Journal of Sound and Vibration 596 (2025) 118788

Contents lists available at ScienceDirect

Journal of Sound and Vibration

journal homepage: www.elsevier.com/locate/jsvi

Methodology for the automated selection of time-frequency representations

Nathaniel DeVol [a], Christopher Saldaña [a], Katherine Fu [b,*]

[a] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, 801 Ferst Drive, Atlanta, GA 30332, United States
[b] Department of Mechanical Engineering, University of Wisconsin-Madison, 1513 University Avenue, Madison, WI 53706, United States

ARTICLE INFO

Keywords:
Time-frequency analysis
Bayesian optimization
Machine learning

ABSTRACT

Data preprocessing is a key step in extracting useful information from sound and vibration data and often involves selecting a time-frequency representation. No single time-frequency representation is always optimal, and no standard method exists for selecting the appropriate time-frequency representation, so selecting the time-frequency representation requires expert knowledge and is susceptible to human bias. To address this, this work introduces a methodology to automate the selection of a time-frequency representation for a dataset using only a subset of the healthy, or normal, class of data. To select the parameters for each type of time-frequency representation, Bayesian optimization is used. With a candidate from each type of time-frequency representation, the average similarity is used to select the final candidate. Additionally, the use of multiple time-frequency representations within a single model is explored. Because there is currently no objective method to compare the selected time frequency representations against, the proposed methodology is evaluated in two case studies. In the case studies, the time frequency representations are used as inputs to a simple convolutional neural network that achieved 100% accuracy in classifying bearing faults and 94% accuracy in classifying the contact tip to workpiece distance in wire arc additive manufacturing. Additionally, the proposed methodology presents a 75% and 94% reduction in the data size for the two case studies. This offers further benefits for reducing costs of data transmission and storage in modern digital manufacturing architectures.

1. Introduction

The recent rise in the Industrial Internet of Things has led to connected devices generating large amounts of data that characterize industrial and manufacturing processes. To make use of this data and provide insights into monitored processes, machine learning (ML) is being increasingly used in industrial applications [1,2]. One challenge with the increased adoption of ML is that the performance of a model greatly depends on the parameters of the model. One such parameter is the preprocessing that is performed on a dataset before training an ML model. The amount of preprocessing can range from using the raw signal to feature extraction, where the original dataset is transformed into a reduced number of features. Feature extraction is common when dealing with signals like vibration and can include statistical features, such as skew and kurtosis, or performing a time-frequency transformation [3]. Feature extraction is beneficial because the reduced dimensionality of the input helps to avoid the curse of dimensionality [4], but the extracted features must be carefully chosen since the quality of the extracted features strongly influences the results of a machine

* Corresponding author.
E-mail address: kfu26@wisc.edu (K. Fu).

https://doi.org/10.1016/j.jsv.2024.118788
Received 7 May 2024; Received in revised form 18 October 2024; Accepted 21 October 2024
Available online 22 October 2024

learning model [5]. The quality of the extracted features is especially important in unsupervised learning, which is of interest in industrial health monitoring because it is not always feasible to have labeled data for all possible failure modes [5].

In industrial ML, feature extraction from common data streams like vibration varies significantly from the major ML areas of image processing and natural language processing. Therefore, additional research is needed to understand how to optimize the preprocessing stage of industrial ML. The challenge with analyzing the time and frequency of a signal separately is that components may be misrepresented or smeared, particularly in non-stationary signals. The time-frequency domain offers the ability to capture both time and frequency components [6] and thus indicates the most generalized solution. In industrial applications, various time-frequency representations (TFR) have been used, and additional details can be found in review articles [7,8].

With the recent increase of machine learning in industrial areas, TFRs of sensor data have been shown to be effective in preprocessing sensor data for use in convolutional neural networks (CNN) [9]. Selecting the proper TFR, however, requires an expert with domain knowledge because there is no single TFR that is always better than the others [7,10]. Within the industrial domain, prior work has been done to evaluate the efficacy of different TFRs. Gao and Yan explored the use of the short-time Fourier transform (STFT), discrete wavelet transform (DWT), wavelet packet transform (WPT), and Hilbert-Huang transform (HHT) for representing the state of a bearing [11]. They found, from visual inspection, that the STFT and WPT were able to identify the frequency shifts associated with the defect growth. Kim et al. compared the fast Fourier transform (FFT), STFT, Wigner-Ville distribution (WVD), and DWT for representing the state of a grinding spindle [12]. By visual inspection between a normal and crack condition, the DWT was concluded to be the best. Verstraete et al. evaluated the STFT, continuous wavelet transform (CWT), and HHT by training a CNN for each TFR. From the accuracy of the trained CNNs, it was found that the CWT performed the best except one case where the STFT was the best performer [9]. While these works provide an evaluation of TFRs for their problems, they do not provide users with general guidance and require expert intervention since the TFR is chosen from visual inspection, or the TFR and appropriate parameters must be chosen by an expert. Additionally, multiple classes are needed before these methods can be applied.

In the TFR selection work reviewed here, the focus is on common time-frequency methods. This is only a subset of time-frequency methods available, and new time-frequency methods are continually introduced which offer improvements in particular use cases [13,14]. The challenge of selecting a TFR is further highlighted in recent unsupervised machine anomaly detection that uses sound and transformer-based networks. In the work, log-mel spectrograms were augmented with a 1D convolutional encoder and passed to a transformer together because the spectrograms were not found to be sufficient alone [15].

In addition to selecting the time-frequency method to use, the parameters for the method must be selected. The parameters that must be chosen depend on the TFR and affect the quality of the TFR. For example, when performing a STFT, the width of the window must be selected, and this width influences both the time and frequency resolutions due to Heisenberg's uncertainty principle [8]. Because of the importance of selecting these parameters, work has gone into optimal parameter selection for TFR, but before optimizing a TFR, the quality of a TFR must be quantified. The Shannon entropy has been used to optimize various aspects of the wavelet transform, including selecting the proper β value for the Morlet wavelet when performing signal denoising [16]. The Shannon entropy has also been used in combination with the energy at a particular wavelet scale [17–19]. This approach required the scale of interest to be known *a priori,* so its application is limited. Additionally, when starting with only healthy data, the scale at which a defect manifests may not be known. While the Shannon entropy has been shown to be useful, it does not work when negative values are present, which occur in some TFR, such as Cohen's class. To combat this, the Rényi entropy is also considered. The Rényi entropy is used by Sang and Williams to develop an optimal kernel for a reduced inference distribution, which reduces cross terms [20].

When selecting a mother wavelet for a wavelet transform, a common heuristic is that the mother wavelet should match the signal [21]. Following this approach, some works optimize wavelet selection based on the correlation between wavelet coefficients and the original signal [22,23]. However, matching the mother wavelet to the signal has been demonstrated to be non-optimal [21].

This prior work has shown that the TFR can be optimized using previously developed evaluation metrics, but it is limited because multiple states must be seen beforehand, which cannot always be assumed in industrial data. Additionally, past work only considers a single application area and the applicability to other domains is not known. Finally, because of the rather large search space, prior work has focused on a subset of TFRs. With a limited search space, the optimum may be missed. This work aims to fill these gaps by introducing a methodology for selecting a TFR to represent a dataset using only data from the healthy or normal class. In this, Bayesian optimization is investigated to search through parameters used in generating the TFR for each TFR class. This research thus provides practitioners with an objective method to select a TFR that may be used in downstream ML tasks, such as classification and regression.

2. Background

2.1. Time-frequency transformations

Analyzing a signal in the time domain as it is collected neglects all frequency information, which make it difficult to understand the signal. A FFT could be performed on a signal to analyze it in the frequency domain, but in doing so, temporal information is lost. Time-frequency analysis allows for an understanding of how the frequency information is changing in time. There is a myriad of time-frequency transformations that can achieve this. In this work, four of the most common transformations used in industrial condition monitoring are used. The TFRs used are the STFT, DWT, HHT, and WVD. Each of these TFRs are introduced below.

2.1.1. Short-time Fourier transform

The STFT is a simple method of transforming a signal into the time-frequency domain, where the frequency content is calculated on a short window of data. The STFT is described by the Eq. (1) [24]:

$$S(t,\tau) = \int_{-\infty}^{\infty} x(t)w(t-\tau)\exp\left(-\mathrm{j}2\pi f t\right)\mathrm{d}t \tag{1}$$

where $x(t)$ is the time signal and $w(t-\tau)$ is the window centered at time τ.

