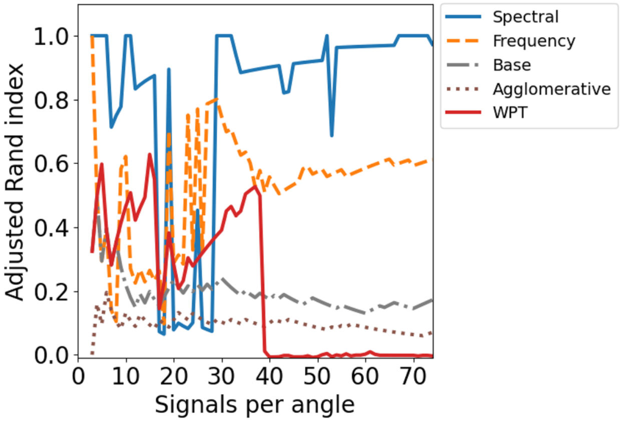

Fig. 7 Adjusted Rand index vs. number of signals per angle. Signals from $\theta_0 = 20°$ and $\theta_b = 26°$ ($\Delta\theta_0 = 6°$) were clustered using an increasing number of signals per angle. As the number of signals increased, the performance of frameworks becomes independent to the addition of new signals, indicating enough data is present to capture stochastic waveform variations

Table 1 Investigated framework summaries

Framework	Feature set	ML algorithm
Base [17]	Time domain	k-means
Agglomerative [29]	Mixed	Agglomerative
Spectral [21]	Frequency domain	Spectral clustering
Frequency [4]	Frequency domain	k-means
WPT [20]	WPT Energies	k-means

Toward better feature selection, universal features should be prioritized, and when possible, choice of feature set should be explicitly motivated. If it is possible to construct cases where a given feature cannot reliably discriminate between signals emitted by two unique sources, then the feature is likely convolved with factors unrelated to the source mechanism and are therefore not universal. The use of such non-universal features must be treated with caution. For example, although small amplitudes and large rise times have been correlated with Mode II cracks, these features are not universal because they are also a strong function of the source-to-sensor distance [50]. This makes it possible to construct a dataset where unique signals can appear artificially similar resulting from little to no statistical difference between features (Fig. 6).

Although universally salient features will not change between material systems or loading configurations, the values of the features might vary. For example, partial power appears to be a universally salient feature [4, 21, 47], but every frequency band does not provide equal discriminating power. As demonstrated by the parametric removal of Frequency framework features, the frequency band from 150–300 kHz was the most useful for signal discrimination. In this case, 150–300 kHz was useful for discrimination between two PLB signals; however, different frequency bands will be useful when the material system or loading conditions changes [12, 26].

Finally, publicly available standardized datasets should be used for quantitative benchmarking of frameworks. Although these types of benchmarking tools are common in other fields [51–54], they are absent in the AE community. Development and continued maintenance of these types of datasets will provide the tools necessary to assess the strengths and limitations of AE-ML frameworks and will allow for detailed discussions regarding the specific strengths and weaknesses of individual frameworks. In turn, this will provide transparency and trust in the results obtained from such frameworks, promoting their broader use in both scientific and engineering applications.

Appendix A

Let $\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}' \in \mathbb{R}^d$ with

$$\mathbf{x} = \sum_i^d x_i \mathbf{e}_i$$

$$\mathbf{y} = \sum_i^d y_i \mathbf{e}_i$$

$$\mathbf{x}' = \sum_i^d \sqrt{\lambda_i} x_i \mathbf{e}_i$$

$$\mathbf{y}' = \sum_i^d \sqrt{\lambda_i} y_i \mathbf{e}_i$$

and λ_i be the eigenvalue associated with the basis vector $\mathbf{e}_i$. Now, consider the standard Euclidean metric:

$$d_1(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2} \tag{6}$$

and the modified metric:

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i \lambda_i (x_i - y_i)^2} \tag{7}$$

It is sufficient to show that scaled vectors under the standard metric, $d_1(\mathbf{x}', \mathbf{y}')$, are equivalent to original vectors under the modified metric, $d_2(\mathbf{x}, \mathbf{y})$.

$$\begin{aligned} d_1(\mathbf{x}', \mathbf{y}') &= \sqrt{\sum_i \left(\sqrt{\lambda_i} x_i - \sqrt{\lambda_i} y_i\right)^2} \\ &= \sqrt{\sum_i \lambda_i \left(x_i - y_i\right)^2} \\ &= d_2(\mathbf{x}, \mathbf{y}) \end{aligned}$$

Acknowledgements C. Muir and B. Swaminathan gratefully acknowledge financial support from the NASA Space Technology Graduate Research Opportunities fellowship (Grants: 80NSSC19K1164 and 80NSSC17K0084 respectively). This paper was supported in part by a fellowship award under contract FA9550-21-F-0003 through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, sponsored by the Air Force Research Laboratory (AFRL), the Office of Naval Research (ONR) and the Army Research Office (ARO). S. Daly and T. M. Pollock gratefully acknowledge financial support from the National Science Foundation (Award: 1934641) as part of the HDR IDEAS^2 Institute. Additionally, N. Tulshibagwale was awarded an NDSEG fellowship partway through this project.

