Simulation-driven Deep Classification of Bearing Faults from Raw Vibration Data

Martin Hemmer[1], Andreas Klausen[2], Huynh van Khang[3], Kjell G. Robbersmyr[4], and Tor I. Waag[5]

[1,2,3,4] *Department of Engineering Sciences, University of Agder, 4877 Grimstad, Norway*
martin.hemmer@uia.no
andreas.klausen@uia.no
huynh.khang@uia.no
kjell.g.robbersmyr@uia.no

[1] *MHWirth AS, Butangen 21, 4639 Kristiansand, Norway*
martin.hemmer@mhwirth.com

[5] *NORCE Norwegian Research Centre AS, 4877 Grimstad, Norway*
towa@norceresearch.no

ABSTRACT

The industry is moving towards maintenance strategies that consider component health, which require extensive collection and analysis of data. Condition monitoring methods that require manual feature extraction and analysis, become infeasible on an industrial scale. Machine learning algorithms can be used to automatically detect and classify faults, however, obtaining sufficient data for training is required for deep learning and other data-driven classification approaches. Data from healthy machine operation is generally available in abundance, while data from representative fault- and operating conditions is limited. This limits both development and deployment of deep learning-based CM systems on an industrial scale. This paper addresses both the challenges of automated analysis and lack of training data. A deep learning classifier architecture utilizing 1-dimensional dilated convolutions is proposed. Dilation of the convolution kernel allows for analysis of raw vibration signals while simultaneously maintaining the receptive field of the classifier enough to capture temporal patterns. The proposed method performs classification in time domain on signal segments of 1 second or shorter. With knowledge of the bearing specification, artificial vibration signals with similar characteristics as an actual bearing fault can be created. In this work, generated fault signals are combined with healthy operational data to obtain training data for a deep classifier. Parameters of the vibration model is chosen as distributions rather than fixed values. By using a range parameters in the vibration model, the classifier learns to recognize temporal features from the training data that generalize to unseen data. The effectiveness of the proposed method is demonstrated by training classifiers on generated data and testing on real signals from faulty bearings at both low and high speed. One dataset containing seeded faults and three run-to-failure tests are used for the demonstration.

1. INTRODUCTION

Rolling element bearings (REBs) play a fundamental part in most types of rotating machinery by reducing friction, transferring forces and constraining motion. Even under operation within design specification, REBs have a finite lifetime due to fatigue induced by cyclic loading from the moving rolling elements. Expected fatigue lifetime can be estimated accurately for large populations of bearings under identical operating conditions. However, the lifetime of a single bearing might be shorter or longer than the estimated value. Therefore, performing maintenance based on elapsed time or operating hours is not optimal. Shorter maintenance intervals do not utilize component lifetime, while longer intervals may result in unexpected failures and downtime. This motivates the development of condition monitoring (CM) methods that are able to diagnose and evaluate component health. Condition based maintenance (CBM) and prognostics and health management (PHM) maintenance regimes aim to utilize the knowledge of machine health to perform maintenance when required.

While time domain features such as root mean square (RMS), peak-to-peak and kurtosis are useful for fault detection through trending, they cannot be used for diagnosing fault types. Single-point surface defects in bearings can be modeled as periodic excitation of the system resonance frequency (McFadden & Smith, 1984). Amplitude demodulation is effective for detecting such defects, and a fast implementation is possible using the Hilbert envelope (Marple, 1999). To improve the effectiveness of envelope analysis, the signal should be preprocessed. The fast Kurtogram (Antoni, 2007b) utilizes spectral kurtosis to identify a frequency band for demodulation (Antoni, 2006), and order tracking reduce the amount of spectral leakage due to speed variations (Fyfe & Munck, 1997; Randall & Antoni, 2011). Bearing vibration is random in nature (Antoni & Randall, 2002), and is thus separable from other vibration sources using time synchronous averaging (TSA) (Bechhoefer & Kingsley, 2009; Hecke, Yoon, & He, 2016), cepstral editing (Randall & Sawalhi, 2011; Borghesani, Pennacchi, Randall, Sawalhi, & Ricci, 2013) and other methods (Randall, Sawalhi, & Coats, 2011). While these tools are effective for bearing fault diagnosis at any speed, diagnosis of low speed applications is more challenging. There is less energy in the system and the fault signature is easily masked in noise. Characteristic fault frequencies are also closer in absolute frequency, requiring longer acquisition time to achieve a spectral resolution where faults are distinguishable (Bechhoefer, Schlanbusch, & Waag, 2016). In any case, a skilled data analyst is required to process data and diagnose the bearing.

With the emergence of industry 4.0, the amount of available sensor data is increasing rapidly (Diez-Olivan, Del Ser, Galar, & Sierra, 2019). Proper usage of the aforementioned tools requires feature engineering and skilled manpower, which leaves manual analysis infeasible on a large scale and motivates a more data-driven, automated approach. Deep learning has proven to be well suited to data-driven feature extraction, classification and prognostics for rotating machinery (Zhao et al., 2019). Autoencoders (X. Li, Liu, Qu, & He, 2018), deep neural networks (Jia, Lei, Guo, Lin, & Xing, 2018), support vector machines (X. Li, Yang, Pan, Cheng, & Cheng, 2019), deep belief networks (Gan, Wang, & an Zhu, 2015; Shao et al., 2018), self-organizing maps (He & He, 2017) and convolutional neural networks (CNNs) (Guo, Chen, & Shen, 2016; G. Li et al., 2019; Jiang, He, Yan, & Xie, 2019) are all applied to bearing fault detection and diagnostics. Time series analysis using deep learning has applications in speech recognition and modeling (Hinton et al., 2012), translation (Wu et al., 2014), and audio recognition and generation (Van Den Oord et al., 2016). In high resolution time series, it is necessary to consider dependencies that are further apart than

the kernel size. One strategy is to use larger kernels and add more convolutional and pooling layers, but this is computationally heavy, and the weight training may be difficult due to vanishing gradients. Long short-term memory recurrent networks are also commonly applied in analysis and modeling of speech and audio (Sak, Senior, & Beaufays, 2014). The memory of previous states in recurrent network has to propagate through the entire time series, which is a limitation for retaining long-term memory.

