



Ball bearing multiple failure diagnosis using feature-selected autoencoder model

Ren-Chi Cheng¹ · Kuo-Shen Chen¹

Received: 16 December 2021 / Accepted: 11 March 2022 / Published online: 21 March 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

Recently, with the advance in information technology, pure data-driven approaches such as machine learnings have been widely applied in status diagnosis. However, the accuracy of those predictions strongly relies on the original data, which largely depends on the selected sensors and signal features. Furthermore, for unsupervised machine learning schemes, although it could avoid the concern of labeling in training, it lacks a quantified evaluation of the prediction results. These concerns significantly limit the effectiveness of modern machine learning and thus should be investigated. Meanwhile, ball bearings are fundamental key machine elements in rotating machinery and their condition monitoring should be critical for both quality control and longevity assessment. In this paper, by utilizing ball bearing failure diagnosis as the main theme, the flow of feature selection and evaluation, as well as the evaluation flow for multiple failure diagnosis, is developed for accessing the status of bearings in their imbalance, lubrication, and grease contamination levels based on unsupervised machine learning. The experimental results indicated that with proper feature selection, the failure identification could be more definite. Finally, a novel model based on the second norm to quantify the classification level of each cluster in hyperspace is proposed as the measure for unsupervised machine learning as the basis for performance evaluation and optimization of unsupervised machine learning schemes and should benefit related machine reliability evaluation studies and applications.

Keywords Ball bearing · Unsupervised machine learning · Feature extraction · Autoencoder · Status diagnosis

1 Introduction

Quality control of machine components is vital for the whole system performance and subsequent product assurance. Traditionally, such a plan is usually executed by either constant service loading- or constant service time-based scheduling [1–3] in past decades. Although such an approach has been proven highly effective based on the upkeep results of past decades, it gradually loses its advantage in modern equipment and manufacturing management due to different business model and product characteristics. Thus, new paradigms to replace the existed scheduled maintenance are currently taking place. With the expeditious development of modern sensory and communication technologies, Industry 4.0 paradigm is currently adapted to enhance productivity and reduce manufacturing costs [4]. For production quality

concerns, the spirit of Industry 4.0 also alters the traditional scheduled maintenance approach into predictive maintenance manner by analyzing signals from embedded sensors using big data and machine learning [5].

Among all machine elements, ball bearing, which can be found virtually in any movable mechanisms, is one of the most commonly used fundamental components. Its status directly impacts the associated machine and even manufacture quality assurance. Therefore, its reliability assessment, failure characteristics, and life predictions have been widely investigated during the past decades [6–8]. Due to its importance, generality, and relatively rich domain knowledge, it is therefore an ideal platform for developing associated failure characterization schemes and machine learning-based prediction as the initial effort to realize general diagnosis and status prediction schemes in generalized machine elements.

Bearings may have their particular signatures during service [9] and could reflect their reliability and remaining life based on a rational flow including the following: signal data collection, feature extraction, and finally condition assessments with signal processing routes such as

✉ Kuo-Shen Chen
kschen@mail.ncku.edu.tw

¹ Department of Mechanical Engineering, National Cheng-Kung University, Tainan, Taiwan, Republic of China

short-time-Fourier-transform (STFT), empirical mode decomposition (EMD), Hilbert-Huang transform (HHT) [10], and continuous wavelet transform [11]. In addition, utilizing fundamental physical knowledge in actual diagnosis occasions is also common. For example, Schoen et al. [12] monitored motor bearing failures by examining the stator current. Elasha et al. [13] utilized acoustic emission for diagnosing bearing faults. Furthermore, in real services, bearing failures could contain multiple causes and their responses could be similar. As a result, this implies that hiring multiple transducers for proper fault detection is critical since it is possible to integrate multiple information to provide a more precise diagnosis. However, in practice, such a procedure strongly depends on the coupling level and the experience of the investigator. Meanwhile, virtually, all physical models to describe the causality relationship between causes and consequences are one-to-one mapping. This further increases the challenge to obtain reliable diagnosis based on the above physical domain knowledge-based diagnosis.

On the opposite, with recent rapid advancements in computer science, data-driven approaches such as big data and artificial intelligent (AI) become practical choices for fault classifications in diagnosis. A number of machine learning schemes [14–16] such as neural network [17] and support vector machines (SVM) [18] has been hired for damaged bearing diagnosis [17, 18] but the performances were not sufficient as expected, possibly attributed to improper selection of trained features. Furthermore, AI schemes addressed above are essentially supervised learning models, where the data for training must be properly labeled through considerate experimental plan, which could be unrealistic in practical use. In response of this fact, unsupervised learning methods are now widely used since their nature without the need of labeling. Particularly, autoencoder [19–21], a dimensionality reduction and information reconstruction scheme, has been applied in various aspects ranged from speech and pattern recognition [22, 23], cyber security [24], to fault diagnosis of rolling bearings [19, 20, 25] for data classifications. By such a procedure, data dimensions could be effectively reduced for systematically classification.

However, the accuracy of pure data-driven based diagnosis largely depends on the selected sensors and indexes [26]. Improper selection could result in a waste investment, excessive training data, and poor accuracy and even the possibility of misjudgment. It is recommended that the selection of sensors and indexes should have strong background in related domain knowledge. However, to the best of our knowledge, virtually all previous data-driven based machine status works ignored this issue and concentrated on the development of status prediction based on their selected sensors/indexes.

Previously, the above concerns have been proposed by us [27] and using wear of miller cutter as the test platform by

hiring using typical supervised machine learning scheme artificial neural network (ANN) to examine and demonstrate those matters. The experimental and analysis results supported the arguments. That is, one should select proper sensors and signatures based on domain knowledge or experiments to optimize the performance of chosen machine learning schemes. Here we like to further address the concern and develop a rational flow to cover unsupervised machine learning by using journal bearings as the test platform. In addition, although it does not require on data labeling, unsupervised scheme usually cannot have a quantified evaluation on evaluating clustering. This could result in certain ambiguity. Here, a performance index based on the separation level of clusters in hyperspace would be proposed and evaluated as the first step to quantify the prediction performance of unsupervised machine learning schemes. To the best of our knowledge, there are no similar works reported to address the above two issues.

Motivated by the above arguments, this paper is therefore to evaluate the effect of sensor and feature selections on the prediction performance of unsupervised machine learning schemes and to propose a rational approach to quantify the performance in event clustering by utilizing ball bearing status diagnosis as the test platform. Autoencoder scheme is selected as the primary unsupervised machine learning scheme to be investigated in this work and it is expected that analysis using more advanced schemes would be explored in the future.

Figure 1 shows the overall investigation flow for realizing the abovementioned goals. The first task is to establish a general purpose ball bearing test platform with multiple sensors installed for collecting necessary data. Here imbalance, lubrication level, and contamination of lubricants are

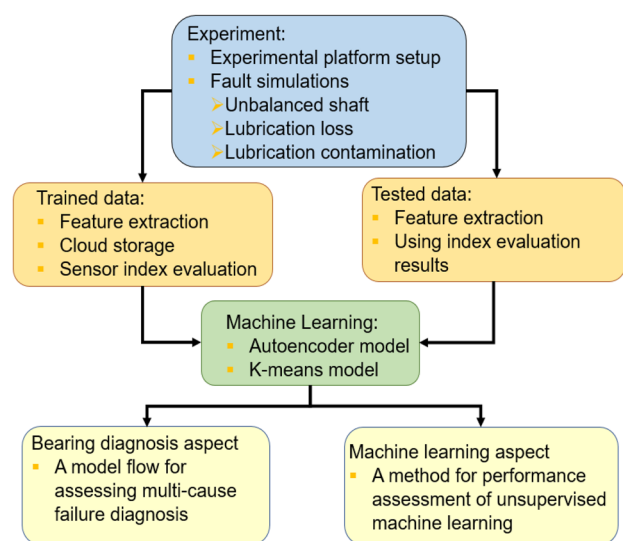


Fig. 1 Investigation flow and the expected outcomes of this work

selected as the major investigation factors. After establishing a testbed installed with multiple transducers for collecting essential information, various time and frequency domain indexes are extracted. Those indexes are then evaluated to select appropriate features to determine their effectiveness. Next, unsupervised machine learning, specifically the autoencoder scheme, is adapted to decrease the feature dimensions and then for visualizing the reduced features on a three-dimensional hyperspace for performance evaluation. In addition, k-means [28], a classification model, would also be adapted to classify the trained results. Based on the setup, one can then evaluate the effectiveness of selecting proper sensor index. Meanwhile, a model proposed by us would be used to quantify the performance of clustering and diagnosis of the used unsupervised machine learning here. Finally, an integrated diagnosis flow is proposed by considering information from cases with a systematic analyzing process.

The abovementioned investigation flow and preliminary results have been presented in conference [29]. In this work, based on our preliminary demonstration [29] and the master thesis of the first author [30], a systematic investigation and discussions, as well as proposing a novel performance indicator, are proposed to provide the most updated results and to elucidate the importance and applications of the proposed quantification evaluation method. The rest of this article thus presents the technical detail. In Sect. 2, the overall system setup and experiment plan is anchored and addressed for guiding the entire study. Essential signal analysis and sensitivity evaluation for picking up effective sensor indexes are presented in Sect. 3. Consequently, in Sect. 4, the signature set and the obtained test data are then used to develop various autoencoder-based unsupervised machine learning schemes for bearing fault diagnosis. A model based on Euclidian distance in hyperspace is proposed in Sect. 5 as the measure for quantifying the clustering performance of unsupervised machine learnings. The key discovery, contribution, and recommended future works are then discussed in Sect. 6. Finally, Sect. 7 concludes this work.