When performing the STFT there are several parameters that must be chosen, including the window length, window overlap, and window function. Because of the uncertainty principle, the time and frequency resolutions are inversely related. Thus, as the window length is reduced, the time resolution will increase, but the frequency resolution will decrease. The window overlap defines how far the window slides with each iteration. The simplest window function is that of a rectangular window, which takes a value of one within the window and zero outside it. The discontinuities caused by the edges of the rectangular window can cause spectral leakage. To minimize this effect, window functions such as the Hanning window, are used. The Hanning window forms a taper using a weighted cosine such that the window smoothly approaches zero at the edges.

2.1.2. Discrete wavelet transform

The wavelet transform decomposes a signal in terms of a family of wavelets. The shape of the wavelet is defined by the mother wavelet $\psi(t)$ used and is translated and scaled in time. The wavelet transform is described by the Eq. (2) [24]:

$$W(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right)\mathrm{d}t \tag{2}$$

where a is the scale and b is the translation.

The wavelet transform can be continuous or discrete, which differ in how they discretize the scale and translation parameters. The DWT uses exponential scales with a base equal to two, while the CWT uses a base smaller than two. This means that the CWT has a higher resolution but redundant information, while the DWT does not contain redundant information and can be quickly computed. Different mother wavelets have different properties. The Daubechies wavelets are compact in the time domain, but in principle, infinite in the frequency domain. The Daubechies wavelet family is comprised of different mother wavelets with varying numbers of vanishing moments. These mother wavelets are referred to as db*A,* where *A* is the number of vanishing moments. The simplest member, db1, is the same as the Haar wavelet. The symmlet and coiflet wavelet families were proposed to combat the asymmetry of the Daubechies wavelets. The symmlet and coiflet wavelets are referred to as coif*A* and sym*A*, where *A* is the number of vanishing moments. The biorthogonal wavelet family is comprised of real, symmetric wavelets, which use one mother wavelet to deconstruct a signal and another to reconstruct it. The biorthogonal wavelets are referred to as bior*A.B,* where *A* and *B* are the number of vanishing moments in the decomposition and reconstruction wavelets respectively. For more details on wavelets and mother wavelets please refer to [25]. Much like there is no single method for selecting a TFR, there is also no standard method for selecting the appropriate mother wavelet. One common approach is to select the wavelet that matches the condition that is to be detected [24]; however, this has been shown to be ineffective [21].

2.1.3. Hilbert-Huang Transform

The HHT is computed by decomposing a signal using empirical mode decomposition to get its intrinsic mode functions (IMFs) and calculating the instantaneous frequency of each IMF [26]. Extracting IMFs is an iterative process called sifting. The number of iterations represents the number of IMFs generated and continues until a stopping criterion is reached. In this work sifting is stopped when the IMF is less than a prespecified threshold. In addition to the sift stopping criterion, the method of calculating the instantaneous frequency and frequency scale are also parameters that must be selected when computing an HHT. Instantaneous frequency can be calculated using several methods, including Hilbert transform, normalized Hilbert transform (NHT), and using the quadrature directly [27].

2.1.4. Winger-Ville distribution

The WVD is formed by calculating the Wigner Distribution on the analytic signal $z(t)$ of a real signal $s(t)$. The WVD seemingly violates the uncertainty principle but suffers from cross terms in multicomponent signals. The WVD is defined by the Eq. (3) [28]:

$$W_z(t,f) = \int_{-\infty}^{\infty} z\left(t+\frac{\tau}{2}\right)z^*\left(t-\frac{\tau}{2}\right)\exp\left(-\mathrm{j}2\pi f\tau\right)\mathrm{d}\tau \tag{3}$$

where the superscript asterisk (*) denotes the complex conjugate.

2.2. Bayesian optimization

Bayesian optimization is a powerful optimization technique that allows for the joint optimization of configuration parameters. The goal of Bayesian optimization is to select parameters that minimize an objective function while testing as few combinations of parameters as possible. Because the number of evaluations is minimized, it is a useful method when the objective function is expensive to evaluate. Bayesian optimization revolves around a surrogate and acquisition function. The surrogate function is the approximation of

Table 1
Search space of TFRs and parameters.

TFR	Parameter (units)	Values
STFT	Window Length (samples)	[16, 1024]
	Window Overlap (%)	[0, 99]
	Window Function	'boxcar', 'hann'
DWT	Daubechies	'db1','db2', …, 'db10'
	Coiflets	'coif1', 'coif2', …, 'coif10'
	Symmlets	'sym2', 'sym3', …, 'sym10'
	Biorthogonal	'bior1.1', 'bior2.2', 'bior3.3', 'bior4.4', 'bior5.5', 'bior6.8'
HHT	Sift Stop Threshold	[1e-9, 1e-7]
	Frequency method	'Hilbert', 'quadrature', 'NHT'
	Frequency scale	'linear', 'log'
WVD		

the objective function. The acquisition function is used to probabilistically determine where to evaluate the objective function. The optimization is performed in an iterative fashion. At each iteration, the point from the acquisition function that is most likely to minimize the objective function is evaluated. The surrogate function is refit with the additional datapoint, and the acquisition function is used to evaluate the next point to test [29].

3. Methodology

In this work, a methodology for selecting a TFR to represent a dataset is introduced and tested. Specifically, the STFT, DWT, HHT, and WVD are evaluated. Each of these TFRs, except for the WVD, have their own set of parameters, which are described in the Background section. The values for each parameter that were explored in this work are listed in Table 1. To evaluate the fidelity of the parameters, Rényi entropy was used. The Rényi entropy was chosen because of its use in prior work and ability to handle negative values, which appear in Cohan's class [30]. The Rényi entropy (H_α) treats the TFR (p) as a joint probability distribution and measures the concentration of information as shown in Eq. (4) [31]:

$$H_\alpha(p) = \frac{1}{1-\alpha}\log_2\sum_i p_i^\alpha \tag{4}$$

where $\alpha > 0$ and $\alpha \neq 0$. The α term represents the order of the Rényi entropy. In this work, the third-order Rényi entropy was used because it is defined for a broad class of TFRs [32].

To reduce the computation time required to explore the search space defined by the TFRs and their parameters, Bayesian optimization was investigated. Bayesian optimization is chosen because it has been shown to meet or surpass an expert human at tuning learning parameters and model hyperparameters for machine learning modeling [33]. Because the selection of a TFR and its parameters parallels the selection of a machine learning model and its hyperparameters, it is hypothesized that Bayesian optimization can be used to automate the selection of a TFR. Bayesian optimization selects the optimal time-frequency parameters by forming a surrogate objective function and an acquisition function that decides which parameters to test based on the greatest probability of improving the objective function. The quality of the selected TFR is evaluated, the objective function is updated, and the acquisition function selects the next set of parameters to evaluate [34]. When performing Bayesian optimization, the user must select the maximum number of iterations to perform. The convergence of the optimization can be observed by monitoring the cumulative minimum loss. The cumulative minimum must be used instead of just the most recent loss because the optimization will continue to explore un-explored areas, so the most recent loss may not be the lowest. In this work, 20 iterations were performed because this was found to be sufficient for convergence.

Differences between the TFR types make it infeasible to directly compare the entropy results between TFR types. For example, the WVD is known to suffer from cross terms for multicomponent signals. This will inherently increase the entropy but has been suggested to aid in classification accuracy [8,35]. For this reason, Bayesian optimization is performed on each TFR type separately. This results in four TFRs. To determine the best TFR to represent the dataset from these four TFRs, the similarity within the optimization dataset samples is used. The similarity is calculated as the average mean squared error (MSE) between all combinations. To account for variations in scale across the different TFR classes, each TFR is scaled by its maximum value. The optimal TFR (TFR_o) is thus determined by the Eq. (5):

$$\mathrm{TFR}_o = \underset{a}{\operatorname{argmin}} \sum_{j=i+1}^{n}\sum_{i=0}^{n} \mathrm{MSE}\left(\frac{\mathrm{TFR}_{a,i}}{\max\left(\mathrm{TFR}_a\right)}, \frac{\mathrm{TFR}_{a,j}}{\max\left(\mathrm{TFR}_a\right)}\right) \tag{5}$$

where n is the number of samples in the optimization dataset and a are the four different TFR types.