However, a common problem is the need for representative fault data during training. To counteract this, it has been proposed to use simulated bearing faults for training classifiers (Ho & Randall, 2000; Sawalhi & Randall, 2008; Sobie, Freitas, & Nicolai, 2018). A wide range of analytical and FEM-based dynamic models for bearing faults exist, taking clearances, elastohydrodynamic lubrication effects, race waviness, defect size and several other parameters into consideration (X. Li et al., 2019). However, the results presented in this paper show that modeling the impulse responses and pseudo-cyclostationary behavior (Antoni, 2009; Randall & Antoni, 2011) is sufficient to capture the characteristics of a real vibration signal.

This paper proposes to train a deep convolutional neural network on vibration time series data based on combinations of healthy and simulated fault data. The architecture is based on hierarchically dilated 1D convolutions. A dilated architecture was applied to bearing fault detection in (Khan, Kim, & Choo, 2018). However, the proposed method in this paper includes low-level feature extraction without dilation, and a global average pooling layer to replace fully connected layers. Additionally, training in (Khan et al., 2018) was done on actual fault data as opposed to simulated data in this paper. Varying the parameters of the simulation model allows the classifier to learn features that are generalized enough to detect and diagnose actual bearing faults. The approach reduces the need for manual feature extraction and allows the analysts to focus the attention on detected faults for closer evaluation. The proposed method is applied to shorter signal segments of raw vibration data. This is particularly beneficial for low speed bearings that require longer acquisition time for spectral resolution. The proposed method reduces the need for data processing, storage and transfer by performing classification in time domain.

The remaining of the paper is organized as follows: Section 2 describes the generation and processing of data that leads to the classification. The datasets used in analysis are presented in Section 3. Analysis results and discussion are provided in Section 4. Final conclusions are given in Section 5.

2. METHODOLOGY

It is assumed that single-point surface defects excite transient pulses of amplitude modulated resonance frequencies of the support structure when internal bearing parts roll over a fault. Modelling a signal requires several assumptions of unknown or uncertain variables. The proposed approach acknowledges this, and generates data from parameters in a range rather than fixed values. Section 2.1 defines how these uncertainties are used to generate the fault signal.

2.1. Bearing Vibration Model

Faults in outer race (OR), inner race (IR) and rolling element (RE), result in periodic impacts with ball pass frequency outer race (BPFO), inner race (BPFI) and ball spin frequency (BSF), respectively. The fundamental train frequency (FTF) is mainly presented as a modulating frequency for rolling element (RE) faults. The nominal period T is the inverse of the characteristic fault frequency of the fault type and is denoted T_{OR}, T_{IR}, T_{FTF} and T_{RE}. Formulas for calculating the periods are provided in Eqs. (1) through (4). The number of rolling elements is denoted by n_r, d and D are the roller and pitch diameter, respectively, and ϕ is the contact angle.

$$T_{OR} = BPFO^{-1} = \left(\frac{f_r n_r}{2} \left(1 - \frac{d}{D} \cos\phi \right) \right)^{-1} \quad (1)$$

$$T_{IR} = BPFI^{-1} = \left(\frac{f_r n_r}{2} \left(1 + \frac{d}{D} \cos\phi \right) \right)^{-1} \quad (2)$$

$$T_{FTF} = FTF^{-1} = \left(\frac{f_r}{2} \left(1 - \frac{d}{D} \cos\phi \right) \right)^{-1} \quad (3)$$

$$T_{RE} = BSF^{-1} = \left(\frac{f_r D}{2d} \left(1 - \left(\frac{d}{D} \cos\phi \right)^2 \right) \right)^{-1} \quad (4)$$

In reality, the impact period is slightly random due to slip, typically around have 1-2 % (Randall & Antoni, 2011). By considering slip as a random process without memory, the bearing signal becomes pseudo-cyclostationary. The slip is accounted for by modeling time between impacts as a random variable $\varepsilon \sim \mathcal{N}(0, \sigma_T^2)$, where $3\sigma_T = 0.02$ (Antoni, 2007a). This keeps the impact periods within approximately $\pm 2\%$ of the nominal period. Thus, the k-th impact period is simply calculated as in Eq. (5).

$$T_k = T(1 + \varepsilon) \quad (5)$$

In a discrete time series, the sample number of the K-th impact, n_K, is calculated as the cumulative sum of previous impact periods multiplied by the sample frequency, and rounded to the nearest integer with the $nint(\cdot)$ operator. The location of the first impact n_0 is determined by sampling a uniform distribution for the first impact time T_0, where $T_0 \sim U(0, T)$. Then, the nearest corresponding sample is calculated as $n_0 = nint\{F_s T_0\}$, where F_s is the sample frequency.

$$n_k = n_0 + nint\{F_s f_r \sum_{k=1}^{K-1} T_k\} \quad (6)$$

The pulse amplitude is also considered a random variable, where the randomness is modeled by $\xi \sim \mathcal{N}(0, \sigma_p^2)$. The discrete pulse train $p[n]$ is created by setting $p[n_k] = 1 + \xi$ for $k \in [0, K)$ pulses and 0 elsewhere. IR and RE faults have additional amplitude modulation as the fault passes through the load zone. This phenomenon is modeled by multiplying a periodic function with the pulse train. As the modulation index m_1 is unknown, it is sampled from an interval for generation of each time series. Rolling element faults typically appear at $2 \cdot BSF$, as the fault strike both the inner and outer per roller revolution. This also creates additional amplitude modulation at BSF with modulation factor m_2. The impulse response is modeled as bandpass-filtered white Gaussian noise $w_{bp}[n]$, with center frequency f_c and bandwidth bw. A Butterworth filter of order 5 is used to make the band-pass filter. The impulse response modulation function for a time series is obtained by convolving the pulse train with a window function $h[l]$ where $l \in [0, L)$, where the pulse is given in Eq. (7). The number of samples L are determined by sampling $L = nint(F_s t_h)$, where $t_h \sim U(0.5\,\text{ms}, 5\,\text{ms})$ is a uniformly distributed variable to model different pulse durations.

$$h[l] = e^{-5l/L} \quad (7)$$

The resulting generated fault signal x_F is given as in Eq. (8).

$$x_F = ((m_1 + m_2) p * h) w_{bp} \quad (8)$$

2.2. Data preprocessing

The following section describes how the generated fault signal x_F and healthy signal x_{HE} are processed before used for training and testing. Each dataset consists of N records with duration t_r. In this paper, the records are either used directly, or segmented using a rectangular rolling window with duration t_w and a stride of t_s. Depending on the available data, either complete samples or shorter segments can be extracted. Because the simulated data is combined with real, healthy data, it is necessary to know which records are healthy. In seeded fault datasets this information is available, but in run-to-failure tests, true condition is not known. This is solved by using the first N_{HE} records as a reference for healthy condition. As this paper utilizes data that has also been analyzed by other researchers, it is possible to choose N_{HE} small enough to be confident that the bearing is actually healthy. All records

selected as healthy are normalized to unit RMS. From the normalized healthy data, the desired number of samples are drawn with replacement for used in training. As the fault data is simulated, it is possible to generate an arbitrary amount of training data. Table 1 shows the number of generated samples for each dataset. Healthy data use for fault simulation was drawn randomly with replacement.