2 Experimental system setup

2.1 Experimental platform setup

The experimental system, as shown in Fig. 2a, is basically a high speed spindle driven shaft supported by an end journal ball bearing. The PMSM spindle provides accurate speed control. Two payload disks (shown in Fig. 2b) are mounted on the aluminum shaft (diameter 20 mm) for providing designated imbalance loading by attaching some eccentric masses on them. Meanwhile, essential faults would be created on the ball bearing intentionally. For example, Fig. 2c

and d show a normal and a lack of lubricated bearings, respectively.

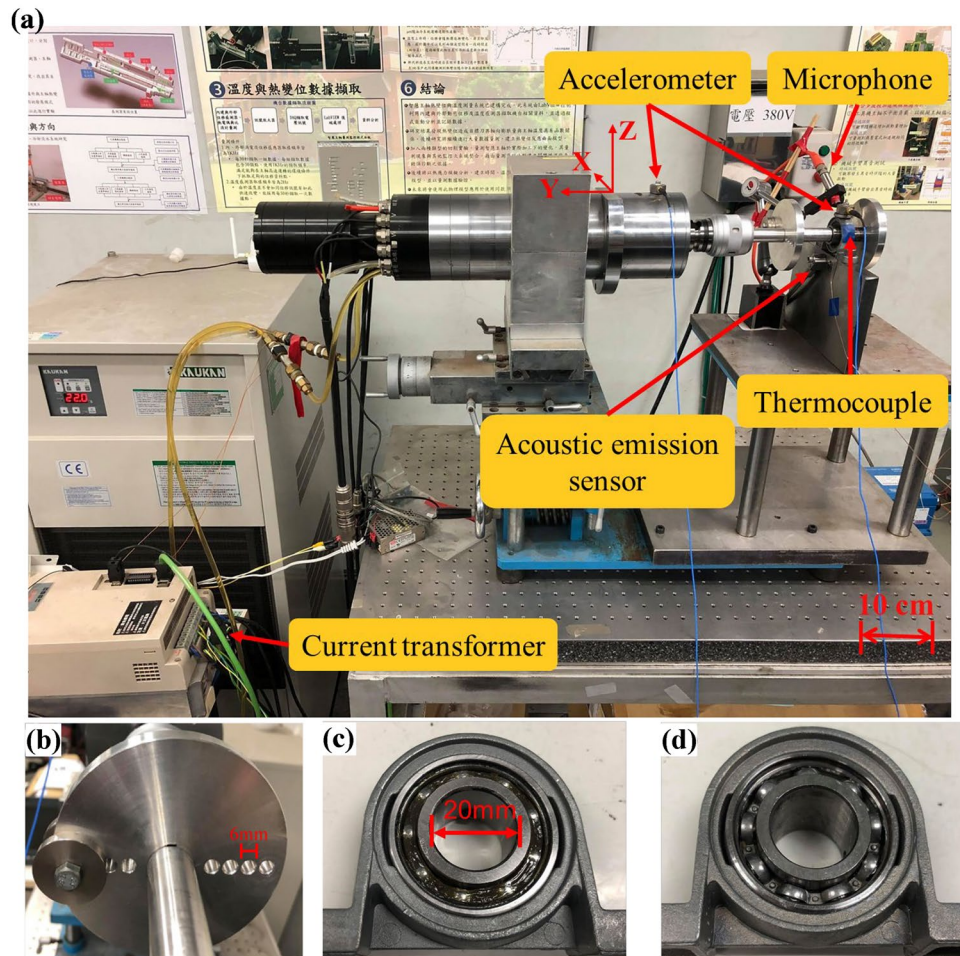
Next, as already shown in Fig. 2a, various transducers, including one PCB piezo microphone, two PCB tri-axial piezo accelerometers, a Kistler acoustic emission (AE) sensor, a K-type thermal couple, plus a current transformer, are set up for accessing acquired physical signals. One of the two tri-axial accelerometers is attached on the spindle for monitoring the spindle vibration during operation and the other is attached on the rim of the ball bearing for detecting abnormal bearing vibration. On the other hand, the microphone is positioned nearby the end ball bearing. The AE sensor is mounted on the bearing foundation for monitoring elastic wave generated and the current sensor hangs on the power cable of the spindle for measuring the current and torque. Finally, the thermal couple is installed on the bearing surface for accessing the friction heating generated during long time operation. Table 1 lists the major specs and the associated data acquiring systems. The sensed data can be acquired synchronically with the sampling rates set as 2 MHz for the AE sensor; 51.2 kHz for the microphone, the accelerometers and the current sensor; and 1 Hz for the thermal couple during the subsequent signal acquisition and processing.

Figure 3 shows a typical result of bearing temperature measurement. Under a rotating speed of 2000 rpm, the temperature of bearing increases from 27.9 to 38.2 °C in 15 min. It is expected that the characteristics of bearing temperature would vary with different operating conditions and different faults. However, the characteristic periods for thermal and vibration analysis signals have significant difference, i.e., 10 (10 min.) vs. 1 (1 ms). This imposes practical difficulties for experiment design and implementation. As a result, we shall focus on addressing the impact of vibration related phenomena on bearing fault status. The temperature related studies would be left as the immediately future work.

2.2 Operating and ball bearing fault condition setup

Rotating speed of the spindle and designated operation faults in the bearings are served as the key controlled input source variables in this work. Four rotating speeds (1000, 1500, 2000, and 2500 rpm) are operated during the test. Fault scenarios created include shaft unbalance, lubrication loss and lubrication contamination of bearings. For ease of classification, three levels (i.e., minor, severe, and healthy) are defined depends on the amount of added flaws. The rotor imbalance is achieved by placing small steel blocks on the disk shown in Fig. 2b. The imbalance level can be adjusted by the number of lumped mass and attached location. In this study, the healthy condition means no extra imbalance is added and the imbalance is solely from the system itself. On the other

Fig. 2 Setup for tests: **a** the platform for conducting experiments, **b** unbalance shaft with an eccentric imbalanced mass and ball bearings with **c** fully lubricated and **d** with no lubrication



hand, the minor imbalance is set by a lump of 19 g placed at a radius of 42 mm (i.e., the imbalance is 80 g-cm), and the severe imbalance is set by a 66 g lump at a radius of 42 mm (imbalance = 277 g-cm).

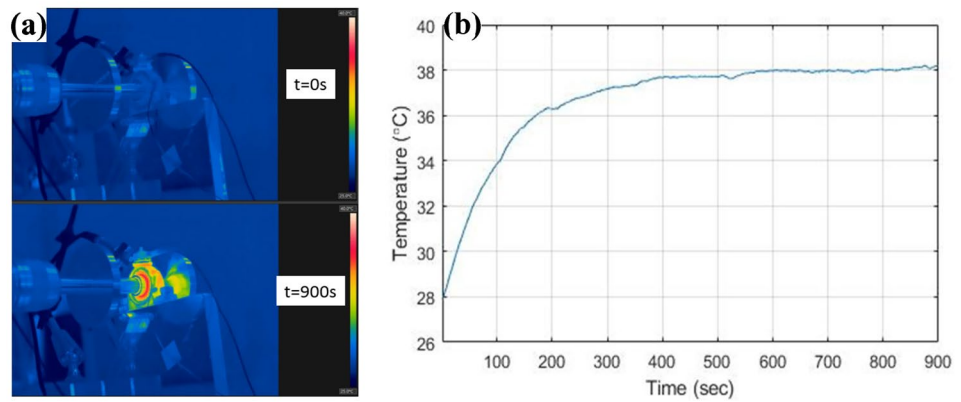
Meanwhile, to realize different levels of lubrications, cleaning naphtha is used to remove the greases in the bearing and re-adding different amount of grease (SKF LGMT 2/0.4) into the bearing. Figure 2c and d show the

fully lubricated bearing and the one with no lubricants. In this study, the severe and minor lacking of lubrication are defined as a bearing with less than 0.01 g and 0.2 g of lubricant, respectively, which corresponds to less than 0.5% and 10% of nominal lubrication level. Finally, for dealing with the issue of lubrication contamination, Inconel 718 metal powders (average diameter approximately 50 μm) are used to mix with the lubricant. The weight ratio between

Table 1 Transducers used in this investigation

Transducers	Spec	Remarks
PCB 354C03 3-axis accelerometer	Bandwidth 6 kHz, sensitivity 100 mV/g	w/ PCB 483C15 charge amplifier
PCB 130F20 Microphone	Bandwidth 16 kHz, gain 41.7 mv/Pa	w/ PCB 483C15 charge amplifier
CTT-10-CLS-CV-5 Current transformer	sensitivity 1 V/A	
Kistler 8152C1030300 AE sensor	Bandwidth 0.1–0.9 MHz, gain 48 dB	w/ Kistler 5125C charge amplifier
TT-K-30-SLE K-type thermocouple	Range –200 ~ 260 °C	
FLIR T530 Infrared camera	Range –20 ~ 650 °C	
DAQ Modules	NI 9234 51.2 kHz/ch NI USB 6361, 2 MHz NI 9213 75 Hz	6361: AE; 9213: for thermocouple

Fig. 3 Temperature of a healthy bearing measured by **a** a thermal image camera and **b** a thermocouple



the lubricant and the Inconel 718 powder is 2% and 25% in minor and severe contamination situations, respectively.