Using an entire dataset in the optimization process would be time and resource consuming, so instead, a randomly chosen subset of samples were used. Specifically, a random subset of the normal class is used. This is done because, in industrial fault monitoring, data for all fault cases may not be initially available, so this work aims to select a TFR based only on the normal class. The complete process is summarized in Fig. 1. After the TFR and parameters are chosen from the random samples, it is applied to the full dataset. To verify

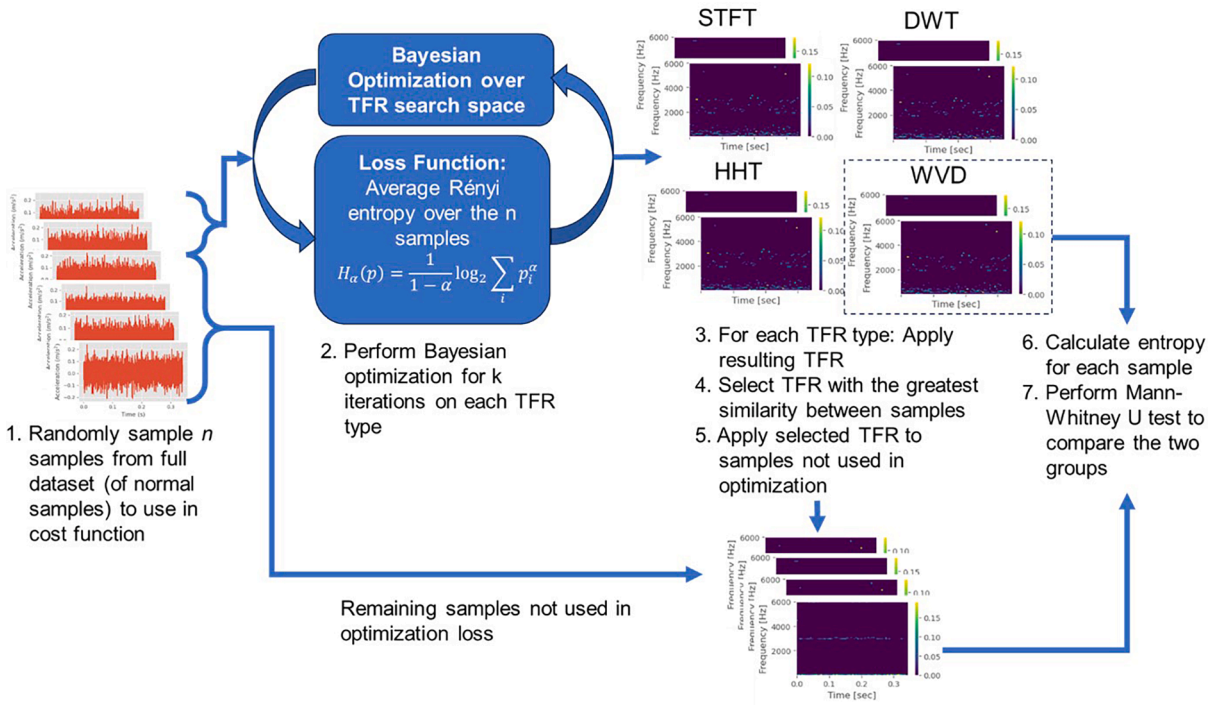

Fig. 1. Summary of the optimal TFR selection methodology.

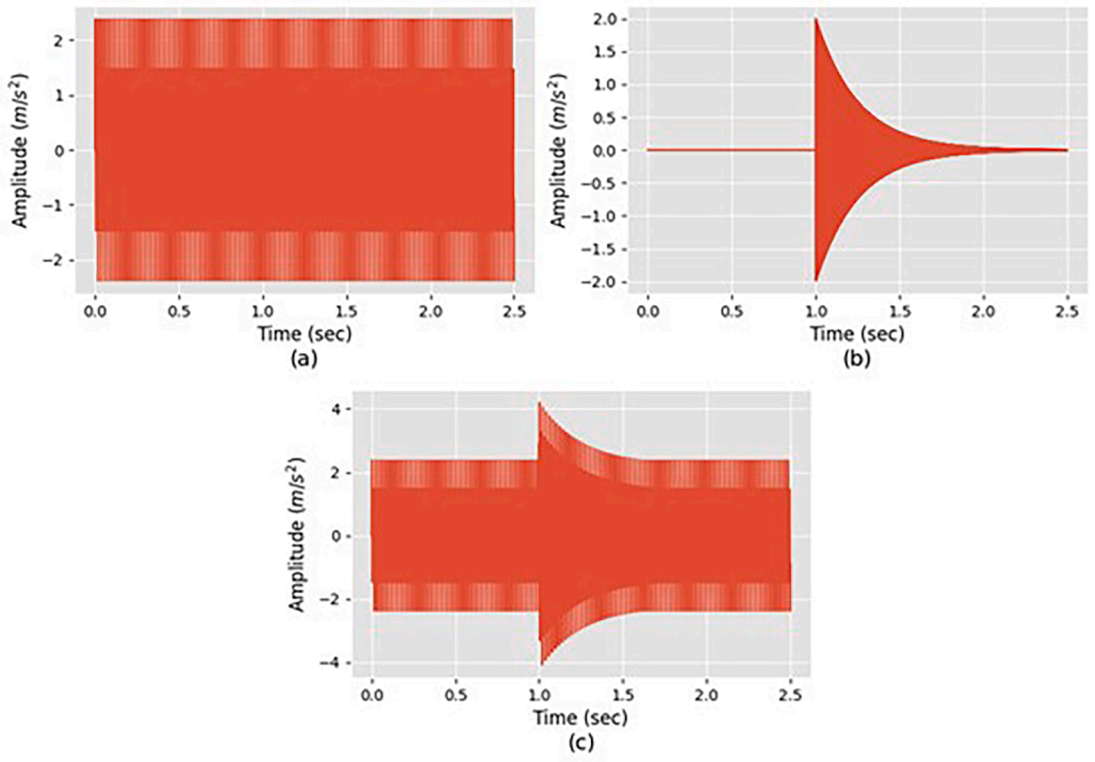

Fig. 2. Example signal comprised of (a) a base signal with (b) an impulse added to it to get the (c) resulting signal.