		CWRU	IMS	UiA
HE	Real	967	256	920
HE	Simulated	2048	2048	2048
IR	Simulated	2048	2048	2048
RE	Simulated	2048	2048	2048
OR	Simulated	2048	2048	2048

Table 1. Number of training records

In the seeded fault datasets, all training and test data was normalized to unit RMS to let the network learn patterns from data in a predictable range. First, the simulated faulty data x_F is normalized with a factor $\lambda \operatorname{RMS}\{x_F\}$ to model differences in damage severity. Each segment, x_{HE}, is augmented with additive white noise $w \sim \mathcal{N}(0, 0.05^2)$ and normalized to unit RMS. The sum of these signals is denoted x'_S, as shown in Eq. (9). This signal is again normalized to unit RMS, as shown in Eq. (10).

$$x'_S = \frac{x_{HE} + w}{\operatorname{RMS}\{x_{HE} + w\}} + \frac{x_F}{\lambda \operatorname{RMS}\{x_F\}} \qquad (9)$$

$$x_S = \frac{x'_S}{\operatorname{RMS}\{x'_S\}} \qquad (10)$$

In run-to-failure datasets, all records are normalized with the mean RMS of the N_{HE} reference records. Again, each healthy segment, x_{HE}, is augmented with additive white noise $w \sim \mathcal{N}(0, 0.05^2)$ and normalized to the RMS of the original segment. The simulated faulty data x_F is also normalized with the same factor $\lambda \operatorname{RMS}\{x_F\}$. This time, the two parts are summed directly to obtained the simulated signal x_S, as shown in Eq. (11). This approach also lets the classifier associate higher RMS with a fault, and was included to help the classifier discriminate noise originating from increasingly severe faults and background noise in healthy records. Figure 1 shows the components of a generated rolling element fault and an actual fault signal.

$$x_S = \frac{(x_{HE} + w) \operatorname{RMS}\{x_{HE}\}}{\operatorname{RMS}\{x_{HE} + w\}} + \frac{x_F}{\lambda \operatorname{RMS}\{x_F\}} \qquad (11)$$

2.3. Classification Network

The classifier is aimed at raw vibration time series input from one or more channels. In this section, we consider single-channel measurements for simplicity. The network consists of three main parts: A low-level feature extraction part, an intermediate feature extraction part with dilated convolutions for capturing long term feature dependencies. Finally, a classification part weights the features and predicts an output. This hierarchical configuration allows the network to first learn long-term relationships with less computational demand. The network configuration and parameters used in this paper are given in Table 2. Development was done in Keras with Tensorflow backend (Chollet & Others, 2015), therefore Keras layer names are used in the table.

The low-level feature extraction part consists of two blocks with same structure, as shown in Figure 2, each containing a 1D convolutional layer, a max pooling layer and a batch normalization layer. This layer extracts features that are close in time by convolving the signal with a kernel that is small compared to the input. Rectified linear unit (ReLU) activation functions (Nair & Hinton, 2010) are used throughout the network, except for final class output. Max pooling is used to reduce the dimension while maintaining the most prominent features. Batch normalization speeds up training by reducing internal covariate shift (Ioffe & Szegedy, 2015). In this paper, separate classifiers were trained for all channels. Principally, the network structure can also handle multiple data channels simultaneously, but this was not investigated in the experiments.

Dilated convolutions is an efficient way to increase the receptive field of the network. By dilating the convolution kernel, smaller filter sizes can capture long-term dependencies in the data with relatively few layers. The approach has been successful in modeling of high resolution time series (Van Den Oord et al., 2016). The concept is shown in Figure 3. Input data is shown in green, active intermediate features in blue, and the output feature in red. By doubling the dilation rate for each new layer, the receptive field grows exponentially.

One-dimensional global average pooling is applied to the output of each filter after all convolutions. This is in contrast to the fully connected layers, which were typically found in the final layers of a CNN. Global average pooling reduces overfitting and makes classification less sensitive to the temporal translation that is introduced by windowing the time series (Lin, Chen, & Yan, 2014). The output is then fed directly to the final, fully connected classification layer with softmax activation. The objection function to be optimized is categorical crossentropy loss between predicted and true label. The networks were trained for 10 epochs with batch size of 32, using an Adam optimizer with learning rate $1e^{-4}$, except dataset 2 which used a batch size of 16 due to GPU memory constraints.
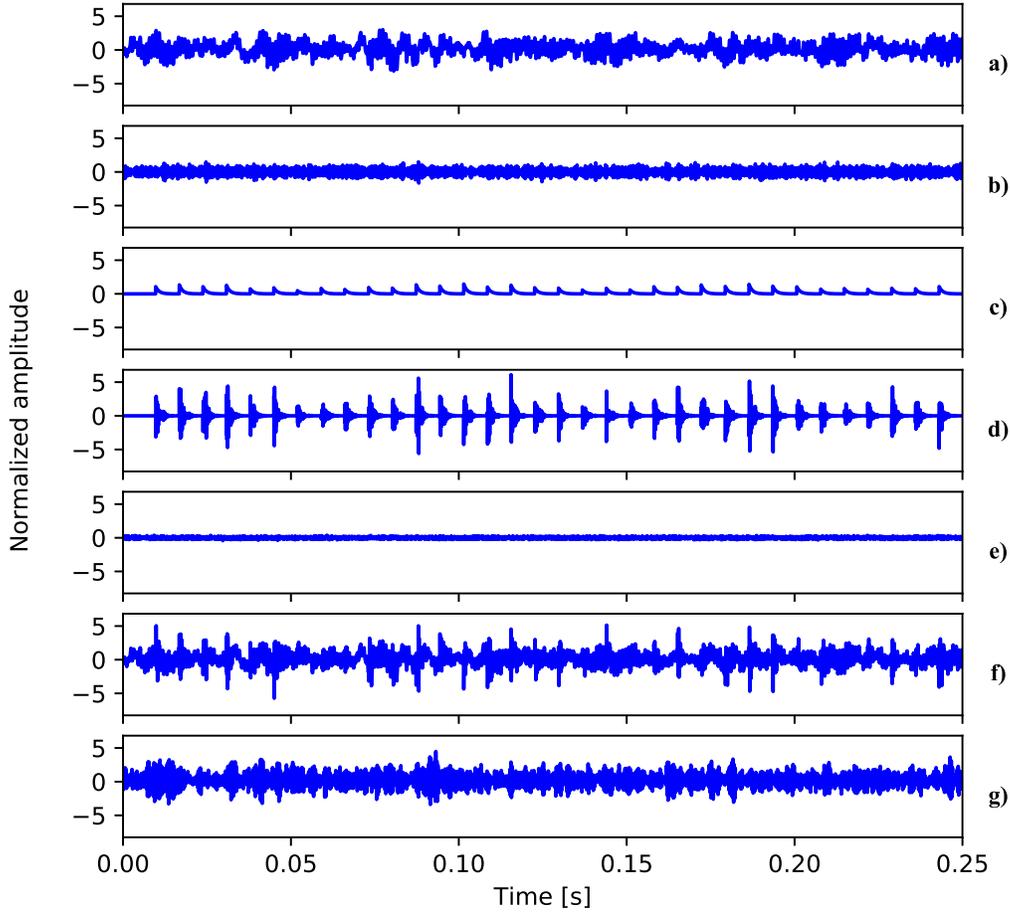
Figure 1. Components of generated rolling element fault compared to actual fault signal. **a)** Healthy signal x_{HE}. **b)** Band-pass filtered noise w_{bp}. **c)** Modulating envelope. **d)** Generated fault impacts x_p. **e)** Additive noise w. **f)** Generated fault signal x_s. **g)** Actual fault signal.