In summary, the experimental design creates one healthy and four faulted cases (i.e., two lubrication levels plus two contamination levels) for addressing bearing damage in this study. For each category, normal and flawed bearings are placed on the test module with four rotating speeds, three imbalanced levels, and five bearing tribology states. Therefore totally 60 scenarios are resulted (i.e., 4 speed × 3 unbalanced conditions × 5 cases) should be investigated. Experiments associated with each scenario are performed repeatedly 100 times and to ensure the repeatability and the sufficient of data. Finally, 100 sets of data are collected for each operation scenario. Each set contains collected data in 10 s. In which 90% of them would be used for training and the rest 10% are for validation. Table 2 lists the detail information regarding the experiment design and data collection plans.

3 Signal processing and analysis

After establishing experimental system and performing tests, the acquired data are then processed for extracting essential features for subsequent effectiveness evaluation and machine learning development. Since the amount of

Table 2 Design of experiment with bearings at different operating conditions

Bearing status (total 5)	Healthy/lubrication loss (minor and severe)/lubrication contamination (minor and severe)			
	1000 rpm	1500 rpm	2000 rpm	2500 rpm
Rotating speed (total 4)				
No unbalance	100	100	100	100
Minor unbalance	100	100	100	100
Severe unbalance	100	100	100	100

Data length: for accelerometer, microphone, current transformer (51.2 kHz for 2 s, totally 102.4k/scenario) and for acoustic emission sensor (2MHz for 2 s, totally 1M/ scenario)

original signals obtained is huge, processing data reduction and extracting effective features is essential. The definitions of key sensor features are introduced first, followed by the evaluation of effective sensor index.

3.1 Feature extraction

From vibration and statistics, it is possible to reduce obtained data into a few meaningful features or indexes to describe the key characteristics of the collected data. Each index could possibly have different sensitivities to specific conditions. Those indexes can be classified into time and frequency domains. The time-domain features utilized in this study are briefly explained below [31]:

- Mean value (μ): μ is the average quantity of a data set.

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \tag{1}$$

- Standard deviation (STD, σ): Standard deviation is a measure of data scattering of a set of data.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu)^2}{n - 1}} \tag{2}$$

- Root mean square (*rms*): *rms* is defined as the square root of the mean square of a set of data.

$$rms = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} \tag{3}$$

- Kurtosis: Kurtosis represents a measure of platykurtic or sharpness of a data set.

$$Kurtosis = \frac{1}{n} \sum_{i=1}^n \frac{E(x_i - \mu)^4}{\sigma^4} \tag{4}$$

- Skewness: Skewness serves as a measure of data symmetry.

Table 3 Characteristic frequency of the octave band

Band No	1	2	3	4	5
Frequency span (Hz)	0–16	16–32	32–64	64–125	125–250
Band No	6	7	8	9	10
Frequency span (Hz)	250–500	500–1000	1000–2000	2000–4000	4000–8000

$$\text{Skewness} = \frac{1}{n} \sum_{i=1}^n \frac{E(x_i - \mu)^3}{\sigma^3} \tag{5}$$

- Entropy: Can be treated as the average level of information uncertainty inherent in the variable’s possible outcomes

$$\text{Entropy} = - \sum_{i=1}^n p(x_i) * \log_2 p(x_i) \tag{6}$$

- Crest factor: which is defined as the ratio between peak amplitude and the *rms* value of a data set for representing the extreme level of a waveform.

$$\text{Crest factor} = \frac{|X_{peak}|}{x_{rms}} \tag{7}$$

- In Eqs. (1)–(7), *n* represents the size in a signal set, *x_i* is the *i*th value in the data set, *E* the expectation value, and *p* the normalized histogram counts. All indexes computations are performed using MATLAB. The indexes mentioned above are applied to data collected by different sensors. Specifically, kurtosis, *rms*, skewness, crest factor, and entropy, are used for the data collected from the accelerometers, the microphone, and the AE sensor. μ and σ are used for the data obtained from current sensor based on subsequent sensitive analyses.

Meanwhile, frequency-domain information such as power spectrum can be obtained by performing STFT of time-domain signals. Here, major frequency-domain indexes are extracted from the signals acquired by the accelerometers and the microphone and are briefly addressed in below.

- Magnitude of the rotating-speed-frequency: This value corresponding to the rotating speed set at the spindle.
- Octave band strength: This indicates the summed energy of a spectrum based on octave band [26]. Here, a full spectrum is divided into ten bands according to the definition listed in Table 3 and the strength is then obtained by summing the spectrum within the band..

After deciding the indexes that being utilized. Proper signal processing can be conducted to evaluate the effectiveness of the selected signatures for reducing the dimension of signals for subsequent storage and machine learning applications.

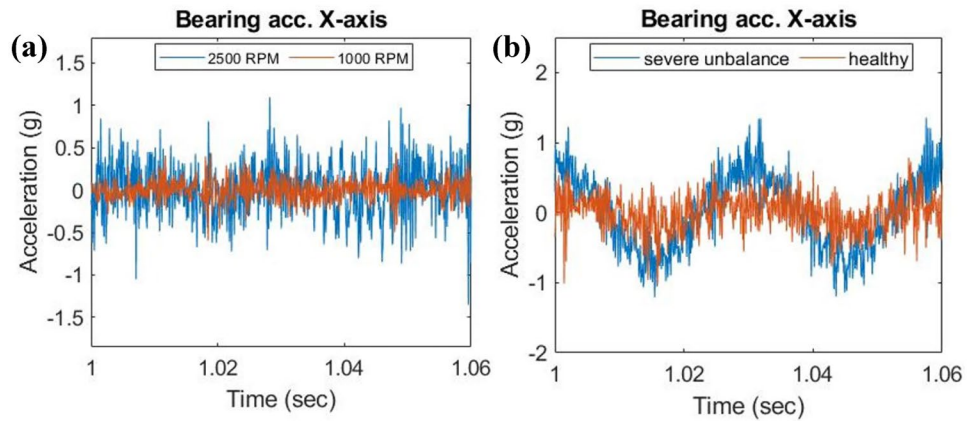
3.2 Sensor index evaluation

The complete list of possible control factors and features of the experiments is shown in Table 4. There are 12 input control variables (and 60 scenarios) and 44 features after primary

Table 4 The control and key feature index for all experiments

Sensors/controls	Features
Bearing operating parameters (total 12)	Shaft rotating speed (4) Unbalance level (3) Lubrication loss/lubrication contamination levels (5)
Tri-axis accelerometer at spindle (total 18)	RMS (3) Entropy (3) Spectrum at rotating frequency (3) Octave band induction No. 8, 9, 10 (9)
Tri-axis accelerometer at bearing (total 18)	RMS (3) Entropy (3) Spectrum at rotating frequency (3) Octave band induction No. 8, 9, 10 (9)
Microphone (total 4)	Crest factor (1) Octave band induction No. 8, 9, 10 (3)
Acoustic emission sensor (total 1)	Entropy
Current transformer (1)	Mean
Thermocouple (1)	Steady-state temperature
Infrared thermograph (1)	Steady-state temperature

Fig. 4 Accelerations of **a** healthy bearing at 1000 and 2500 rpm and **b** healthy and imbalanced bearings at 2000 rpm



screening from experimental observation and literature survey. However, the sensitivity and correlation of these features should be further identified to pick up the most dominated features. This is not a trivial issue since we believe that if all of the listed indexes are used to train AI model, the computational cost could be significantly increased and the training efficiency could be even worse [26, 27]. Consequently, the evaluation on the effectiveness of sensor features is important and selects proper features to train the subsequent AI models. After performing systematic feature sensitivity analysis, the most dominated features are determined for different types of fault and are shown below.

1. Unbalance of shaft

Rotating unbalance is traditionally an important issue to be addressed and machine learning technique has been applied [32]. After completing the experiments and subsequent sensor index extraction, the database of unbalanced shaft is then established. Figure 4 shows typical vibration of bearings at different rotating speeds and imbalance levels. It can be seen that the acceleration increases with both the speed and the imbalance level. The next issue is to perform correlation and sensitivity evaluation on the database. Eventually, proper sensor

features suitable for subsequent AI modeling are then identified. The effectiveness is mainly determined by experiment for each sensor feature. For example, Fig. 5 shows the mean values and standard deviations of kurtosis and skewness *w.r.t.* unbalance situations. Although these factors may have certain correlations with imbalance level, their results are considered poor in general. Thus, these two indexes are treated as ineffective. In contrast, Fig. 6 shows the acceleration amplitude of both *rms* in time domain and the power spectrum of the rotating speed in frequency domain. Both factors monotonically increase with unbalance and speed levels increases. Therefore, both of them could be treated as effective index the index for training the AI model subsequently. Based on the same examining procedure, the most effective features are extracted and listed in Table 5.

