Journal of Vibration Engineering & Technologies (2024) 12:1991–2001
https://doi.org/10.1007/s42417-023-00959-9

ORIGINAL PAPER

Lab-Scale Vibration Analysis Dataset and Baseline Methods for Machinery Fault Diagnosis with Machine Learning

Bagus Tris Atmaja[1,2] · **Haris Ihsannur**[1] · **Suyanto**[1] · **Dhany Arifianto**[1]

Received: 26 December 2022 / Revised: 24 March 2023 / Accepted: 27 March 2023 / Published online: 27 May 2023

Abstract

Motivation The monitoring of machine conditions in a plant is crucial for production in manufacturing. A sudden failure of a machine can stop production and cause a loss of revenue. The vibration signal of a machine is a good indicator of its condition.

Purpose This paper presents a dataset of vibration signals from a lab-scale machine. The dataset contains four different types of machine conditions: normal, unbalance, misalignment, and bearing fault. Three machine learning methods (SVM, KNN, and GNB) evaluated the dataset, and a perfect result was obtained by one of the methods on a onefold test.

Results The performance of the algorithms is evaluated using weighted accuracy (WA), since the data are balanced. The results show that the best-performing algorithm is the SVM with a WA of 99.75% on the fivefold cross-validations. The dataset is provided in the form of CSV files in an open and free repository at https://zenodo.org/record/7006575.

Keywords Vibration data · Vibration analysis · Predictive maintenance · Machine condition monitoring · Anomaly detection · Machine learning

Introduction

Vibration analysis is the process of evaluating the vibration characteristics of a machine or structure, typically with the goal of identifying any problems or abnormalities that may be present. Vibrations are often indicative of the health and performance of a machine or structure and can provide valuable information about the condition of certain components, such as bearings, gears, and motors. By analyzing the characteristics of vibrations, such as frequency, amplitude, and waveform, it is possible to identify potential problems or failures that may occur in the future. The analysis of vibration is often performed in the frequency-domain, since the pattern of abnormalities in this domain is more obvious than in the time-domain.

Vibration signals convey more information than others for predictive maintenance, a maintenance technique based on the condition of machines. Other techniques are oil (lubricant) analysis [1], infrared thermography [2], and sound pattern analysis [3–5]. Vibration and lubricant analysis were the most common techniques for predictive maintenance (PdM) [6]. PdM, which is developed in the 1970s, is an advancement of preventive maintenance, time-based maintenance from the 1950s [7]. Vibration analysis is a key predictive maintenance technique (among others), since it can identify the problem of machines before they become too serious and cause unscheduled downtime [1].

Current technologies in vibration analysis lack in many aspects. The requirement for a large amount of data is challenging [8] and can be difficult for real machines (for treating machines in different conditions). The accuracy and reliability of vibration analysis depend on the quality of the sensors and measurement equipment being used. Vibration analysis requires specialized knowledge and expertise to interpret the

✉ Bagus Tris Atmaja
b-atmaja@aist.go.jp

Haris Ihsannur
harisihsannur@gmail.com

Suyanto
suyanto@ep.its.ac.id

Dhany Arifianto
dhany@ep.its.ac.id

[1] Department of Engineering Physics, Sepuluh Nopember Institute of Technology, ITS Sukolilo Campus, Surabaya 60111, Jawa Timur, Indonesia

[2] National Institute of Advanced Industrial Science and Technology, Tsukuba 3058560, Japan

data and identify potential problems [8]. Finally, vibration analysis can be expensive due to the specialized equipment and software required. This paper is presented to tackle the limitations of vibration analysis above.

The use of machine learning to replace expert engineers has been tried previously, including deep learning methods. In [8], the authors used self-organizing maps (SOM) for a leak detection problem based on vibration signals. A jump to using deep learning for vibration-based machinery fault detection has been tempted in Refs. [9–11]. In [9], the authors proposed a deep transfer learning based on vibration data converted to images from VGG-16 to their vibration data. In [10], the authors reviewed machine deep learning methods for vibration analysis: auto-encoder (AE), deep belief network (DBN), deep Boltzmann machines (DBM), convolutional neural network (CNN), and recurrent neural network (RNN). In [11], the authors proposed a switchable normalization CNN for bearing fault detection. In this light, we saw a gap between previous old SOM methods and recent deep learning methods.