There is randomness involved in multiple stages of the training process. Network weights are initialized randomly, so two identical models with different random seeds may yield different results. Random shuffling of training data may also affect results. To improve generalization and reduce the effect of randomness, five folds are created from the available data. Each fold contains all the data, but the distribution of data in training, validation and testing splits are different. Each fold is used to train a separate model, resulting in an ensemble of five classifiers. In cases where a record is split in multiple segments, each classifier outputs a decision per record based on a plurality voting scheme. If the record consists of a single segment, there is no voting in this step. The final ensemble classification is also determined through a plurality vote over the individual decisions.

3. EXPERIMENTAL DATA

Vibration data from three datasets have been used in this study. Dataset 1 contains healthy (HE) and seeded fault data from IR, RE and OR, used to verify that the methodology is capable of detecting and diagnosing the different fault types. The other datasets contain run-to failure data. The following sections give a brief description of the test rigs and data that are utilized in this paper. Results from reference publications are used as a baseline for evaluating classifier performance.

3.1. Dataset 1: Case Western Reserve University (CWRU)

Dataset 1 is provided by the Case Western Reserve University (CWRU) and includes vibration records at the drive end,

| No.	Type	Filters	Size	Stride	Padding	Dilation	Activation
0	Input						
1	Conv1D	32	11	1	valid	1	ReLU
2	MaxPooling1D		3	2			
3	BatchNormalization						
4	Conv1D	32	5	1	valid	1	ReLU
5	MaxPooling1D		3	2			
6	BatchNormalization						
7	Conv1D	8	5	1	same	1	ReLU
8	Conv1D	8	5	1	same	2	ReLU
9	Conv1D	8	5	1	same	4	ReLU
10	Conv1D	8	5	1	same	8	ReLU
11	Conv1D	8	5	1	same	16	ReLU
12	Conv1D	8	5	1	same	32	ReLU
13	Conv1D	8	5	1	same	64	ReLU
14	Conv1D	8	5	1	valid	128	ReLU
15	GlobalAveragePooling1D						
16	Dense	4					Softmax

Table 2. Network parameters
s
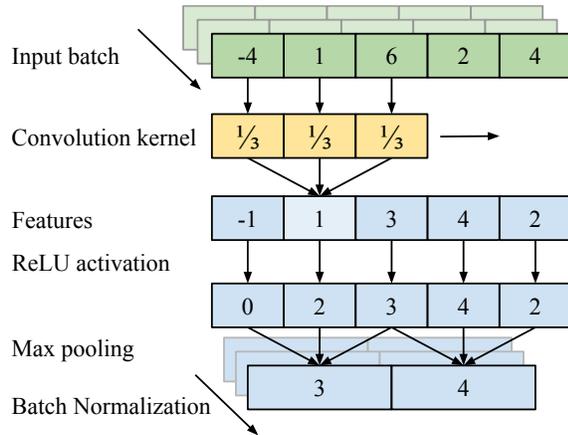
Figure 2. Low-level feature extraction: Input data points (green) are convolved with 1D kernel (yellow), followed by a ReLU activation function and max pooling layer. The batch normalization is applied to the output of the max pooling.
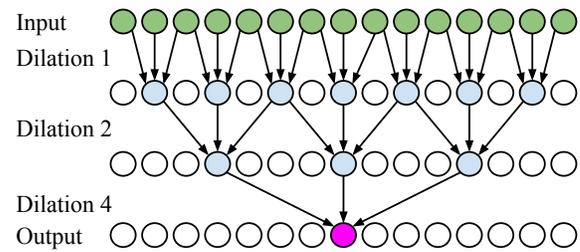

Figure 3. Intermediate level feature extraction: The receptive field (green) and intermediate features (blue) for computing feature output (pink). Dilation allows a wide receptive field with few layers.

fan end and foundation of a motor. Both the drive end and fan end bearing were seeded separately with OR, IR and RE faults. Data from four loads and four damage severities are available, with a sample frequency of 12 kHz. Additionally, data sampled at 48 kHz are available for drive end faults only. The CWRU test setup is shown in Figure 4. The test motor (left) is connected to a dynamometer (right) through an encoder and torque sensor (center) (*Case Western Reserve University Bearing Data Website*, n.d.).

In a reference paper for analysis (Smith & Randall, 2015), authors applied three fault diagnosis metohds to the data: Envelope analysis of the raw signal, cepstrum prewhitening forllowed by envelope analysis, and envelope analysis of a bandpass filtered signal, where spectral kurtosis was used to calcu-

late the optimal passband. Faults in OR and IR were successfully diagnosed using one or more of the methods. However, the RE faults were not diagnosable. Numerous papers using CWRU data report better performance than the reference paper. A review of deep learning algorithms trained on the CWRU datset show that the majority of research papers report 95-100 % accuracy (Zhang, Zhang, Wang, & Habetler, n.d.). However, supervised classification algorithms displaying such high accuracy may be a sign of overfitting and poor generalization performance (Smith & Randall, 2015).