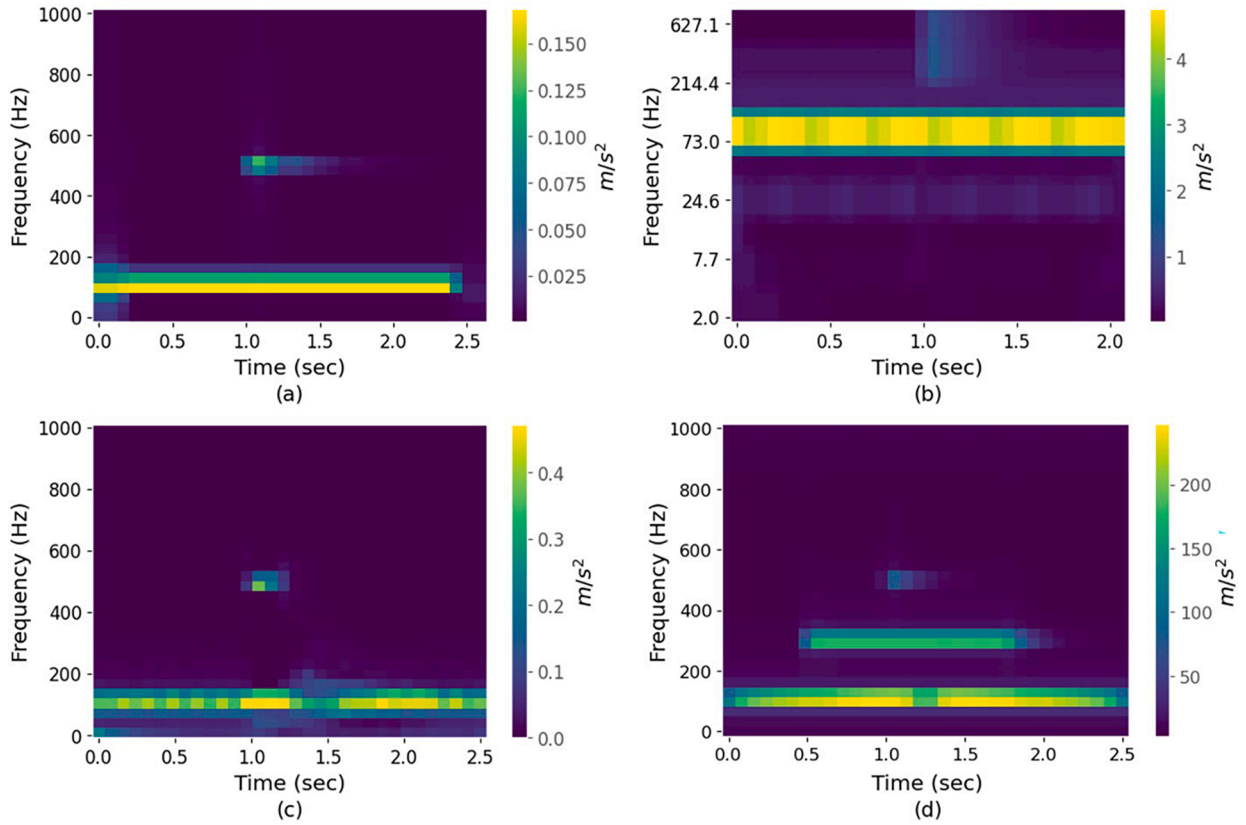

Fig. 3. The result of applying the TFR selection methodology to an example signal demonstrates that the proposed methodology can select appropriate parameters for each TFR type: (a) STFU, (b) DWT, (c) HHT, and (d) WVD.

that the subset of the data used in the optimization dataset was large enough to represent the full dataset, a Mann-Whitney U test is performed on the entropies of the resulting TFRs. Each TFR results in different dimensions based on the parameters chosen. In this work, the output of each TFR is resized to be 32×32 pixels so that the same CNN models can be used. Down sampling is done by averaging over a sliding window. The 32×32 size allows for smaller CNN models to be generated and represents a reduction in the size of the data compared to its raw format.

To illustrate the methodology, it is applied to an example signal comprised of a base signal and a disturbance. The base signal (x_{base}) contains two components as described in Eq. (6), where t is sampled on the interval [0,2.5] at 2000 Hz. A disturbance in the form of an exponentially decaying impulse signal ($x_{\text{disturbance}}$) is added to the base signal starting at $t = 1$. The exponentially decaying impulse signal is defined in Eq. (7). Plots of these signals are shown in Fig. 2. For illustrative purposes, this is a vibration signal with units of m/s^2.

$$x_{\text{base}} = 1.5 * \sin(2 * \pi * 100 * t) + 1.0 * \sin(2 * \pi * 150 * t) \tag{6}$$

$$x_{\text{disturbance}} = 2 * e^{-4*t} * \sin(2 * \pi * 500 * t) \tag{7}$$

The result of applying the proposed TFR selection methodology to the example signal is shown in Fig. 3. In this example, only the first part of the proposed methodology is performed where Bayesian optimization is used to select parameters for each TFR. Additionally, because there is only a single sample, the entropy of this sample is minimized as opposed to the average entropy over the optimization dataset. The resulting TFRs show that all capture the nature of the example signal in that the base signal and impulse are both represented with good localization. Fig. 3c also shows the cross-term that appears in the WVD between the base signal and the impulse. This cross-term causes the entropy of the WVD distribution to be one of the higher ones in the group at 13.3 bits compared to 12.4 bits for the STFT, 13.9 bits for the DWT, and 12.8 bits for the HHT. The cross-term, however, has been suggested to benefit classification tasks [8,35], and thus this example helps to illustrate the use of the similarity as opposed to the use of entropy for selecting which TFR to use.

To evaluate the proposed methodology, and the resulting TFRs, two case studies were performed. This is done because there is currently no objective methodology that can be used to compare the results against. Instead, the selected TFRs are used as inputs to simple CNN models for each case study. If the models which use the TFRs produce good accuracies on their respective datasets, it is concluded that the TFR appropriately represents the dataset.

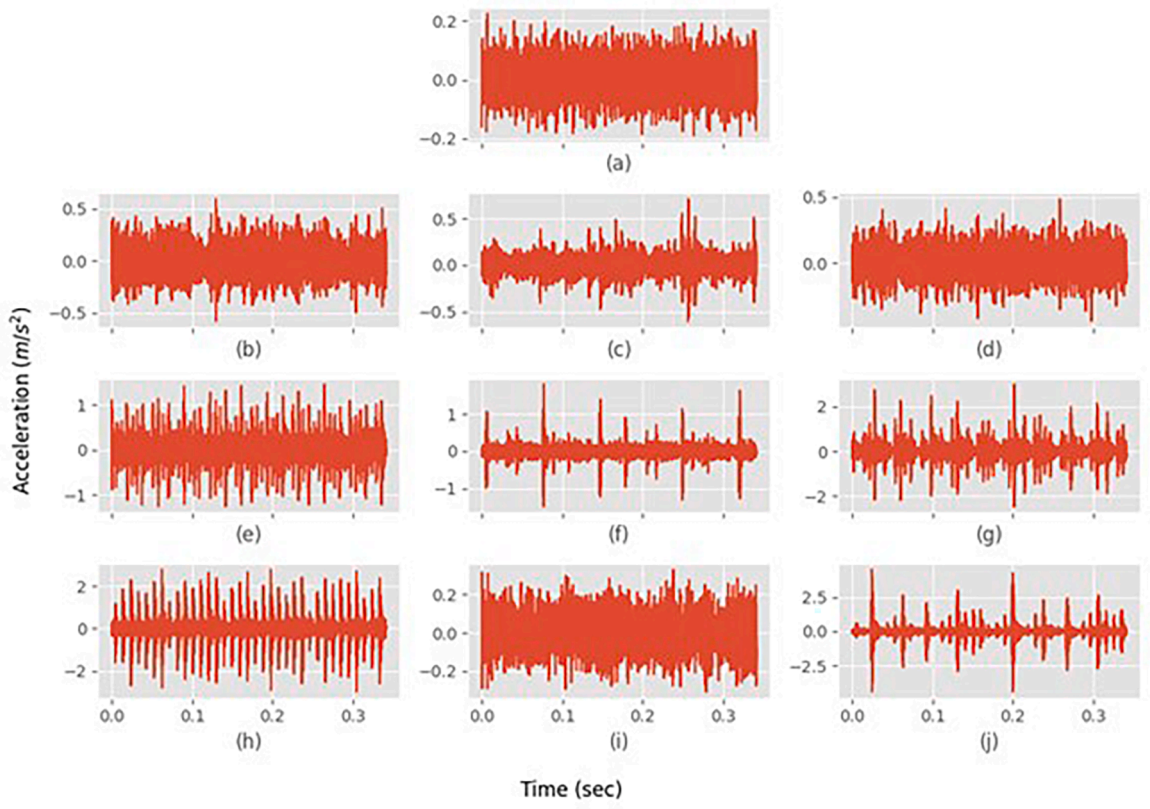

Fig. 4. Example samples for each of the CWRU bearing dataset fault classes including (a) Normal, (b) 0.0178 mm B, (c) 0.0356 mm B, (d) 0.0533 mm B, (e) 0.0178 mm IR, (f) 0.0356 mm IR, (g) 0.0533 mm IR, (h) 0.0178 mm OR, (i) 0.0356 mm OR, and (j) 0.0533 mm OR sampled under a load of 1.47 kW.