2. Loss of lubrication

Similar to the procedure for addressing the case of shaft unbalance, the significance of features is evaluated to pick up dominated indexes for reflecting lubrication loss. Figure 7 shows typical vibrations signal picked up by the bearing accelerometer in both time and frequency domains at 2000 rpm. From Fig. 7a, it can be

Fig. 5 **a** Kurtosis and **b** skewness in unbalanced conditions of the bearing acceleration

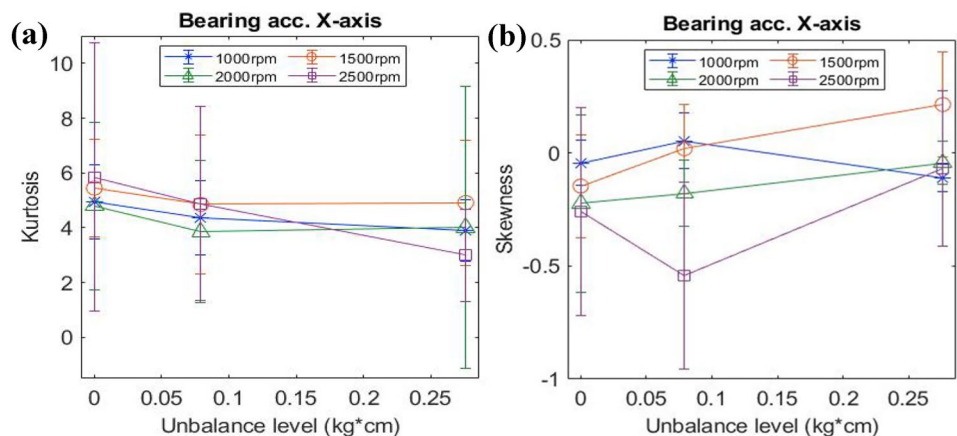
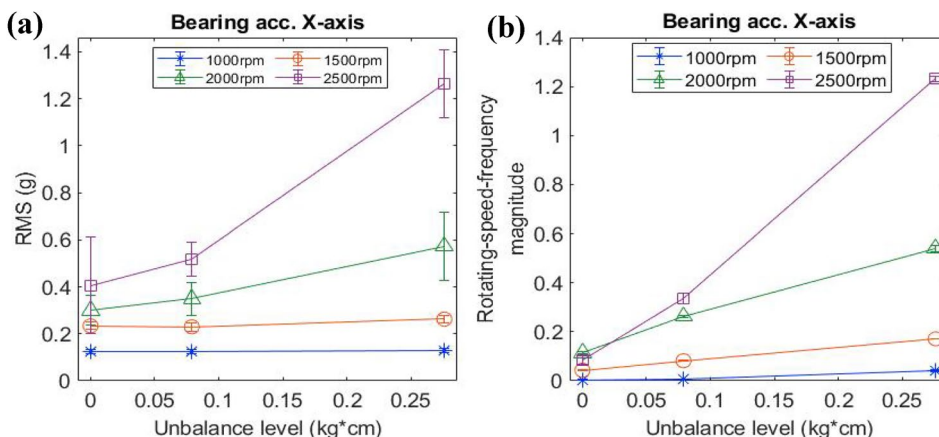


Fig. 6 Effective feature examples: the relationship between unbalance level vs. **a** *rms* and **b** the magnitude of rotating-speed-frequency of bearing acceleration signals



seen that the vibration level significantly increases for bearings without lubrication. On the other hand, as one can see in Fig. 7b, the power spectra distributions are different. These data are then used to extract features for subsequent machine learning. Examples of poor features are shown in Fig. 8, which indicates the correlations between the obtained kurtosis and crest factor of the acceleration signals and the lubrication level are not significant. In the meantime, Fig. 9a shows that the well correlation between the *rms* of the x-acceleration signals of the accelerometer mounted on the bearing and the lubrication level. Furthermore, as shown in Fig. 9b, the associated responses of octave band spectrum are presented. Although there are no obvious differences in low-frequency bands (i.e., band no. 1 to no. 6), high frequency bands (i.e., bands no. 8, 9, 10) indeed exhibit strong differences between different bearing lubrication status. Thus, those high-frequency bands could possibly serve as appropriate indexes. Based on the same procedure, the most effective features for counting lubrication loss are selected and listed in Table 5.

3. Lubrication contamination

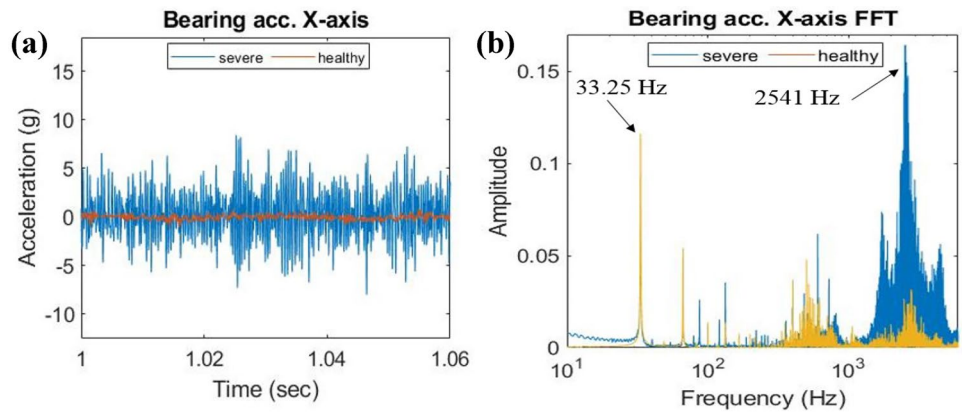
Finally, for the case of lubrication contamination, similar to the above cases, we also perform correla-

tion to find major sensor features. Again, typical data collected by the microphone are presented in Fig. 10. In Fig. 10a, it can be seen that the noise level significantly increases due to the presence of metal particle contaminations. On the other hand, the corresponding frequency power spectra, shown in Fig. 10b, also exhibit different distributions between normal and contaminated cases. These data are then used to extract features for subsequent computations. As shown in Fig. 11, it indicates that both the kurtosis and skewness of microphone signals are failed to be qualified indexes. On the other hand, similar to the lubrication loss case and as appeared in Fig. 12, the *rms* of x-acceleration component of the bearing accelerometer and the octave band spectrum of the microphone signals are treated as effective sensor features. Finally, through a series of correlation analysis, the most effective features are obtained and listed in Table 5. Notice that most of the effective features are the same for both lubrication loss and contamination cases and there are only some tiny differences to distinguish them. For example, as shown in Fig. 13, the tendencies of the *rms* and 8th octave band component for lubrication loss and contamination are similar but with certain quantitative differences. Such a difference might not be

Table 5 The most dominated features for this study

Faults	Sensors	Features
Unbalance	Accelerometers at spindle and bearing	Spectrum at rotating frequency (6)
Lubrication loss	Accelerometers at spindle and bearing	RMS (6) Octave bands 8, 9, 10 (18)
	Microphone	Octave band No. 8, 9, 10 (3)
Lubrication contamination	Accelerometer at spindle	RMS (3) Octave band No. 8, 9, 10 (9)
	Accelerometer at bearing	RMS (3) Entropy (3) Octave bands 8, 9, 10 (9)
	Microphone	Octave bands 8, 9, 10 (3)
	AE sensor	Entropy

Fig. 7 Accelerations of healthy and lack lubrication bearings at 2000 rpm. **a** Time domain data and **b** frequency power spectra



easily distinguished by engineers without the aid of artificial intelligent model. Here, as addressed previously, autoencoder would be hired to proceed this task.

4 Failure diagnosis using autoencoder models

Autoencoder is a kind of signal compression model based on neural network. As schematically shown in Fig. 14, it can reduce data dimension for achieving clustering of events through systematic data encoding and decoding [19]. It is a non-supervised learning network used to replace classical dimensionality reduction method such as principal component analysis (PCA) [33] to achieve better nonlinear feature compression. We utilize Python, and PyTorch [34] to create the autoencoder model for subsequent coding and simulation. In encoding process, the number of inputs equals to the used sensor indexes, which could be different based on faults to be deal. An architecture with two hidden layers with 128 and 64 neurons connected in the first and the second layer is used in this work. The current architecture is based on the recommendation from previous researches [35, 36]. Further alternation on the program architecture is possible toward deep autoencoder for better performance and would be our

immediate future work. In addition, the dimension of the final hyperspace could be arbitrary but is currently set as three for ease of visualization and three clustering variables (x,y,z) would be used to represent the clustering level. The optimal dimension for achieving best clustering still remains unknown and is left as our current study. Meanwhile, the decoding processing is symmetric to the encoding one and the whole process is schematically illustrated in Fig. 14. The models are trained using 128 epochs with a batch size of 64 data sets. The training process including two major steps: i.e., data forward propagation and error back propagation. One can refer to elsewhere for introducing the detail architecture of autoencoder [37]. Finally, as mentioned earlier, the feature dimension is reduced to three and visualized on the hyperspace space to evaluate the trained results.