This paper uses data from the drive end bearing, running with unloaded motor, sampled at 48 kHz. This was chosen to allow the classifier to capture higher frequency amplitude modulation. The smallest fault size, measuring 0.18 mm in diameter and 0.28 mm in depth, was used for all fault types. Healthy data in only available sampled at 12 kHz for 20 seconds. This record is upsampled by a factor of 4 to obtain 48 kHz healthy data. Faulty bearing records are otherwise 5 seconds long. To obtain more training data, a window of 1 second with a stride
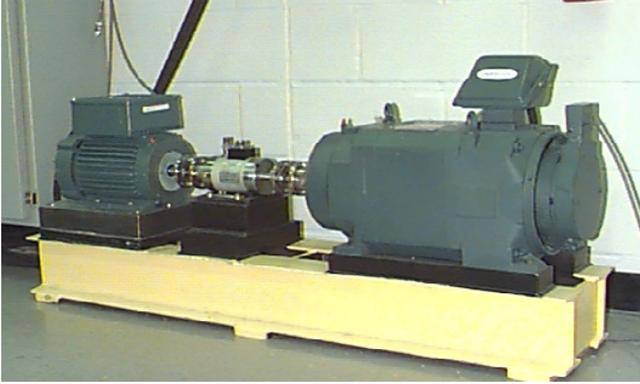
Figure 4. Case Western Reserve University test rig, used for dataset 1.

Figure 5. Schematic drawing of IMS test rig used to collect dataset 2.

of 0.02 seconds is applied to the original record. The files used are listed in table 3.

	HE	IR	RE	OR
File number	97	109	122	135

Table 3. Files from CWRU dataset used in this paper.

3.2. Dataset 2: Center for Intelligent Maintenance Systems (IMS)

Dataset 2 consists of run-to-failure test data, provided by NSF I/UCR Center for Intelligent Maintenance Systems (IMS). A diagram of the test rig is shown in Figure 5. The test setup is further described in the reference paper (Qiu, Lee, Lin, & Yu, 2006). In this paper, data from test 1 and 2 is used. Test 1 had 8 available channels, however only one channel per bearing was used (channel 1, 3, 5 and 7). Test two was recorded using only one channel per bearing. Data was sampled at 20480 Hz for one second. In both cases, the first 256 records were assumed healthy and used in training. As the available records were only on second, no windowing was performed on this dataset. In test 1, an IR fault was found in bearing 3, and an RE fault in bearing 4. An OR fault occurred in bearing 1 in test 2. In both datasets , an increase in RMS and kurtosis is observed towards the end, as shown in Figure 6. This increase is interpreted as an indication of damage occurrence and progression. In test 1, bearing 4 shows slightly raised kurtosis around record 1435, followed by a large increase from record 1610. Bearing 3 kurtosis increases from record 1800. In test 2, an increase is seen from record 530.

3.3. Dataset 3: The in-house test at University of Agder (UiA)

Dataset 3 was collected using an in-house test rig from the University of Agder as shown in Figure 7. The test rig was made for run-to-failure tests at low and variable speed con-
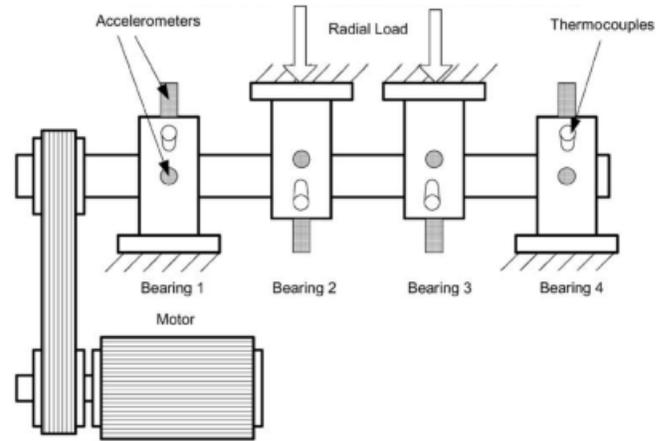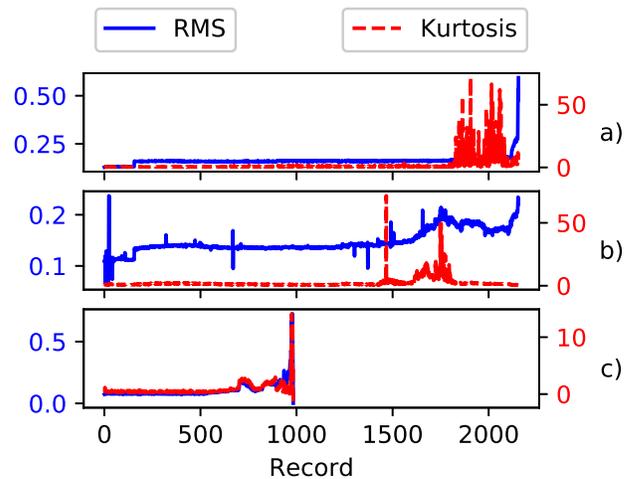

Figure 6. Plot of RMS and kurtosis for the faulty bearings in dataset 2. **a)** Test 1, bearing 3. **b)** Test 1, bearing 4. **c)** Test 2, bearing 1.

ditions under combined radial and axial load. The design and functionality thoroughly described in (Klausen, Folgerø, Robbersmyr, & Karimi, 2017). Vibration data was collected from an accelerometer perpendicular to the load zone, sampled at 51200 Hz. Each record contains 100 revolutions. A 2-second window with 1 second stride was applied, resulting in 23 windows per record. 192 records from the last 5 days of testing was used in this paper. Signs of faults were observed from record 163 and onward. Recordings from the first day, 40 records in total, were used for training.

The reference paper (Klausen, Robbersmyr, & Karimi, 2017) reports a pit in the OR, two damaged rollers and a heavily spalled IR at the end of the test. The authors report to observe signs of RE damage first, starting at 29 recordings from the end of life. Later, indications of OR and IR damage were also

Figure 7. Test rig used in dataset 3.

observed, in that order.

Bearing type and fault frequencies for the datasets are listed in Table 4.