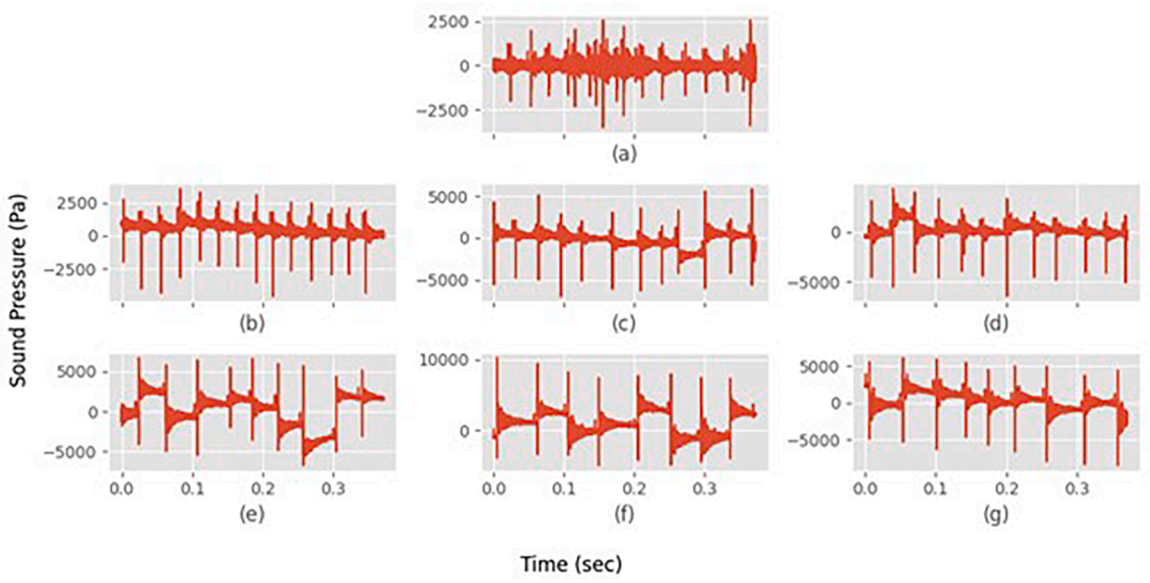

Fig. 5. Example samples from the WAAM dataset for the (a) normal CTWD and deviations from normal of (b) 2.54, (c) 5.08, (d) 7.62, (e) 10.16, (f) 12.7, and (g) 15.24 mm.

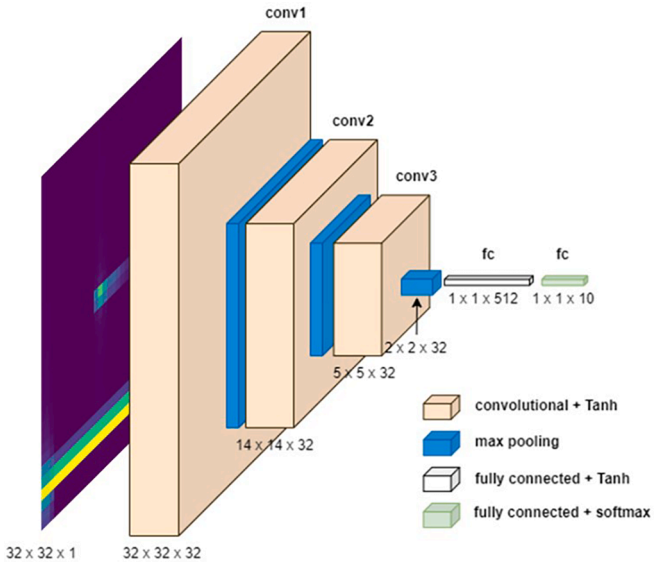

Fig. 6. Convolutional neural network architecture used to train models from a single TFR.

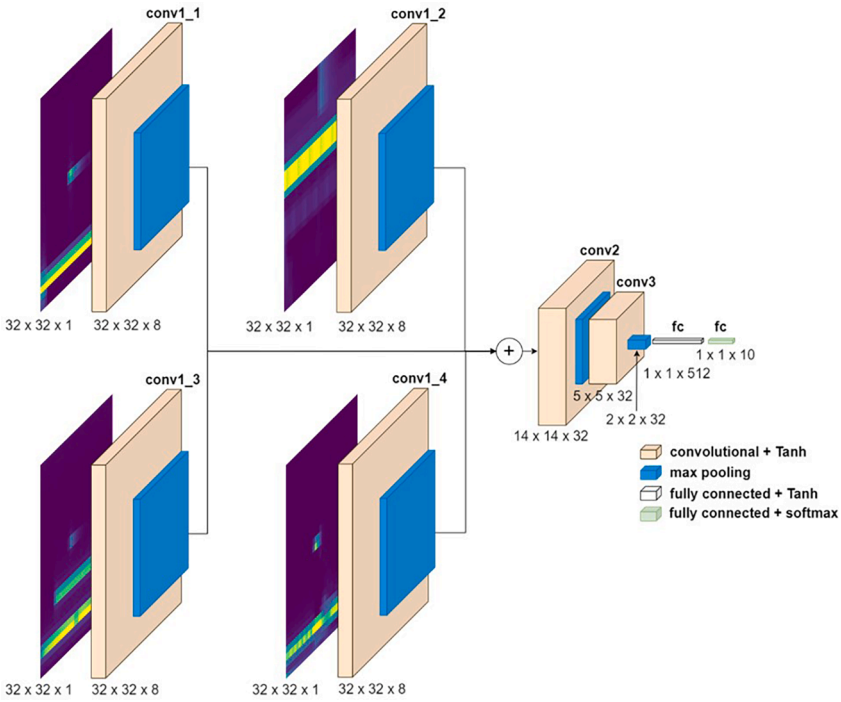

Fig. 7. Composite convolutional neural network architecture used to train a single model from four TFRs.

4. Case studies

The two datasets used for the case studies allowed the methodology to be tested on both vibration and acoustic data. The first case study was done on the Case Western Reserve University (CWRU) bearing dataset, which contains vibration data for three types of seeded faults and three sizes each [36]. Seeded faults are introduced to the inner raceway (IR), outer raceway (OR), and rolling elements (B). The seeded vaults are of sizes 0.0178, 0.0356, and 0.0533 mm (0.007, 0.014, and 0.021 inches). Each fault size is applied to each component resulting in nine fault conditions. Including the normal condition, ten classes are present. For each of these ten conditions, vibration data is collected for the system under four load conditions. The loads tested are 0, 0.735, 1.47, and 2.21 kW (0, 1, 2, and 3 horsepower), and data is sampled at 12 kHz. The problem explored with this dataset is to classify the fault condition using the

Table 2
CWRU results for individual and composite models.

TFR	Entropy	MSE within Normal	CNN Avg Test Accuracy
STFT	15.0	0.0045	1.0
DWT	15.1	0.013	0.98
HHT	13.1	0.0043	0.95
WVD	15.3	0.0037	1.0
Composite	–	–	1.0

Table 3
WAAM Results for individual and composite models.

TFR	Entropy	MSE within Normal	CNN Avg Test Accuracy
STFT	15.4	0.0011	0.94
DWT	14.9	0.0022	0.41
HHT	11.7	0.0027	0.23
WVD	15.4	0.0089	0.23
Composite	–	–	0.92

vibration data. An example of a 0.34 s sample of vibration data for each class is shown in Fig. 4.

The second case study was performed on an acoustic dataset for monitoring the contact tip to work-piece distance (CTWD) in wire arc additive manufacturing (WAAM) [37]. This dataset contains acoustic data for seven different CTWD values. Maintaining a constant CTWD is important for the quality of WAAM parts, but cumulative errors from modeling inaccuracies and thermal conditions can cause the CTWD to increase during a build. The seven CTWD values include the baseline, which represents the healthy value, and six fault conditions that deviate from normal by 2.54 mm (0.1 inches) each step. The problem explored with this dataset is to classify the CTWD using the acoustic data. An example of a 0.37 s sample of acoustic data for each is shown in Fig. 5.

4.1. Evaluation

To evaluate the effectiveness of the generated TFRs, a simple CNN architecture was used to create trained models of each TFR. A CNN was used because it is well suited to handle the TFRs as input and has been shown to be successful for intelligent fault diagnostics [38]. The model architecture used is based on the LeNet-5 architecture and is comprised of three pairs of convolutional and max pool layers [39]. Convolutional kernels are of size (3,3) and the max pool kernel is (2,2); the convolutional and max pool layers both use a stride of (1,1). The first convolutional layer is zero-padded while the following convolutional layers are not padded. After the last max pool layer, there is a single fully connected layer (size 512) before the output layer. All hidden layers use the hyperbolic tangent activation function. The output layer for the CWRU bearing dataset contains ten neurons for the nine fault classes and one normal class. The output for the WAAM dataset is seven for the seven different CTWD values tested. In total, this results in models with 89,994 trainable parameters for CWRU models and 88,455 trainable parameters for WAAM models. This architecture is illustrated in Fig. 6.