Figure 15a shows the trained classification results for the three unbalance levels operated under the same lubrication loss at 1500 rpm. The models are trained by all addressed indexes without evaluations to exclude out improper indexes. One could see that the clustering result is relatively poor, where one could hardly identify the clustering results by the distribution in the hyperspace. Meanwhile, by utilizing the dominated indexes listed in Table 5, different imbalance levels can be clearly clustered as shown in Fig. 15b. Within

Fig. 8 Ineffective feature examples: **a** kurtosis of the accelerometer on the spindle and **b** crest factor of the acceleration on the bearing at different lubrication loss levels

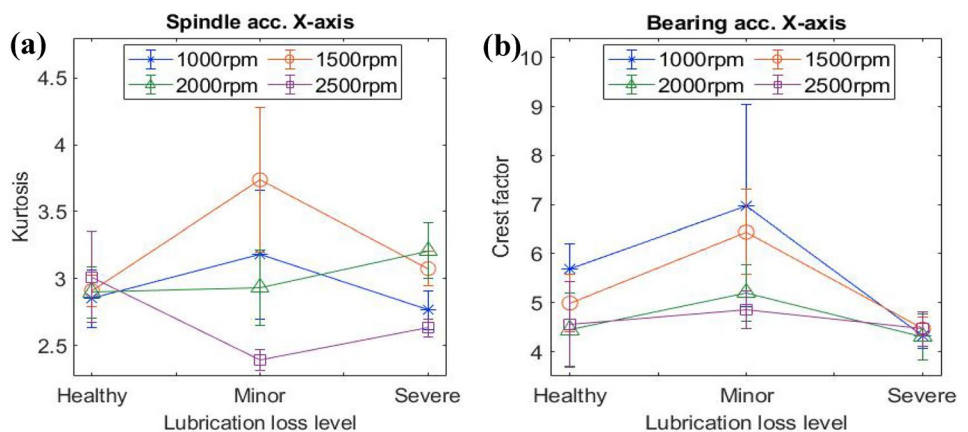
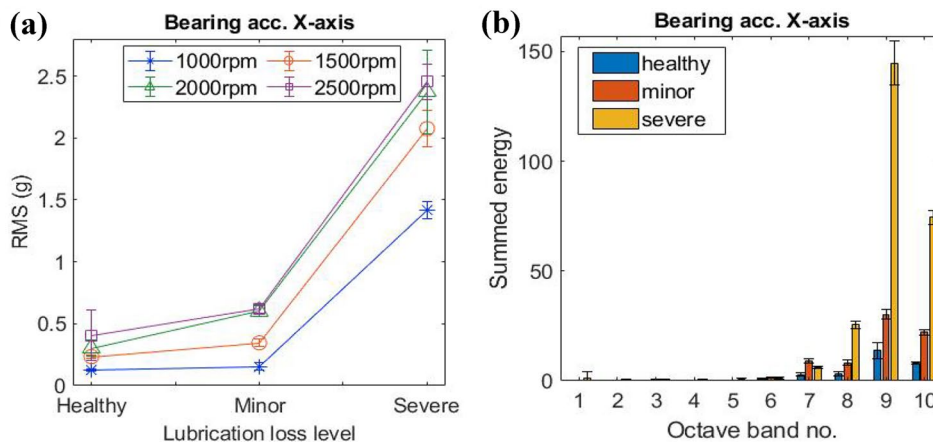


Fig. 9 Effective index examples: **a** *rms* and **b** octave band strength of the bearing accelerations in different lubrication loss levels



the hyperspace, clustered data for each imbalance levels concentrates at specific regimes and there are only a few extreme values away from those regions. This demonstrates the importance of using appropriate indexes for machine learning.

Next, it is possible to utilize the same setting to deal the lubrication loss and contamination problems for checking the effect of input features. The results for lubrication loss problem are shown in Fig. 16 and the model inputs have been listed in Table 5. The result indicates that if the most effective features are selected, the clusters are mainly concentrated in three well-separated regimes in the hyperspace, which implies that the three different lubrication levels can be well-classified apparently (seen in Fig. 16b). In contrast, for using all features as inputs, the separation level would not be so clear (seen in Fig. 16a). Same conclusion applies for the lubrication contamination case. As shown in Fig. 17, the results indicate the importance of proper selection of sensor features.

It is interesting to point out here that although the harshness levels of different types of bearing faults (i.e., lubrication loss and lubrication contamination in this work) could be clearly distinguished within their own domains, it is questionable if the scheme can distinguish two different types of fault. This concern is not a

trivial since the used indexes and the trends of them are extremely similar. However, as shown in Fig. 18, by utilizing the similar training process, different fault types in different levels could still be classified and the root-cause could be clearly identified.

Finally, for model validations, fresh experimental data are fed into those trained models (10% of all experimental data) and the prediction results are shown in Fig. 19, the prediction results of those new data for validation (labeled in color solid) are well grouped at the same regimes as the training data (labeled in hollow gray) in the hyperspace. This result validates the reliability of the designed autoencoder models.

5 Proposed model for quantify unsupervised machine learning schemes

Although the visualization in 3D hyperspace is straightforward for qualitatively examining the clustering performance, a quantitative evaluation is still required based on the following reasons: First, if the clustering is performed in a hyperspace with dimension larger than 3, the above-used visualization techniques cannot be used. Second, the judgment on the goodness of clustering are

Fig. 10 Microphone signal of healthy and contaminated bearings at 2000 rpm. **a** Time domain data and **b** frequency power spectra

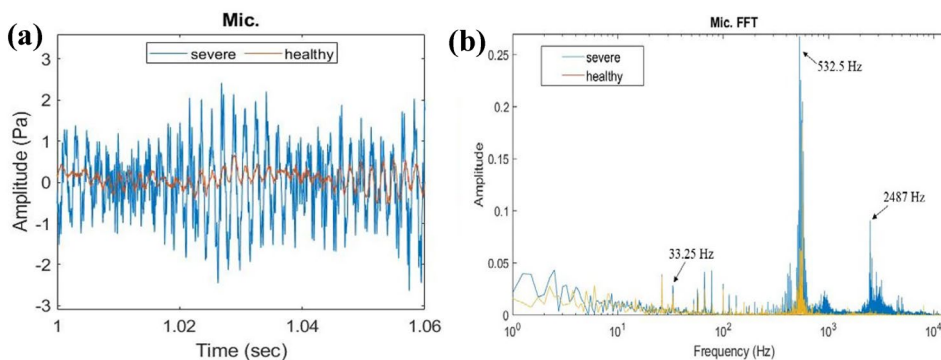
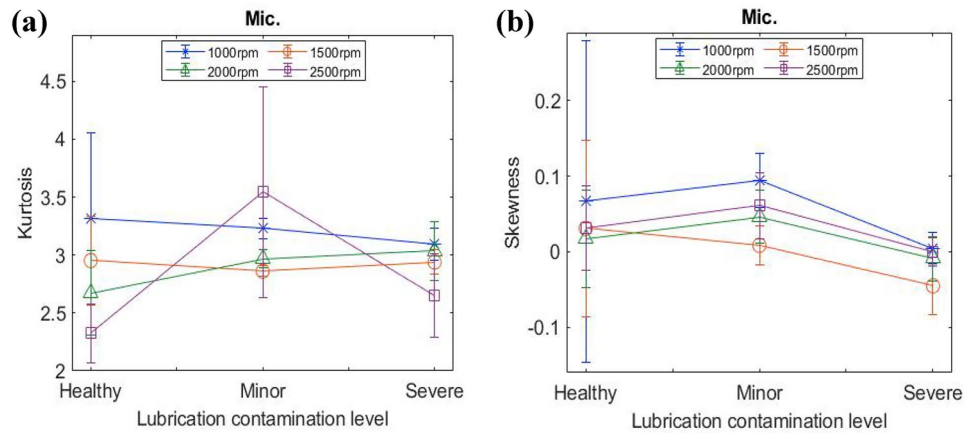


Fig. 11 Ineffective indexes examples: **a** kurtosis values and **b** skewness in different lubrication contamination levels of the microphone signals



highly subjective. In certain sense, such a subjective judgment could be difficult. For example, for the 2-D clustering cases schematically shown in Fig. 20a, it is hard to judge which classification is better. As a result, it requires a quantitative evaluation to address the performance of the non-supervised autoencoder model. Surprisingly this issue is not reported in previous works. To response this concern, a quantitative indicator, Q_{ij} , served as the cost function, is proposed for evaluation the clustering performance and is expressed as

$$Q_{ij} = \frac{d_{ij}}{2(std_i + std_j)} \tag{8}$$

In Eq. (8), d_{ij} means Euclidian distance between the centroid of clusters i and j ; std represents standard deviation of the position data in each cluster in the hyperspace; and indexes i, j represents different clusters (here $i, j = 1, 2$ or 3 and they represent healthy, minor fault, and severe fault data clusters) respectively. For a Gaussian distribution, the probability of data apart with its mean value within two standard deviation is 95.5%. Therefore, Eq. (8) is an indicator to measure the cleanness of separation level between clusters. That is, if $Q_{ij} > 1$, it is reasonable to declare that the two

clusters (i.e., i and j) are completely separated. On the other hand, for a Q_{ij} less than 1, it means that there could have certain probabilities these two clusters are tangled together and causes misjudgments. Figure 20b shows the schematic diagram in a 3-D hyperspace. Here, three different center to center distances d_{ij} and the standard deviation of each data set σ_i are calculated. The smallest Q_{ij} represents the worst separation level between two clusters and this value is chosen as the clustering performance. Table 6 shows the quantitative performance comparisons based on the proposed model for the three fault cases discussed in this work. For each fault cases, three runs are performed. It can clearly be seen that the Q factors for the situation with proper indexes are larger than those without feature screening, which indicates that a better classification. In particular, for the unbalance case, the Q factors are less than unity if the indexes are not properly screened. This implies that the clustering performance is inferior and one can compare the results shown in Fig. 16 and Table 6 for the argument.