	CWRU	IMS	UiA	Unit
Make	SKF	Rexnord	SKF	
Model	6205-2RSJEM	ZA-2115	6008-2RS1	
BPFO	107.37	236.38	21.32	Hz
BPFI	162.18	296.90	28.57	Hz
BSF	141.18	139.92	13.86	Hz
FTF	11.92	14.77	1.77	Hz

Table 4. Bearing specifications for the test datasets

3.4. Vibration Model Parameters

Fault simulation requires selection of model parameters described in section 2.1. Sample frequency F_s, record duration t_r and shaft rate f_s are known system parameters, while window duration t_w and stride t_s are user-defined, limited by sample frequency and the length of available data. For the remaining parameters, optimal values are not known. Therefore, instead of attempting to determine the optimal values, the parameters are defined as either uniform or Gaussian distributions. This papers aims to demonstrate that by defining the distributions wide enough to encompass the assumed true values, a subset of the simulated datapoints will approximate the true faulty condition. Thus, the network is able to classify real, unseen data.

Window duration should be set long enough to capture periodicity from the impacts. The lowest frequency component of interest is typically the FTF, so t_w should at least exceed $1/FTF$. Window length t_w should be set sufficiently long to capture at least a few shaft revolutions, as one would if signal processing were done in the frequency domain. However, it is assumed that longer windows will improve classifier performance. There is no upper limit of t_w, but computational load increase with segment length.

As the test rig in dataset 2 operates at low speed, the window is longer than for the high-speed datasets. Window stride t_s is of less importance, but affects the total number of unique windows. As a rule of thumb, t_s should be selected smaller than t_w to have some overlap of the windows and thus capture more variations of the signal.

The combined center frequency f_c and bandwidth bw must not violate the Nyquist criterion. A conservative limit of $F_s/2.56$ is used. It was found empirically that the impulse response duration t_h should be short enough to not have overlap between impulses in the lower end of the range. A summary of the bearing specifications is shown in Table 4. Impact arrival time jitter ξ, was set to approximately $\pm 2\%$ to account for slip. Otherwise, the classifier struggles to separate fault types. The upper limit on RMS ratio λ was set to 4 for dataset 1, as it was known that rolling element faults were hard to detect. This did however, not result in successful diagnosis. A low RMS parameter should encourage the network to learn weak signatures, but at the risk of making damage and healthy too similar. Other parameters were set wide across the datasets. A systematic parameter grid search was not performed due to the associated computational load. Table 5 shows the dataset-specific parameters, while the parameters common for all datasets are listed in Table 6. Distributions for center frequency and bandwidth are deliberately chosen wide to reduce the chance of achieving good results purely by coincidence. Other parameters are simply set by making a qualified guess, and has not been tuned specifically to improve performance except what is already noted for parameter λ and t_h.

	CWRU	IMS	UiA	Unit
F_s	48	20.48	51.2	kHz
f_s	29.95	33.33	4.17	Hz
t_w	1	1	2	s
t_s	0.02	0.1	1	s
f_c	$U(1.5, 15)$	$U(1.5, 8)$	$U(1.5, 15)$	kHz
bw	$U(0.5, 5)$	$U(0.5, 5)$	$U(0.5, 5)$	kHz
λ	$U(1, 4)$	$U(1, 2)$	$U(1, 2)$	-
N_{HE}	-	256	40	-

Table 5. Model-specific bearing vibration model parameters.

Parameter	Value	Unit
t_h	$U(0.5, 5)$	ms
ϵ	$\mathcal{N}(0, \sigma_T^2)$	s
$3\sigma_T$	0.02	s
ξ	$\mathcal{N}(0, \sigma_p^2)$	-
$3\sigma_p$	0.1	-
m_1	$U(0.1, 0.5)$	-
m_2	0.25	-

Table 6. Common vibration model parameters.

4. RESULTS

This section presents the results of training the classifier on healthy and simulated fault data and testing on actual fault data. For dataset 1, which has seeded faults, performance can be evaluated quantitatively through classification accuracy. In datasets 2 and 3, there is no ground truth available, which makes a quantitative performance evaluation impossible. Classifier outputs are compared to results in reference papers, and evaluated qualitatively. All predicted classes shown are the result of a plurality vote in an ensemble of 5 classifiers.

4.1. Dataset 1 (CWRU)

Dataset 1 contained samples of single-point faults in OR, IR and OR. The classifier successfully diagnosed the IR and OR fault, but did not manage to diagnose RE faults. Instead, this fault type was consistently misclassified as IR damage by the ensemble. The confusion matrix for is shown in Figure 8. As seen in the confusion matrix, there were no false alarms, and no damaged bearings were classified as healthy. It's worth noting that in the reference paper, RE faults were not identifiable using any of the applied analysis methods, and did not show the same classical behavior as IR and OR faults (Smith & Randall, 2015). The achieved result of the proposed method is therefore on par with the reference paper. While other deep learning algorithms have been able to diagnose the RE fault as well, the authors of this paper are not aware of any algorithms achieving this with simulated training data.

Figure 8. Confusion matrix for CWRU test data. Predicted label is a result of a plurality voting in the classifier ensemble.

4.2. Dataset 2 - Test 1

In this run-to-failure test, bearing three (B3) and four (B4) were damaged at the end of the experiment with IR and RE damage respectively. In the run-to-failure experiments, each record is classified separately by the ensemble. The kurtosis trend, shown in Figure 6, has a slight increase around record 1435, so this is the time when damage is expected to start in bearing 4. The outcome of the plurality voting for each record is shown in Figure 9. Record 1525 in bearing 4 is the first to be classified as faulty. It is classified with a RE fault, which is the damage that is found at the end of life. However, from record 1554, IR damage is indicated, and from that point and onward the classifier outputs both IR and RE damage.

In bearing 3, the classifier also identifies RE and IR damage, with a majority of IR damage from record 2000. Towards the end of life, it is classified as OR fault. The IR fault indication is in accordance with the findings in the reference paper (Qiu et al., 2006). The authors also note that the IR appeared to be severely spalled, so any impact impulses may not be as prominent as in the training data.

OR damage is indicated in bearing 1 and 2 towards the end of life, but this is assumed to be caused by faults in the other bearings.
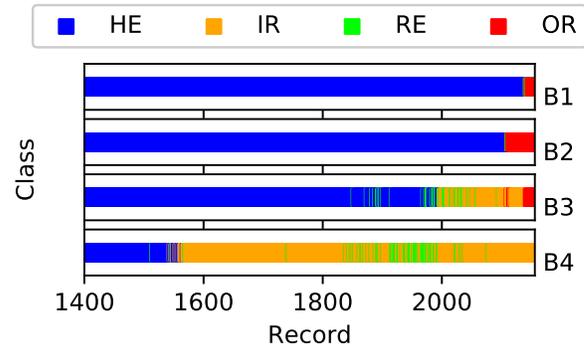

Figure 9. Dataset 2, test 1, recording 1400 until end of life. Classification by ensemble plurality voting.