In addition to generating a CNN for each TFR type, a single CNN was trained that takes all four TFRs as an input. Each input is the same as in the previous model that takes a single TFR as input, so the input is 4 TFRs, each of size 32×32. This is referred to as the composite model. The first convolution and max pool layer are independent for each TFR. To maintain the same total number of trainable parameters as the models for each TFR, the first convolutional layers contain a quarter of the number of filters. The outputs for each TFR are then combined before going through the following two convolution layers. This is done because the frequency axis of the TFRs will not align, so stacking TFRs like color channels in an image does not make physical sense. This architecture is illustrated in Fig. 7. While this increases the complexity of the model, it does not require a single TFR to be chosen. All models are trained for 40 epochs with categorical cross-entropy loss and the Adam optimizer. The TensorFlow default values for the Adam optimizer are used including a learning rate of 0.001, β_1 of 0.9, β_2 of 0.999, and ε of 1e-7. For both datasets, 80% of the data was used for training, and 20% was used for testing.

For the composite models, the permutation importance was calculated to determine if the importance of each TFR correlates with the performance of the single input models and the TFR similarities. Permutation importance is the decrease in model accuracy when a single input feature is randomly shuffled [40]. The permutation importance for one of the input TFRs is determined by first randomly shuffling the values of that TFR while leaving the remaining TFRs unchanged. The accuracy of the corrupted dataset is then computed. This is repeated for K iterations and the importance, i, of the TFR under investigation is calculated as shown in Eq. (8) [40]:

$$i = \mathrm{acc} - \frac{1}{K}\sum_{k=1}^{K} \mathrm{acc}_k, \tag{8}$$

where acc is the accuracy of the uncorrupted dataset, and acc_k is the accuracy of the corrupted dataset at iteration k. In this work, 30 iterations were used.

Table 4
P values for the post hoc analysis comparing the MSE similarity values.

TFR 1	TFR 2	CWRU	WAAM
STFT	DWT	<0.001	.0020
STFT	HHT	.35	<0.001
STFT	WVD	<0.001	<0.001
DWT	HHT	<0.001	<0.001
DWT	WVD	<0.001	<0.001
HHT	WVD	.0039	<0.001

4.2. Results

For both case studies, 20 samples were randomly chosen from the normal class of data to be used to select the TFR. This represents 10% of the normal class of the CWRU bearing dataset and 4.4% of the WAAM dataset. The results of applying the TFR selection method to the CWRU bearing dataset are summarized in Table 2. Results for the WAAM dataset are shown in Table 3. These tables show the result of the models for each TFR separately, as well as the composite model that takes all four TFRs as input.

The average entropies of each TFR are given in Tables 2 and 3 for reference, but these values are not intended to be compared between TFR types. Instead, the average MSE within the normal class is used to quantify the relative quality of each TFR before training models. When comparing the CNN average accuracy for the CWRU bearing dataset, it is seen that all TFRs result in a high accuracy, so any of the four TFR types could be used to represent the dataset. For the WAAM dataset, only one TFR, the STFT, results in a high accuracy from the trained CNN model. This TFR is also the one with the lowest MSE within the normal class, indicating that it is the one preferred by the introduced methodology. A Friedman test was performed on the MSE values, and the result was found to be statistically significant for both the CWRU ($p<.001$) and WAAM ($p<.001$) datasets. Post hoc analysis with Wilcoxon signed-rank tests was performed with Bonferroni correction applied, resulting in an adjusted significance level at $p<.0125$. The results of this post hoc analysis showed that the only combination of TFRs that had a did not have a statistically significant difference were the STFT and HHT for the CWRU bearing dataset. All p values are reported in Table 4.

To further illustrate the correlation between the CNN accuracy and the similarity within TFRs in the normal class (as measured by the MSE), the values are plotted in Fig. 8 for both the CWRU and WAAM datasets. For the CWRU bearing dataset, no correlation is observed, but all TFR types perform well with accuracies of at least 0.95. The WAAM dataset shows a clear downward trend until the CNN accuracy hits a floor near 0.20.

Comparing the composite models results to the results of the individual models in Tables 2 and 3 indicates that selecting a single TFR could be skipped in favor of using a model that takes multiple TFRs as input. For the CWRU dataset, the composite model matches that of the best performing individual model. For the WAAM dataset, the performance of the composite model is within 3% of the best single TFR model. The permutation importance for each TFR in the composite model are reported in Tables 5 and 6 for the CWRU and WAAM datasets respectively.

After the TFR was selected and applied to the entire dataset, a Mann-Whitney U test was performed on the entropies of the samples within and outside the optimization dataset to verify that the optimization dataset was large enough to represent the full dataset. The Mann-Whitney U test was used because some of the entropy distributions were not normal. The results, shown in Table 7, verify that for all four TFR types tested, the optimization dataset represented the normal class.

5. Discussion

The results from these case studies demonstrate that the proposed methodology is able to select TFRs that, when combined with a simple CNN, are able to perform well on their respective tasks. For the CWRU bearing dataset, the proposed methodology matches the performance of other published results, which use custom designed architectures [41]. For the WAAM architecture, the proposed methodology exceeds the performance of previous work, which used statistical features and decision trees [37]. This presents a promising result, since user input and possible user bias is reduced without a loss in model accuracy, providing a pathway for automation.

In the case studies in this work, the introduced methodology was applied to four common time-frequency methods: STFT, DWT, HHT, and WVD. These four common, and distinctly different, time-frequency methods provide an initial evaluation of the introduced methodology. Given the success in the case studies presented in this work, future work should extend this work by including a wide range of time-frequency methods and evaluating the results across additional case studies. This will help in understanding the extents of the applicability of the methodology, but it will require testing a sufficiently large set of case studies such that the applicability of each time-frequency method can be observed.

The results of the composite model indicate that multiple pathways exist for making use of the TFRs returned by the Bayesian optimization. Tables 2 and 3 show that the MSE within the optimization dataset can be used to select a single TFR to use or all TFRs can be used in a single model with minimal loss of performance, even when three of the four TFRs do not perform well on their own, as with the WAAM results. Considering the feature importance of each TFR in the composite model for the CWRU bearing dataset, Table 5 shows that the composite model has high importance on one of the inputs. Since the individual TFR models all had high accuracy, the inputs can be considered correlated, so the feature importance is less meaningful [42]. The results for the WAAM dataset shown in

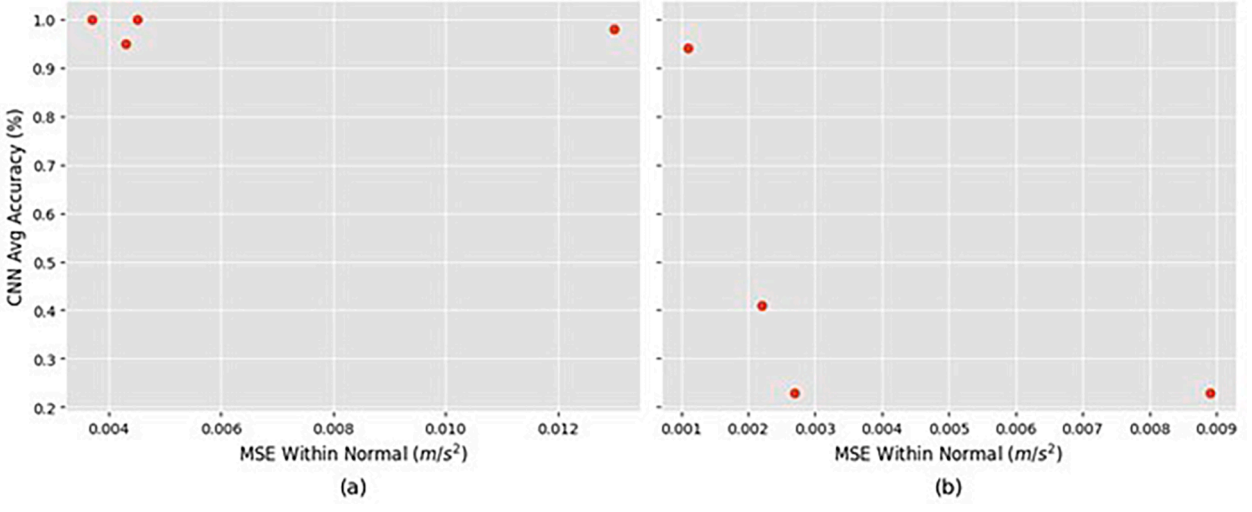

Fig. 8. Comparison of CNN average accuracy and MSE within the normal class for the (a) CWRU bearing dataset and (b) WAAM dataset.