It is also important to point out that the proposed approach can be extended to hyperspaces with any dimensions thus it provides an universal approach. With the proposed evaluation model, it is possible to examine the effect of different model parameters such as the number

Fig. 12 Effective indexes: **a** rms and **b** octave band strength of the acceleration and microphone signals on the ball bearing with different lubrication contamination degrees

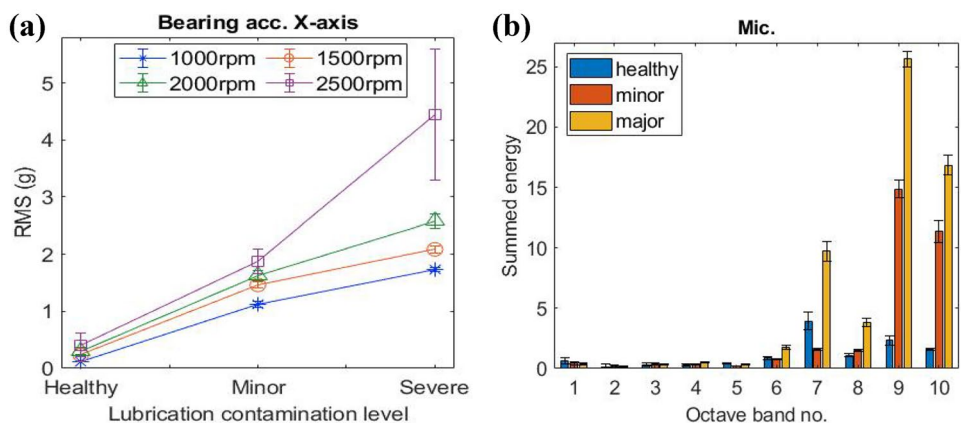
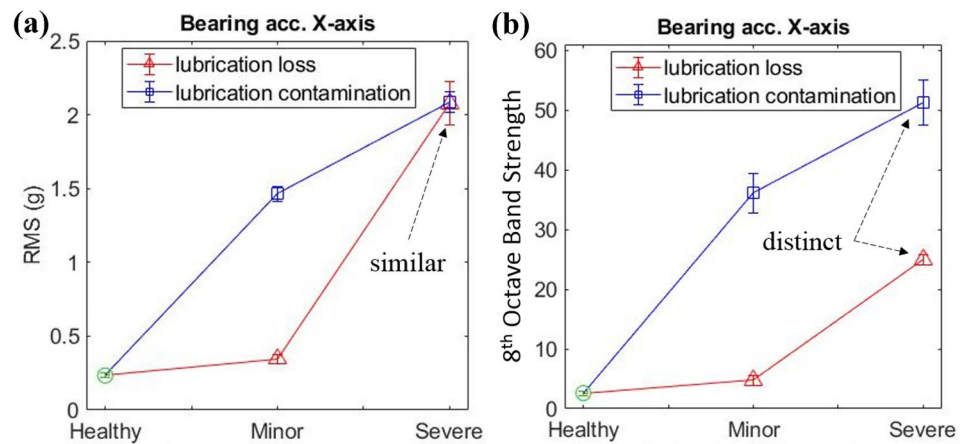


Fig. 13 The relationship between acceleration features of faulted bearing in different fault level of lubrication loss and contamination: **a** *rms* and **b** 8th octave band



of hidden layers, the number of neurons, and the final dimension of hyperspace, as well as the effect of input data features to the performance of machine learning schemes. Furthermore, as schematically sketched in Fig. 21, it is also possible to perform several autoencoder models sequentially to identify the possible multiple faults. Taking the investigated problem to identify a bearing fault containing both minor imbalance and lubricant contamination as an example, the first step is to examine if the system contain unbalance fault (and the severity). Then, one can use the same data with different features to examine the possibility of lubrication loss and subsequently, the lubrication contaminations by examining if the data is well within the corresponding clusters. This provides a systematic flow for investigating faults containing multiple causes in the future.

6 Discussions

Using machine learning in diagnosis of machine and components has been widely investigated and is believed as a promising approach in modern system longevity

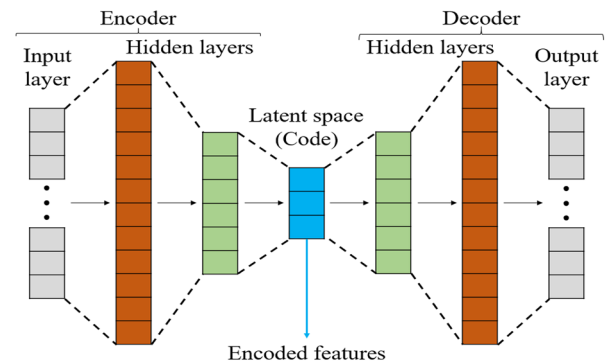


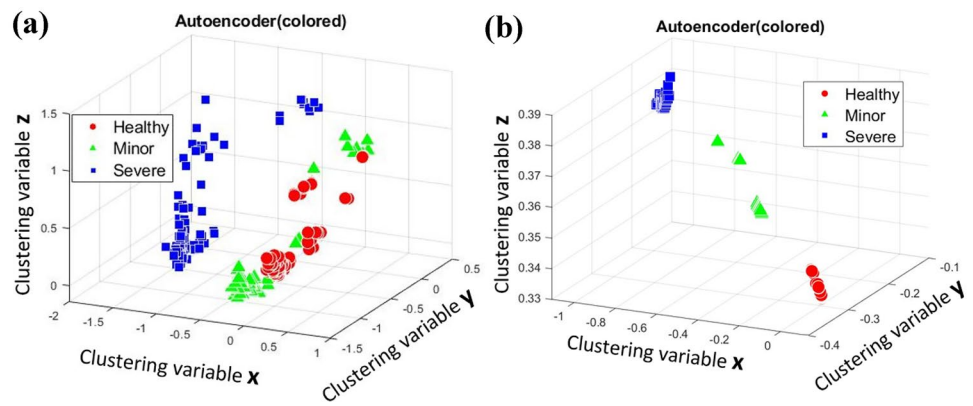
Fig. 14 A schematic diagram for the procedure of autoencoder

assessment and maintenance scheduling. However, the performance still largely depends on the selected sensor features. In addition, the unsupervised machine learning approach, although it does not need labeling in training, the clustering results lack a quantification comparison and thus it brings an obstacle for subsequent scheme optimization.

Here a platform of high speed rotating shaft with multiple sensors is set up for performing the investigation and ball bearings with five different lubrication fault conditions or levels are designed. These bearings are worked with 3 unbalance levels at 4 rotating speeds to generate data for subsequent autoencoder model development. The study results clearly indicate the importance of choosing effective sensor features on the machine learning performance. With effective sensor features determined via experiments, difference imbalance or bearing fault levels can be clearly classified. The model is powerful such that different fault category can also be identified. To better quantify the clustering performance, an index based on the measure of separation level between clusters in hyperspace is proposed. This provides further opportunities to optimizing the clustering performance of unsupervised machine learning schemes.

To the best of our knowledge, there is no similar work presented before for addressing the above two concerns and this could be the major innovative contribution of this work. One could further optimize autoencoder or hire other unsupervised machine learning schemes to study the bearing failure problems or utilize the concept proposed in this work in dealing with related engineering fault analyses. Notice that although some sensor indexes seem not effective in this work, it could possibly become critical if other types of faults are considered. Nevertheless, one can follow the same procedure developed for evaluating the situation and discussing

Fig. 15 Trained clustering result with **a** all indexes and **b** most sensitive features for unbalance-only situation at 1500 rpm



fault scenario individually to make the diagnosis system more comprehensive.

In short, we believe that this work provides the several novelty and contribution to the related fields. First, it provides an experimental flow in finding significant features for bearing and to prove that using correct features are critical. Second, it utilizes unsupervised machine learning scheme to classify both different types of cause and the degree of damage in failure diagnosis. Finally, this work proposes a new qualification method to measure the goodness of classification for machine damages.

In the near future, for the current bearing fault diagnosis problem, investigations on the effect of hyperspace dimension and the effect on the limitation of available sensors would be investigated and optimized based on

the proposed Q parameter. Finally, it is hoped that the proposed approach is helpful for quantifying performance of unsupervised machine learning schemes and thus imposing contributions on other engineering system fault diagnosis. A specific example is that the proposed methodology could be used for machine tools and high speed spindles failure predictions, where the failures of ball bearings are identified as a dominated failure mode. Since the selection of proper sensors and their index, as well as the install locations, are crucial to the performance, our developed sensing module [38] could possibly be mounted at the interior of spindle for further optimizing the sensor performance. By such an integration, the diagnosis performance should be improved in comparison with our pervious used ANN scheme in machine tool wear experiment [27].

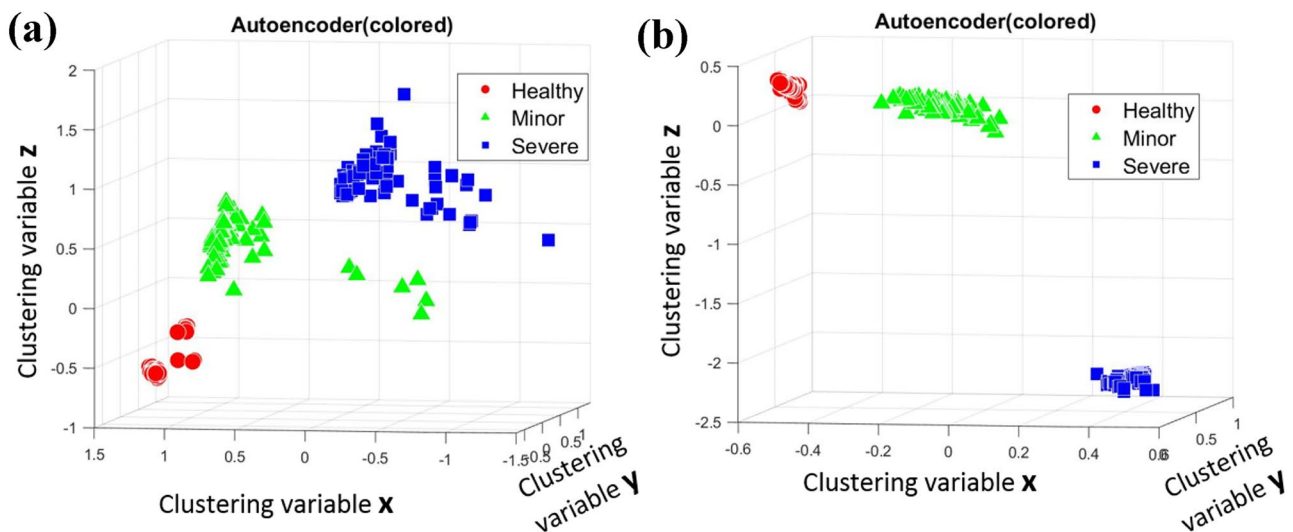


Fig. 16 Trained clustering results using **a** all indexes and **b** most sensitive features for lubrication loss-only condition at 1500 rpm