Data Availability The PLB benchmarking dataset described herein along with all code used to generate and analyze results is freely available at https://github.com/Muir-UCSB/AE_ML_Benchmarking.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Elsley RK, Graham LJ (1987) Pattern recognition in acoustic emission experiments. Pattern Recognit Acoust Imaging 0768:285
2. de Groot PJ, Wijnen PA, Janssen RB (1995) Real-time frequency determination of acoustic emission for different fracture mechanisms in carbon/epoxy composites. Compos Sci Technol 55(4):405–412
3. Johnson M, Gudmundson P (2000) Broad-band transient recording and characterization of acoustic emission events in composite laminates. Compos Sci Technol 60(15):2803–2818
4. Sause MG, Gribov A, Unwin AR, Horn S (2012) Pattern recognition approach to identify natural clusters of acoustic emission signals. Pattern Recogn Lett 33(1):17–23
5. Morscher GN, Godin N (2015) Use of acoustic emission for ceramic matrix composites. In: Bansal NP, Lamon J (eds) Ceramic matrix composites: materials. Modeling and technology, 1st edn. Wiley, New York, pp 571–590
6. Whitlow T, Jones E, Przybyla C (2016) In-situ damage monitoring of a SiC/SiC ceramic matrix composite using acoustic emission and digital image correlation. Compos Struct 158:245–251
7. Swaminathan B, McCarthy N, Almansour A, Sevener K, Pollock T, Kiser J, Daly S (2021) Microscale characterization of damage accumulation in cmcs. J Eur Ceram Soc 41(5):3082–3093
8. Hilmas AM, Sevener KM, Halloran JW (2020) Damage evolution in SiC/SiC unidirectional composites by x-ray tomography. J Am Ceram Soc 103(5):3436–3447
9. Maillet E, Singhal A, Hilmas A, Gao Y, Zhou Y, Henson G, Wilson G (2019) Combining in-situ synchrotron X-ray microtomography and acoustic emission to characterize damage evolution in ceramic matrix composites. J Eur Ceram Soc 39(13):3546–3556
10. Huang W, Rokhlin SI, Wang YJ (1995) Effect of fibre-matrix interphase on wave propagation along, and scattering from, multilayered fibres in composites. Transfer Matrix Approach, Ultrasonics 33(5):365–375
11. Wilcox PD, Lee C, Scholey JJ, Friswell MI, Wisnom M, Drinkwater B (2006) Progress towards a forward model of the complete acoustic emission process. Adv Mater Res 13–14:69–76