Table 5
Feature importance of the composite model for the CWRU bearing dataset.

TFR	CNN Avg Accuracy (individual models)	Permutation Importance (composite model)
STFT	1.0	0.000
DWT	0.98	0.001
HHT	0.95	0.000
WVD	1.0	0.803

Table 6
Feature importance of the composite model for the WAAM bearing dataset.

TFR	CNN Avg Accuracy (individual models)	Permutation Importance (composite model)
STFT	0.94	0.601
DWT	0.41	0.000
HHT	0.23	0.000
WVD	0.23	0.000

Table 7
Mann-Whitney U test p-values comparing the entropies for the samples used in the optimization and those not used.

TFR	CWRU	WAAM
STFT	.52	.82
DWT	.88	.20
HHT	.34	.19
WVD	.39	.52

Table 6 show that the permutation importance corresponds with the TFR that yields the greatest accuracy in an individual model. In summary, the permutation importance of the composite model can indicate the appropriate TFR, but only if a single TFR is appropriate for a problem. When all TFRs are viable inputs, the composite model permutation importance is not effective for selecting a single TFR. In either case, the composite model could be used in place of selecting a single TFR. With new TFR methods being continually introduced [43,44], a natural extension of this work is to expand the search to include additional TFR types. Doing so may result in the composite model becoming less feasible.

Contrastive methods are another common approach for TFR selection but require multiple classes to be present. These include the discrimination between classes [45,46], and making use of the genetic algorithm [47,48]. The use of only healthy data in the proposed methodology introduces several benefits. Firstly, it enables model training to be initiated sooner, since multiple classes of data do not need to be collected. This allows for unsupervised anomaly detection to be used before additional classes are present. Additionally, TFRs can start being collected instead of raw data sooner. This allows for a reduction in the amount of data transmitted and stored, as discussed later.

Another alternative to selecting a TFR is to use the raw sensor data directly [49]. The use of the TFR instead of the raw sensor data greatly reduces the amount of data transmitted and stored. For example, in the CWRU case study, 0.34 s samples containing 4096 floating point numbers are reduced to 1024 floating point numbers in the 32×32 TFR. In the WAAM case study, 0.37 s samples containing 16,384 floating point numbers are reduced to 1024 numbers in the 32×32 TFR. These present 75% and 94% data reductions. This has implications for edge and hybrid data architectures, where the TFR can be computed locally and transmitted in place of the raw data to reduce network traffic and any downstream data storage [50].

6. Conclusion

This work introduces a methodology to automate the selection of a TFR for a dataset of sound or vibration data using only healthy data. The selection of parameters for each TFR is performed with Bayesian optimization with the average entropy used as the cost function. The optimal TFR is then chosen using the average MSE between all the normal samples used in the optimization. This serves as a measure of similarity. Because no objective method exists for TFR selection, the methodology is evaluated by performing two case studies, where the introduced TFR selection methodology is used with a simple CNN to predict bearing faults from vibration data and the contact tip to workpiece distance in a wire arc additive manufacturing process from sound data. Models trained on the resulting TFRs achieve 100% and 94% test accuracy, demonstrating that the proposed method is able to match or exceed previously published results. This research thus provides practitioners with an objective method to select a TFR to represent a dataset. This is valuable because current approaches for TFR selection require domain expertise or trial-and-error, both of which have the potential for user induced bias and do not extend well to new domains. Given the continued innovation and development of new time-frequency methodologies, the challenge of selecting a time-frequency will continue to grow, so the methodology introduced in this work presents a significant improvement to state-of-the-art approach which requires a user with domain expertise. Additionally, by only using

healthy data in the proposed methodology, failure modes do not need to be known ahead of time, and the generated TFRs can be used to reduce the amount of data generated, transmitted, and saved within a modern digital manufacturing ecosystem by as much as 94%.

CRediT authorship contribution statement

Nathaniel DeVol: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Christopher Saldaña:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Katherine Fu:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding: This work was supported by the U.S. Department of Defense (DOD), Office of Local Defense Community Cooperation (OLDCC), Industry Resilience Program, Award #ST1449–21–03 2016–2166 and Georgia AI Manufacturing (GEORGIA -AIM) Technology Corridor, Award #AWD-004285.

Data availability

The authors do not have permission to share data.

References

[1] R.A. Khalil, N. Saeed, M. Masood, Y.M. Fard, M. Alouini, T.Y. Al-Naffouri, Deep learning in the industrial internet of things: potentials, challenges, and emerging applications, IEEe Internet Things J. 8 (2021) 11016–11040, https://doi.org/10.1109/JIOT.2021.3051414.
[2] R.Y. Zhong, X. Xu, E. Klotz, S.T. Newman, Intelligent manufacturing in the context of industry 4.0: a review, Engineering 3 (2017) 616–630, https://doi.org/10.1016/J.ENG.2017.05.015.
[3] N. Tandon, A. Choudhury, A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings, Tribol. Int. 32 (1999) 469–480, https://doi.org/10.1016/S0301-679X(99)00077-8.
[4] C.M. Bishop, Neural Networks For Pattern Recognition, Oxford University Press, Cambridge, 1995.
[5] E. Zio, Prognostics and Health Management (PHM): where are we and where do we (need to) go in theory and practice, Reliab. Eng. Syst. Saf. 218 (2022) 108119, https://doi.org/10.1016/j.ress.2021.108119.
[6] L. Cohen, Time-frequency distributions-a review, Proc. IEEE 77 (1989) 941–981, https://doi.org/10.1109/5.30749.
[7] E. Sejdić, I. Djurović, J. Jiang, Time–frequency feature representation using energy concentration: an overview of recent advances, Digit. Signal Process. 19 (2009) 153–183, https://doi.org/10.1016/j.dsp.2007.12.004.
[8] Z. Feng, M. Liang, F. Chu, Recent advances in time–frequency analysis methods for machinery fault diagnosis: a review with application examples, Mech. Syst. Signal. Process. 38 (2013) 165–205, https://doi.org/10.1016/j.ymssp.2013.01.017.
[9] D. Verstraete, A. Ferrada, E.L. Droguett, V. Meruane, M. Modarres, Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings, Shock Vib. 2017 (2017) 1–17, https://doi.org/10.1155/2017/5067651.
[10] B. Boashash, Time-Frequency Signal Analysis Methods and Applications, Longman Cheshire, Melbourne, 1992.
[11] R.X. Gao, R. Yan, Non-stationary signal processing for bearing health monitoring, Int. J. Manuf. Res. 1 (2006) 18–40, https://doi.org/10.1504/IJMR.2006.010701.
[12] B.S. Kim, S.H. Lee, M.G. Lee, J. Ni, J.Y. Song, C.W. Lee, A comparative study on damage detection in speed-up and coast-down process of grinding spindle-typed rotor-bearing system, J. Mater. Process. Technol. 187–188 (2007) 30–36, https://doi.org/10.1016/j.jmatprotec.2006.11.222.
[13] H. Li, T. Liu, X. Wu, Q. Chen, An optimized VMD method and its applications in bearing fault diagnosis, Measurement 166 (2020) 108185, https://doi.org/10.1016/j.measurement.2020.108185.
[14] Y. Cheng, S. Wang, B. Chen, G. Mei, W. Zhang, H. Peng, G. Tian, An improved envelope spectrum via candidate fault frequency optimization-gram for bearing fault diagnosis, J. Sound Vib. 523 (2022) 116746, https://doi.org/10.1016/j.jsv.2022.116746.
[15] J. Yan, Y. Cheng, Q. Wang, L. Liu, W. Zhang, B. Jin, Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts, IEEe Trans. Emerg. Top. Comput. Intell. 8 (2024) 2827–2842, https://doi.org/10.1109/TETCI.2024.3377728.
[16] J. Lin, L. Qu, Feature extraction based on morlet wavelet and its application for mechanical fault diagnosis, J. Sound Vib. 234 (2000) 135–148, https://doi.org/10.1006/jsvi.2000.2864.
[17] P.K. Kankar, S.C. Sharma, S.P. Harsha, Fault diagnosis of ball bearings using continuous wavelet transform, Appl. Soft Comput. J. 11 (2011) 2300–2312, https://doi.org/10.1016/j.asoc.2010.08.011.
[18] R. YAN, R.X. GAO, Base wavelet selection for bearing vibration signal analysis, Int. J. Wavelets Multiresolution Inf. Process. 07 (2009) 411–426, https://doi.org/10.1142/S0219691309002994.
[19] Y. Jiang, B. Tang, Y. Qin, W. Liu, Feature extraction method of wind turbine based on adaptive Morlet wavelet and SVD, Renew. Energy 36 (2011) 2146–2153, https://doi.org/10.1016/j.renene.2011.01.009.
[20] T.H. Sang, W.J. Williams, Renyi information and signal-dependent optimal kernel design, in: Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, IEEE, 1995, pp. p997–1000, https://doi.org/10.1109/ICASSP.1995.480344.
[21] W.K. Ngui, M.S. Leong, L.M. Hee, A.M. Abdelrhman, Wavelet analysis: mother wavelet selection methods, Appl. Mech. Mater. 393 (2013) 953–958, https://doi.org/10.4028/www.scientific.net/AMM.393.953.
[22] R. Yan, R.X. Gao, Impact of wavelet basis on vibration analysis for rolling bearing defect diagnosis, in: Proceedings of the 2011 IEEE International Instrumentation and Measurement Technology Conference, Hangzhou, IEEE, 2011, pp. p1–p4, https://doi.org/10.1109/IMTC.2011.5944209.
[23] D. Jiang, C. Liu, Machine condition classification using deterioration feature extraction and anomaly determination, IEEe Trans. Reliab. 60 (2011) 41–48, https://doi.org/10.1109/TR.2011.2104433.
[24] R.B. Randall, Vibration–Based Condition Monitoring, 2nd ed., Wiley, Hoboken, NJ, 2021 https://doi.org/10.1002/9781119477631.