4.3. Dataset 2 - Test 2

In this test, an OR fault was found in bearing 1 at the end of life, and first signs are expected to appear from approximately record 530. The classifier ensemble correctly outputs OR damage consistently from record 545, as shown in Fig 10, until the final stages of bearing life, where the classifier changes from OR damage to HE. A possible explanation is that at the end of life, fault size increase, and noise masks any periodic impacts. A healthy bearing is also mainly noisy in some frequency bands. Damage is indicated in the other bearings as well, but as in test 1, this is assumed to be cause by vibration induced by the fault in bearing 1.
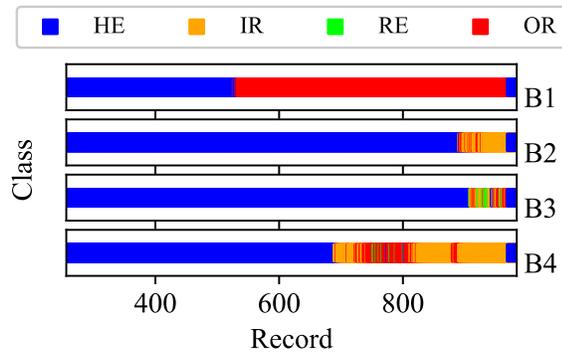
Figure 10. Dataset 2, test 2. Classification by ensemble plurality voting.

4.4. Dataset 3

Dataset 3 is a more challenging case, as the shaft operates at a low speed of 250 rpm. As described in the reference dataset (Klausen, Robbersmyr, & Karimi, 2017), there was extensive bearing damage at the end of the test. The bearing had a small OR pit, two damaged balls, and a larger spalled area in the IR. The authors observed signs of damage in the 20 last measurements, with damage on balls appearing first. Figure 11 shows RE damage from record 171, then HE from record 179, and RE again from record 188 until record 192.
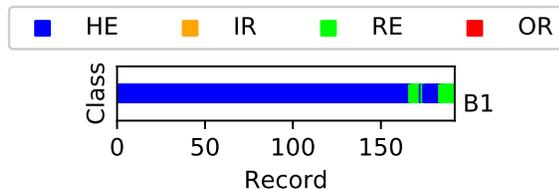

Figure 11. Dataset 3 test classification by ensemble plurality voting.

These results are partly in accordance with the reference paper. Signs of damage occur at the same time, and the type of damage is coinciding. The classifier does not indicate other damage than RE, but all fault types were present at the end of testing. This shows that the initial classification was in accordance with the reference paper, but the classifier struggles when multiple faults are present later in the test. This behavior is expected, as training data was only simulated with a single fault at a time.

5. CONCLUSION

In this work, a deep learning classifier for raw vibration signals was trained on simulated data and evaluated on actual fault data. By stacking several layers of 1D convolutions in a hierarchical, dilated structure, the classifier is able to cover a wide receptive field with few layers. This type of architecture

has been used successfully in speech and audio modeling, but has yet seen limited application in machine condition monitoring. The presented results show that simulation-driven training of deep neural networks for bearing fault detection and classification has clear potential for industrial applications where automated analysis of time series is beneficial and the access to actual fault data is scarce. Overall, the classifier was able to detect the presence of faults, but tended to misclassify fault types. This may be acceptable in some industrial applications, where detection is more important than diagnosis. In dataset 1, RE damage was classified as IR. In the reference paper, RE damage was also not identifiable, and was said to not exhibit classical fault behavior. This may be the cause of misclassification, as the simulated bearing vibration was simply not representative of this failure mode. Other deep learning classifiers have achieved near perfect accuracy on the same dataset, but not without using real fault data for training. Detection performance is also comparable to results in reference articles in run-to failure dataset 2 and 3, but diagnosis is inconsistent. Training data quality and consequently classifier performance will improve with more information about the system. For example, an estimate of the resonance frequency can be obtained through a bump test. More advanced and accurate models for bearing vibration could also improve performance, and should be investigated further.

REFERENCES

Antoni, J. (2006). The Spectral Kurtosis: A Useful Tool for Characterising Non-Stationary Signals. *Mech. Syst. Signal Process.*, *20*(2), 282–307. doi: 10.1016/j.ymssp .2004.09.001

Antoni, J. (2007a). Cyclic Spectral Analysis of Rolling-Element Bearing Signals: Facts and Fictions. *J. Sound Vib.*, *304*(3-5), 497–529. doi: 10.1016/j.jsv.2007.02 .029

Antoni, J. (2007b). Fast Computation of the Kurtogram for the Detection of Transient Faults. *Mech. Syst. Signal Process.*, *21*(1), 108–124. doi: 10.1016/j.ymssp.2005 .12.002

Antoni, J. (2009). *Cyclostationarity by Examples* (Vol. 23) (No. 4). doi: 10.1016/j.ymssp.2008.10.010

Antoni, J., & Randall, R. B. (2002). Differential Diagnosis of Gear and Bearing Faults. *J. Vib. Acoust.*, *124*(2), 165. doi: 10.1115/1.1456906

Bechhoefer, E., & Kingsley, M. (2009). A Review of Time Synchronous Average Algorithms. In *Annu. conf. progn. heal. manag. soc.* (pp. 24–33).

Bechhoefer, E., Schlanbusch, R., & Waag, T. I. (2016). Fault Detection on Large Slow Bearings. In *Phme 2016* (Vol. 7, pp. 1–8).

Borghesani, P., Pennacchi, P., Randall, R. B., Sawalhi, N., & Ricci, R. (2013). Application of Cepstrum Pre-Whitening for the Diagnosis of Bearing Faults Under Variable Speed Conditions. *Mech. Syst. Signal Process.*, *36*(2), 370–384. doi: 10.1016/j.ymssp.2012.11 .001

Case Western Reserve University Bearing Data Website. (n.d.).

Chollet, F., & Others. (2015). *Keras.* \url{https://keras.io}.