Fig. 17 Trained results with **a** all unscreened indexes and **b** most sensitive indexes for lubrication contamination classification with no unbalance at 2000 rpm

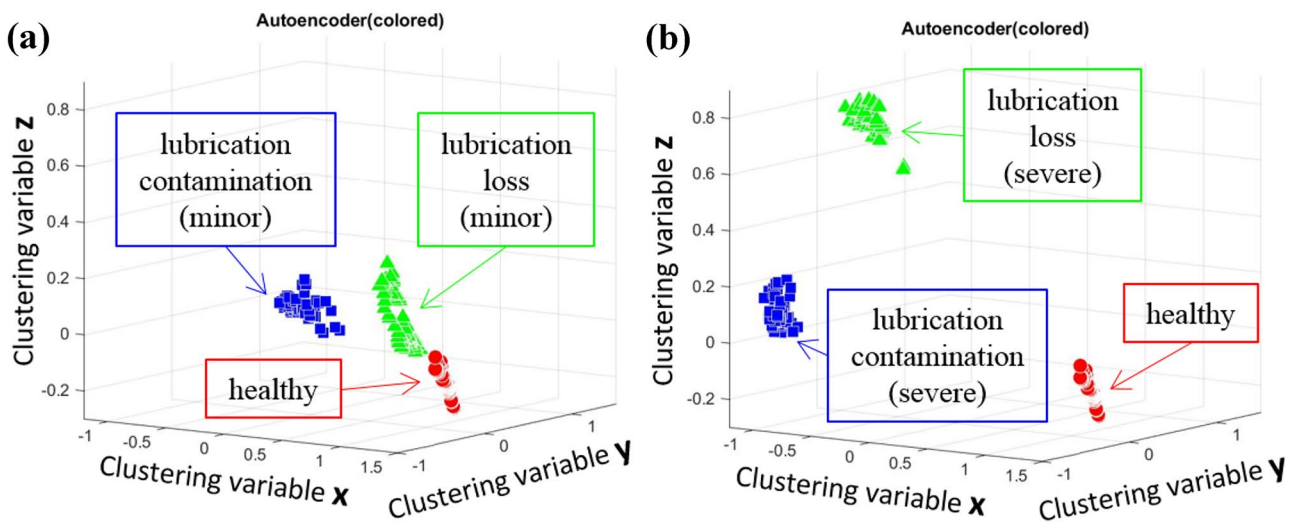
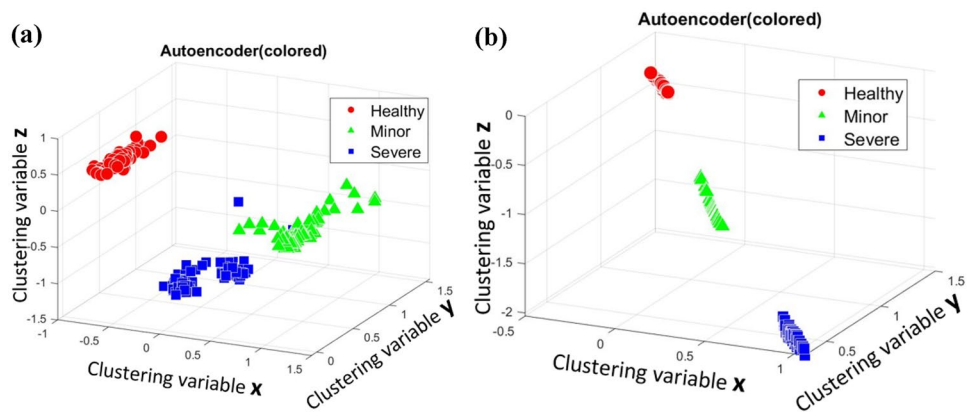
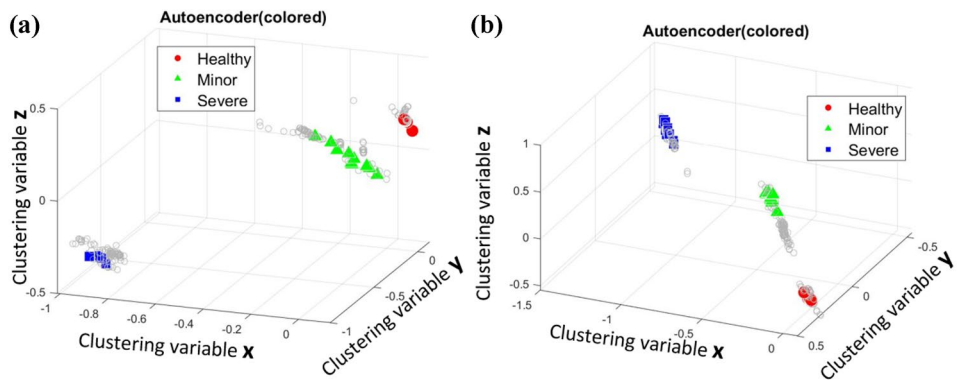


Fig. 18 Trained results for distinction between lubrication loss and contamination: **a** minor and **b** severe cases

Fig. 19 Validated results for **a** imbalance level at 2500 rpm and **b** lubrication loss level at 2000 rpm



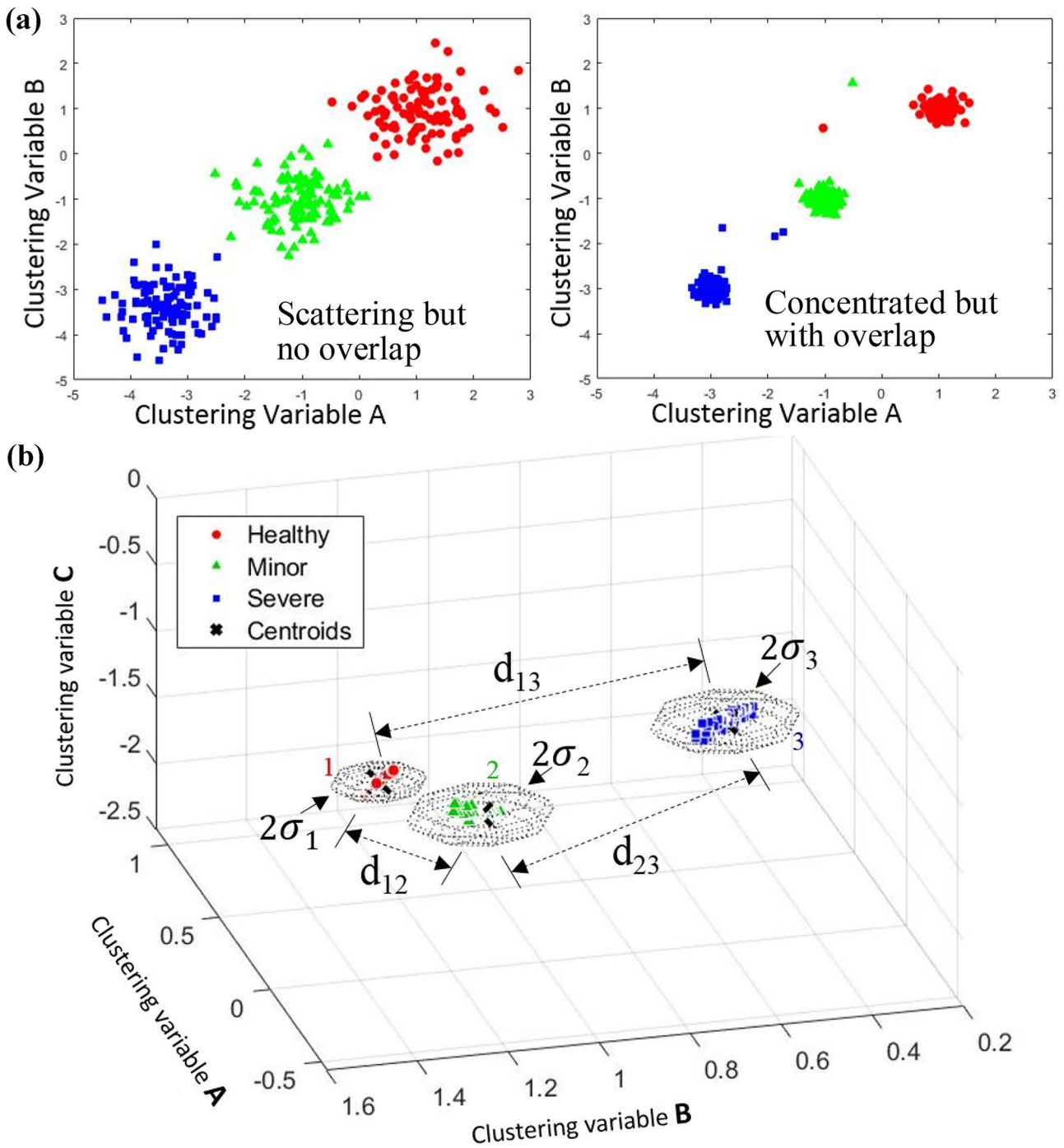


Fig. 20 Schematic diagrams to illustrate **a** the problem on evaluating the clustering performance and **b** physical insight of the proposed quantitative indicator

Table 6 The average Q -factor of three cases in both selected and unselected feature situations of bearings runs at 1500 rpm

	Unbalance fault	Lubrication loss	Lubrication contamination
Avg. Q (selected features)	1.51	1.81	3.32
Avg. Q (all features)	0.45	1.55	2.01

Acknowledgements Technical supports from Prof. M-C Tsai of NCKU are also greatly appreciated.