[25] P.S. Addison, The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance, Second Edition, 2nd ed., CRC Press, Boca Raton, 2017 https://doi.org/10.1201/9781315372556.
[26] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Snin, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. A Math. Phys. Eng. Sci. 454 (1998) 903–995, https://doi.org/10.1098/rspa.1998.0193.
[27] N.E. Huang, Z. Wu, S.R. Long, K.C. Arnold, X. Chen, K. Blank, On instantaneous frequency, Adv. Adapt. Data Anal. 1 (2009) 177–229, https://doi.org/10.1142/S1793536909000096.
[28] B. Boashash, Time-frequency signal analysis and processing - a comprehensive reference. Time-Frequency Signal Analysis and Processing, 2nd ed., Elsevier, San Diego, 2016 https://doi.org/10.1016/B978-0-12-398499-9.09983-0.
[29] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. de Freitas, Taking the human out of the loop: a review of Bayesian optimization, Proc. IEEE 104 (2016) 148–175, https://doi.org/10.1109/JPROC.2015.2494218.
[30] W.J. Williams, M.L. Brown, A.O. Hero III, Uncertainty, information, and time-frequency distributions, in: F.T. Luk (Ed.), Advanced Signal Processing Algorithms, Architectures, and Implementations Ii, San Diego, 1991: p144–156. 10.1117/12.49818.
[31] A. Rényi, On measures of entropy and information, in: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1961, pp. 547–562.
[32] R.G. Baraniuk, P. Flandrin, A.J.E.M. Janssen, O.J.J. Michel, Measuring time-frequency information content using the Renyi entropies, IEEe Trans. Inf. Theory 47 (2001) 1391–1409, https://doi.org/10.1109/18.923723.
[33] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst., Lake Tahoe (2012).
[34] P.I. Frazier, A tutorial on Bayesian optimization, ArXiv Preprint ArXiv:1807.02811 (2018) 1–22.
[35] B. Boashash, G. Azemi, N.A Khan, Principles of time–frequency feature extraction for change detection in non-stationary signals: applications to newborn EEG abnormality detection, Pattern. Recognit. 48 (2015) 616–627, https://doi.org/10.1016/j.patcog.2014.08.016.
[36] W.A. Smith, R.B. Randall, Rolling element bearing diagnostics using the case western reserve university data: a benchmark study, Mech. Syst. Signal Process. 64–65 (2015) 100–131, https://doi.org/10.1016/j.ymssp.2015.04.021.
[37] D. Hebert, A. Thien, C. Saldana, Acoustic process monitoring of contact tip to work piece distance in wire arc additive manufacturing by random forest algorithms, in: Additive Manufacturing; Biomanufacturing; Life Cycle Engineering; Manufacturing Equipment and Automation; Nano/Micro/Meso Manufacturing, 1, American Society of Mechanical Engineers, West Lafayette, 2022, pp. p1–p6, https://doi.org/10.1115/MSEC2022-85404.
[38] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A.K. Nandi, Applications of machine learning to machine fault diagnosis: a review and roadmap, Mech. Syst. Signal Process. 138 (2020) 106587, https://doi.org/10.1016/j.ymssp.2019.106587.
[39] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2323, https://doi.org/10.1109/5.726791.
[40] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[41] S. Zhang, S. Zhang, B. Wang, T.G. Habetler, Deep learning algorithms for bearing fault diagnostics - a comprehensive review, IEEe Access 8 (2020) 29857–29881, https://doi.org/10.1109/ACCESS.2020.2972859.
[42] C. Strobl, A. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinform. 8 (2007) 25, https://doi.org/10.1186/1471-2105-8-25.
[43] B. Chen, Z. Hai, X. Chen, F. Chen, W. Xiao, N. Xiao, W. Fu, Q. Liu, Z. Tian, G. Li, A time-varying instantaneous frequency fault features extraction method of rolling bearing under variable speed, J. Sound Vib. 560 (2023) 117785, https://doi.org/10.1016/j.jsv.2023.117785.
[44] P. Wang, Z. Wu, C. Yang, A modified frequency–time domain method for nonlinear aeroelastic systems with initial conditions, J. Sound Vib. 566 (2023) 117899, https://doi.org/10.1016/j.jsv.2023.117899.
[45] Y.H. Wu, M.X. Shan, Y.N. Qian, X.L. Li, R.Q. Yan, Aeroengine rub-impact fault diagnosis based on wavelet packet transform and the local discriminate bases, Appl. Mech. Mater. 226–228 (2012) 740–744, https://doi.org/10.4028/www.scientific.net/AMM.226-228.740.
[46] Y. Mallet, D. Coomans, J. Kautsky, O. De Vel, Classification using adaptive wavelets for feature extraction, IEEe Trans. Pattern. Anal. Mach. Intell. 19 (1997) 1058–1066, https://doi.org/10.1109/34.625106.
[47] J. Rafiee, P.W. Tse, A. Harifi, M.H. Sadeghi, A novel technique for selecting mother wavelet function using an intelli gent fault diagnosis system, Expert. Syst. Appl. 36 (2009) 4862–4875, https://doi.org/10.1016/j.eswa.2008.05.052.
[48] W.J. Staszewski, Vibration data compression with optimal wavelet coefficients, in: Proceedings of the 2nd International Conference on Genetic Algorithms in Engineering Systems, IEEE, 1997, pp. p186–p190, https://doi.org/10.1049/cp:19971178.
[49] L. Wen, X. Li, L. Gao, Y. Zhang, A new convolutional neural network-based data-driven fault diagnosis method, IEEE Trans. Ind. Electron. 65 (2018) 5990–5998, https://doi.org/10.1109/TIE.2017.2774777.
[50] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, J. Zhang, Edge intelligence: paving the last mile of artificial intelligence with edge computing, Proc. IEEE 107 (2019), https://doi.org/10.1109/JPROC.2019.2918951.