12. Sause MG, Horn S (2010) Simulation of acoustic emission in planar carbon fiber reinforced plastic specimens. J Nondestr Eval 29(2):123–142
13. Gall TL, Monnier T, Fusco C, Godin N, Hebaz SE (2018) Towards quantitative acoustic emission by finite element modelling: contribution of modal analysis and identification of pertinent descriptors. Appl Sci (Switzerland) 8(12):2557
14. McLaskey GC, Glaser SD (2012) Acoustic emission sensor calibration for absolute source measurements. J Nondestr Eval 31(2):157–168
15. Godin N, Huguet S, Gaertner R, Salmon L (2004) Clustering of acoustic emission signals collected during tensile tests on unidirectional glass/polyester composite using supervised and unsupervised classifiers. NDT and E Int 37(4):253–264
16. Kostopoulos V, Loutas TH, Kontsos A, Sotiriadis G, Pappas YZ (2003) On the identification of the failure mechanisms in oxide/oxide composites using acoustic emission. NDT and E Int 36(8):571–580
17. Moevus M, Godin N, Mili MR, Rouby D, Reynaud P, Fantozzi G, Farizy G (2008) Analysis of damage mechanisms and associated acoustic emission in two SiC$_f$/[Si-B-C] composites exhibiting different tensile behaviours. Part II : unsupervised acoustic emission data clustering. Compos Sci Technol 68(6):1258–1265
18. Marec A, Thomas JH, El Guerjouma R (2008) Damage characterization of polymer-based composite materials: multivariable analysis and wavelet transform for clustering acoustic emission data. Mech Syst Signal Process 22(6):1441–1464
19. Gutkin R, Green CJ, Vangrattanachai S, Pinho ST, Robinson P, Curtis PT (2011) On acoustic emission for failure investigation in CFRP: pattern recognition and peak frequency analyses. Mech Syst Signal Process 25(4):1393–1407
20. Maillet E, Godin N, R'Mili M, Reynaud P, Fantozzi G, Lamon J (2014) Damage monitoring and identification in SiC/SiC minicomposites using combined acousto-ultrasonics and acoustic emission. Compos A Appl Sci Manuf 57:8–15
21. Muir C, Swaminathan B, Fields K, Almansour AS, Sevener K, Smith C, Presby M, Kiser JD, Pollock TM, Daly S (2021) A machine learning framework for damage mechanism identification from acoustic emissions in unidirectional SiC/SiC composites. npj Comput Mater 7(1):1–10
22. Muir C, Swaminathan B, Almansour A, Sevener K, Smith C, Presby M, Kiser J, Pollock T, Daly S (2021) Damage mechanism identification in composites via machine learning and acoustic emission. npj Comput Mater 7(1):1–15
23. Saeedifar M, Zarouchas D (2020) Damage characterization of laminated composites using acoustic emission: a review. Compos Part B Eng 195(2019):108039
24. Farhidzadeh A, Mpalaskas AC, Matikas TE, Farhidzadeh H, Aggelis DG (2014) Fracture mode identification in cementitious materials using supervised pattern recognition of acoustic emission features. Constr Build Mater 67(PART B):129–138
25. Maillet E, Morscher GN (2015) Waveform-based selection of acoustic emission events generated by damage in composite materials. Mech Syst Signal Process 52(53):217
26. Ospitia N, Aggelis DG, Tsangouri E (2020) Dimension effects on the acoustic behavior of TRC plates. Materials 13(4):955
27. Sause MGR, Horn SR (2010) Influence of specimen geometry on acoustic emission signals in fiber. In: 29th European conference on acoustic emission testing pp 1–8
28. Hamstad MA (2007) Acoustic emission signals generated by monopole (Pencil Lead Break) versus dipole sources : finite element modeling and experiments. J Acoustic Emission 25:92–106
29. Saeedifar M, Najafabadi MA, Zarouchas D, Toudeshky HH, Jalalvand M (2018) Clustering of interlaminar and intralaminar damages in laminated composites under indentation loading using Acoustic Emission. Compos Part B Eng 144:206–219
30. Sibil A, Godin N, R'Mili M, Maillet E, Fantozzi G (2012) Optimization of acoustic emission data clustering by a genetic algorithm method. J Nondestr Eval 31(2):169–180
31. Morizet N, Godin N, Tang J, Maillet E, Fregonese M, Normand B (2016) Classification of acoustic emission signals using wavelets and random forests: application to localized corrosion. Mech Syst Signal Process 70–71:1026–1037
32. Jain AK (1995) Artificial neural networks for feature extraction and multivariate data projection. IEEE Trans Neural Netw 6(2):296–317
33. Jain AK, Murty P, Flynn P (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
34. Jain AK, Duin R, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37
35. Gorman MR, Prosser WH (1991) AE source orientation by plate wave analysis. J Acoustic Emiss 9(4):283–288
36. Sause MGR (2011) Investigation of pencil-lead breaks as acoustic emission sources. J Acoustic Emiss 29:184–196
37. ASTM International (2020) Standard practice for secondary calibration of acoustic emission sensors. Designation: E1781/E1781M-13
38. Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M (2011) Édouard Duchesnay, Scikit-learn: machine learning in python. J Mach Learn Res 12(85):2825–2830
40. Gates AJ, Ahn YY (2017) The impact of random models on clustering similarity. J Mach Learn Res 18:1–28
41. Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
42. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P (2020) SciPy 1.0 contributors, SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods 17:261–272
43. Fränti P, Sieranoja S (2019) How much can k-means be improved by using better initialization and repeats? Pattern Recogn 93:95–112
44. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 14
45. Watanabe S (1985) Pattern recognition: human and mechanical. Wiley, New York
46. Swaminathan B, McCarthy NR, Almansour AS, Sevener K, Musaffar AK, Pollock TM, Kiser JD, Daly S (2021) Interpreting acoustic energy emission in SiC/SiC minicomposites through modeling of fracture surface areas. J Eur Ceram Soc 41(14):6883–6893
47. Guel N, Hamam Z, Godin N, Reynaud P, Caty O, Bouillon F, Paillassa A (2020) Data merging of ae sensors with different frequency resolution for the detection and identification of damage in oxide-based ceramic matrix composites. Materials 13(20):1–22
48. WenQin H, Ying L, AiJun G, Yuan FG (2016) Damage modes recognition and hilbert-huang transform analyses of cfrp laminates utilizing acoustic emission technique. Appl Compos Mater 23(2):155–178
49. Kim JT, Sakong J, Woo SC, Kim JY, Kim TW (2018) Determination of the damage mechanisms in armor structural materials via self-organizing map analysis. J Mech Sci Technol 32(1):129–138

50. Aggelis DG, Shiotani T, Papacharalampopoulos A, Polyzos D (2012) The influence of propagation path on elastic waves as measured by acoustic emission parameters. Struct Health Monit 11(3):359–366
51. Deng L, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, 2009, pp 248–255, iSSN: 1063-6919
52. Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747
53. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. ArXiv. https://doi.org/10.48550/arXiv.1512.03385
54. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. ArXiv. https://doi.org/10.48550/arXiv.1511.06434