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Inf. Fusion*, *50*, 92–111. doi: 10.1016/J.INFFUS.2018 .10.005

Fyfe, K., & Munck, E. (1997). Analysis of Computed Order Tracking. *Mech. Syst. Signal Process.*, *11*(2), 187–205. doi: 10.1006/MSSP.1996.0056

Gan, M., Wang, C., & an Zhu. (2015). Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings. *Mech. Syst. Signal Process.*, *72-73*, 92–104. doi: 10.1016/j.ymssp.2015.11.014

Guo, X., Chen, L., & Shen, C. (2016). Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis. *Measurement*, *93*, 490–502. doi: 10.1016/J.MEASUREMENT.2016.07.054

He, M., & He, D. (2017). Deep Learning Based Approach for Bearing Fault Diagnosis. *IEEE Trans. Ind. Appl.*, *53*(3), 3057–3065. doi: 10.1109/TIA.2017.2661250

Hecke, B. V., Yoon, J., & He, D. (2016). Low speed bearing fault diagnosis using acoustic emission sensors. *Appl. Acoust.*, *105*, 35–44. doi: 10.1016/j.apacoust.2015.10 .028

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Process. Mag.*.

Ho, D., & Randall, R. B. (2000). Optimisation of Bearing Diagnostic Techniques Using Simulated and Actual Bearing Fault Signals. *Mech. Syst. Signal Process.*, *14*(5), 763–788. doi: 10.1006/mssp.2000.1304

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* (Tech. Rep.).

Jia, F., Lei, Y., Guo, L., Lin, J., & Xing, S. (2018). A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing*, *272*, 619–628. doi: 10.1016/ J.NEUCOM.2017.07.032

Jiang, G., He, H., Yan, J., & Xie, P. (2019). Multiscale Convolutional Neural Networks for Fault Diagnosis of

Wind Turbine Gearbox. *IEEE Trans. Ind. Electron.*, *66*(4), 3196–3207. doi: 10.1109/TIE.2018.2844805

Khan, M. A., Kim, Y.-H., & Choo, J. (2018). Intelligent Fault Detection via Dilated Convolutional Neural Networks. In *2018 ieee int. conf. big data smart comput.* (pp. 729–731). IEEE. doi: 10.1109/BigComp.2018.00137

Klausen, A., Folgerø, R. W., Robbersmyr, K. G., & Karimi, H. R. (2017). Accelerated Bearing Life-time Test Rig Development for Low Speed Data Acquisition. *Identif. Control*, *38*(3), 143–156. doi: 10.4173/mic.2017.3.4

Klausen, A., Robbersmyr, K. G., & Karimi, H. R. (2017). Autonomous Bearing Fault Diagnosis Method based on Envelope Spectrum. *IFAC-PapersOnLine*, *50*(1), 13378–13383. doi: 10.1016/J.IFACOL.2017.08.2262

Li, G., Deng, C., Wu, J., Xu, X., Shao, X., Wang, Y., et al. (2019). Sensor Data-Driven Bearing Fault Diagnosis Based on Deep Convolutional Neural Networks and S-Transform. *Sensors*, *19*(12), 2750. doi: 10.3390/ s19122750

Li, X., Liu, Z., Qu, Y., & He, D. (2018). Unsupervised Gear Fault Diagnosis Using Raw Vibration Signal Based on Deep Learning. In *2018 progn. syst. heal. manag. conf.* (pp. 1025–1030). IEEE. doi: 10.1109/ PHM-Chongqing.2018.00182

Li, X., Yang, Y., Pan, H., Cheng, J., & Cheng, J. (2019). A novel deep stacking least squares support vector machine for rolling bearing fault diagnosis. *Comput. Ind.*, *110*, 36–47. doi: 10.1016/J.COMPIND.2019.05.005

Lin, M., Chen, Q., & Yan, S. (2014). *Network In Network* (Tech. Rep.).

Marple, L. (1999). Computing the Discrete-time "Analytic" Signal via FFT. *IEEE Trans. Signal Process.*, *47*(9), 2600–2603. doi: 10.1109/78.782222

McFadden, P., & Smith, J. (1984). Model for the Vibration Produced by a Single Point Defect in a Rolling Element Bearing. *J. Sound Vib.*, *96*(1), 69–82. doi: 10.1016/ 0022-460X(84)90595-9

Nair, V., & Hinton, G. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proc. 27th int. conf. mach. learn.*

Qiu, H., Lee, J., Lin, J., & Yu, G. (2006). Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *J. Sound Vib.*, *289*, 1066–1090. doi: 10.1016/j.jsv.2005.03.007

Randall, R. B., & Antoni, J. (2011). Rolling Element Bearing Diagnostics - A Tutorial. *Mech. Syst. Signal Process.*, *25*(2), 485–520. doi: 10.1016/j.ymssp.2010.07.017

Randall, R. B., & Sawalhi, N. (2011). Use of the Cepstrum to Remove Selected Discrete Frequency Components from a Time Signal. In *Proc. int. conf. noise vib. eng.* (pp. 451–461). Springer, New York, NY. doi: 10.1007/ 978-1-4419-9428-8_38

Randall, R. B., Sawalhi, N., & Coats, M. (2011). A Comparison of Methods for Separation of Deterministic and

Random Signals. *Int. J. Cond. Monit.*, *1*(1), 11–19.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proc. annu. conf. int. speech commun. assoc. interspeech.*

Sawalhi, N., & Randall, R. B. (2008). Simulating Gear and Bearing Interactions in the Presence of Faults Part I. the Combined Gear Bearing Dynamic Model and the Simulation of Localised Bearing Faults. *Mech. Syst. Signal Process.*, *22*(8), 1924–1951. doi: 10.1016/j.ymssp.2007.12.001

Shao, H., Jiang, H., Zhang, H., Duan, W., Liang, T., & Wu, S. (2018). Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing. *Mech. Syst. Signal Process.*, *100*, 743–765. doi: 10.1016/J.YMSSP.2017.08.002

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.*, *64-65*, 100–131. doi: 10.1016/j.ymssp.2015

.04.021

Sobie, C., Freitas, C., & Nicolai, M. (2018). Simulation-Driven Machine Learning: Bearing Fault Classification. *Mech. Syst. Signal Process..* doi: 10.1016/j.ymssp.2017.06.025

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). *WaveNet: A Generative Model for Raw Audio* (Tech. Rep.).

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2014). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* (Tech. Rep.).

Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (n.d.). *Machine Learning and Deep Learning Algorithms for Bearing Fault Diagnostics-A Comprehensive Review* (Tech. Rep.).

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep Learning and Its Applications to Machine Health Monitoring. *Mech. Syst. Signal Process.*, *115*, 213–237. doi: 10.1016/J.YMSSP.2018.05.050