Funding This study was financially supported by the Ministry of Science and Technology (MOST) of Taiwan government with project IDs: 109-2622-8-006-005, 108-2221-E-006-209, 108-2221-E-006-209, and 110-2221-E-006-172, is acknowledged.

Availability of data and material Available upon request.

Code availability Available upon request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All the authors agree with the publication.

Conflict of interest The authors declare no competing interests.

References

1. Barlow RE, Hunter LC (1960) Optimum preventive maintenance policies. *Oper Res* 8:90–100
2. Pham H, Wang H (1996) Imperfect maintenance. *Eur J Oper Res* 94:425–438
3. Wang H (2002) A survey of maintenance policies of deteriorating systems. *Eur J Oper Res* 139:469–489
4. Qin J, Liu Y, Grosvenor R (2016) A categorical framework of manufacturing for industry 4.0 and beyond. *Procedia Cirp* 52:173–178
5. Wang K (2016) Intelligent predictive maintenance (IPdM) system—Industry 4.0 scenario. *WIT Trans Eng Sci* 113:259–268
6. Zaretsky E, Poplawski J, Miller CR (2000) Rolling bearing life prediction: past, present, and future. NASA Tech Rep 210529
7. Zaretsky E (2013) Rolling bearing life prediction, theory, and application. NASA Tech Rep 2013-215305
8. Harris TA, McCool JI (1996) On the accuracy of rolling bearing fatigue life prediction. *ASME J Tribol* 118:297–309
9. Mishra C, Samantaray A, Chakraborty G (2017) Ball bearing defect models: a study of simulated and experimental fault signatures. *J Sound Vib* 400:86–112
10. Gao Q, Duan C, Fan H, Meng Q (2008) Rotating machine fault diagnosis using empirical mode decomposition. *Mech Syst Signal Process* 22:1072–1081
11. Kankar PK, Sharma SC, Harsha SP (2011) Fault diagnosis of ball bearings using continuous wavelet transform. *Appl Soft Comput* 11:2300–2312
12. Schoen RR, Habetler TG, Kamran F, Bartfield RG (1995) Motor bearing damage detection using stator current monitoring. *IEEE Trans Ind Appl* 31:1274–1279
13. Elasha F, Greaves M, Mba D (2018) Planetary bearing defect detection in a commercial helicopter main gearbox with vibration and acoustic emission. *Struct Health Monit* 17:1192–1212
14. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. *Neurocomputing* 234:11–26
15. Sutrisno E, Oh H, Vasani A, Pecht M (2012) Estimation of remaining useful life of ball bearings using data driven methodologies. *Proc Prognostics Health Manag* 1–7

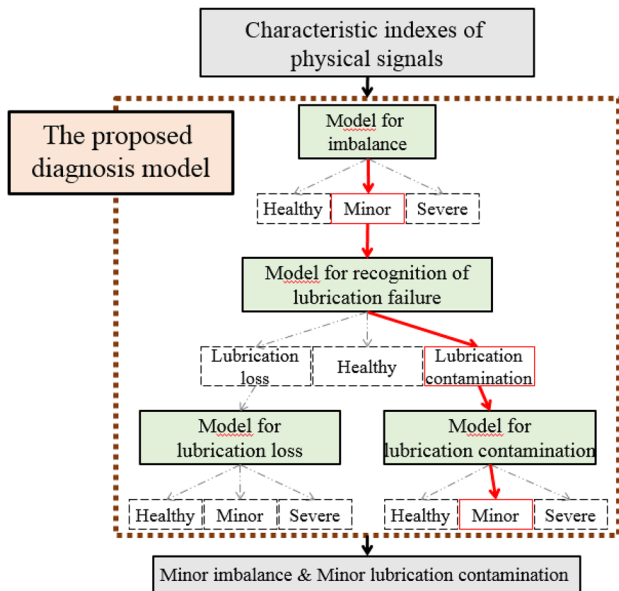


Fig. 21 The proposed status identifying flow chart

7 Summary and conclusion

Machine learning based fault diagnosis gradually plays important roles in modern smart manufacturing. Unsupervised machine learning, owing its nature with no needs on pre data labeling on training, becomes increasingly popular. However, it lacks a quantitative measure to evaluate the effect of feature selection and other settings and parameter adjustments. In this work, by utilizing ball bearing failure and autoencoder models as the test bed, the effect of feature selection based on physical effectiveness evaluation on the AI prediction performance is investigated. It clearly indicates the importance of selecting effective feature on improving the AI diagnosis performance. Meanwhile, a model is also proposed to quantify the clustering performance of unsupervised machine learning. With the model in hand, it is possible to use it as the basis for evaluating the performance and for optimizing the structure of unsupervised machine learning schemes.

16. Guo L, Li N, Jia F, Lei Y, Lin J (2017) A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing* 240:98–109
17. Li B, Chow MY, Tipsuwan Y, Hung JC (2000) Neural-network-based motor rolling bearing fault diagnosis. *IEEE Trans Industr Electron* 47:1060–1069
18. Kankar PK, Sharma SC, Harsha SP (2011) Fault diagnosis of ball bearings using machine learning methods. *Expert Syst Appl* 38:1876–1886
19. Ren L, Sun Y, Cui J, Zhang L (2018) Bearing remaining useful life prediction based on deep autoencoder and deep neural networks. *J Manuf Syst* 48:71–77
20. Lu W, Li Y, Cheng Y, Meng D, Liang B, Zhou P (2018) Early fault detection approach with deep architectures. *IEEE Trans Instrument Measure* 67:1679–1689
21. Ma M, Sun C, Chen X (2018) Deep coupling autoencoder for fault diagnosis with multimodal sensory data. *IEEE Trans Ind Inform* 14:1137–1145
22. Deng J, Zhang Z, Eyben F, Schuller B (2014) Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process Lett* 21(9):1068–1072
23. Lore KG, Akintayo A, Sarkar S (2017) LLNet: a deep autoencoder approach to natural low-light image enhancement. *Pattern Recogn* 61:650–662
24. Merrill N, Eskandarian A (2020) Modified autoencoder training and scoring for robust unsupervised anomaly detection in deep learning. *IEEE Access* 8:101824–101833
25. Shao H, Jiang H, Zhao H, Wang F (2017) A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. *Mech Syst Signal Process* 95:187–204
26. Lee N, Azarian M, Pecht M (2020) Octave-band filtering for convolutional neural network-based diagnostics for rotating machinery. *Ann Conf PHM Soc* 12:9
27. Sun I-C, Cheng R-C, Chen K-S (2021) Evaluation of transducer signature selections on machine learning performance in cutting tool wear prognosis. *Int J Adv Manuf*. <https://doi.org/10.1007/s00170-021-08526-w>
28. Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S (2013) Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern* 43:982–994
29. Cheng RC, Chen KS, Liu YH, Chang LK, Tsai MC (2021) Development of autoencoder-based status diagnosis method for ball bearing tribology status monitoring. *Proc. 9th IIAE International Conference on Industrial Application Engineering 2021 (ICIAE 2021)*, Kitakyushu, Japan, p. 45–52
30. Cheng RC (2021) Development of autoencoder-based unsupervised fault recognition method for application in bearing condition diagnosis. Master Thesis, National Cheng-Kung University, Taiwan
31. Schervish MJ (2012) *Theory of statistics*. Springer Science & Business Media
32. Jia F, Lei Y, Lu N, Xing S (2018) Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mech Syst Signal Process* 110:349–367
33. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. *Chemom Intell Lab Syst* 2:37–52
34. Subramanian V (2018) *Deep learning with PyTorch: a practical approach to building neural network models using PyTorch*. Packt Publishing Ltd
35. Ding Y, Ma L, Ma J, Suo M, Tao L, Cheng Y, Lu C (2019) Intelligent fault diagnosis for rotating machinery using deep Q-network based health state classification: a deep reinforcement learning approach. *Adv Eng Inform* 42:100977
36. Pihlgren GG, Sandin F, Liwicki M (2020) Improving image autoencoder embeddings with perceptual loss. *Proc. 2020 IEEE International Joint Conference on Neural Networks (IJCNN)*, p. 1–7
37. Abouzid H, Chakkor O (2020) Autoencoders in deep neural network architecture for real work applications: convolutional denoising autoencoders. In *Handbook of Research on Recent Developments in Electrical and Mechanical Engineering*, IGI Global, p. 214–236
38. Tsai JM, Sun IC, Chen KS (2021) Realization and performance evaluation of a machine tool vibration monitoring module by multiple MEMS accelerometer integrations. *Int J Adv Manuf Technol* 114:465–479

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.