Since the vibration patterns (the extracted feature) of each machine condition are distinct [8], it is arguably better and enough to use machine learning instead of deep learning. Machine learning seeks to find relatively small data patterns given the features instead of patterns in the data itself, as in deep learning (which also was usually used to extract the features). If the extracted features from vibration signals are informative enough to distinguish the machine conditions, machine learning can be used to detect the machine conditions. In this paper, we evaluated three machine learning, namely support vector machine (SVM), *K*-nearest neighbors (KNN), and Gaussian Naive Bayes (GNB), to detect the machine conditions-based features extracted from vibration signals.

Instead of evaluating machine learning only, the need for a free dataset for vibration analysis is mainly addressed. The dataset is collected from a lab-scale vibration analysis experiment. Lab-scale data are often used in vibration analysis to validate and verify the results of simulations and calculations, as well as to confirm the performance of new designs or technologies. Lab-scale testing allows researchers and engineers to study the behavior of a system under controlled conditions, which can provide valuable insights into the problem and its potential solution for the system. Lab-scale testing can also be used to assess the feasibility and reliability of a design or technology before it is implemented on real data. This lab-scale data can help and allow designers and engineers to make necessary adjustments or improvements for the development of vibration analysis tools.

This paper, hence, contributes in two aspects. First, we provided a free vibration dataset in CSV format that can be downloaded directly from the open repository. Second, we provided baseline methods with machine learning to detect the machine conditions based on the vibration signals. Furthermore, we show that our evaluation of the dataset and methods achieves a near-perfect accuracy on fivefold cross-validation and a perfect accuracy on onefold test data based on a set of distinct features, highlighting the effectiveness of the proposed dataset and methods.

Previous Works: The Available Datasets

Research on vibration analysis has been conducted progressively over the years. However, the lack of reports on the available datasets for vibration analysis is a challenge. The available datasets are not well documented, and some datasets are not available for free. The focus of this section is to review the datasets for vibration analysis based on the literature.

Machinery fault diagnosis (MaFaulDa) is the openly available vibration analysis dataset from the Signals, Multimedia, and Telecommunications Laboratory, Universidade Federal do Rio de Janeiro. The dataset contains 1951 samples from six conditions: normal, horizontal misalignment, vertical misalignment, imbalance, underhang bearing, and overhang bearing. The apparatus for data collection is SpectraQuest's Machinery Fault Simulator (MFS) Alignment-Balance-Vibration (ABVT) system. The data are collected at a sample rate of 51.2 kHz from two IMI sensors (model 601A01 and 604B31). In addition to vibration data, the dataset also provides tachometer signals (to estimate rotation frequency) and microphone data (Shure SM81). This dataset was used in several pieces of research, such as in [12–15].

Case Western Reserve University (CWRU) [16] bearing fault dataset is a dataset designed for examining normal and faulty conditions of the ball bearing. The experiments to obtain the dataset were conducted using a two-horsepower (hp) Reliance Electric motor, and acceleration data were measured at locations near to and remote from the motor bearings [17]. The faults on the motor bearing were created using electro-discharge machining (EDM). The CWRU dataset contains bearing faults ranging from 0.007 inches in diameter to 0.040 inches in diameter and was introduced separately at the inner raceway, rolling element (i.e., ball), and outer raceway. The vibration data were recorded for motor loads, in which faulted bearings were installed, of 0–3 horsepower (motor speeds of 1797–1720 RPM). This dataset were used in [11, 14, 15, 18].

Another toy dataset is the Accelerometer Dataset, which is developed by Mackenzie Presbyterian University, Sao Paulo, Brazil, for predictions of the failure time of a cooling fan [19]. As an apparatus is a cooling fan (Akasa AK-FN059) with weights on its blades was used to generate vibrations. This fan cooler was attached an accelerometer, MMA8452Q accelerometer, to collect the vibration data.

There are 153,000 records (lines) of vibration data in this dataset.

It is interesting to see that each dataset has a different goal. The MaFaulDa dataset is the closest to our dataset for machine condition diagnosis, with differences in apparatus and number of data. Here, we used a lab-scale vibration analysis apparatus (real electrical motors) to collect the vibration data. The vibration data are collected from five machines with different conditions. We designed the dataset to fill the gap in the available datasets for vibration analysis. Another dataset, such as the CWRU dataset, focuses on bearing only, while Accelerometer Dataset focuses on predicting the failure time of a cooling fan.

Dataset: VBL-VA001

In this section, we present the VBL-VA001 dataset, our first public lab-scale vibration analysis (VA) dataset developed at VibrasticLab (VBL), Department of Engineering Physics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia.

Apparatus and Recorded Conditions

The need for a Lab-scale vibration analysis dataset is triggered by the difficulties of obtaining the actual data from a real plant. Treating a machine to fail will cost a lot of money and time. Using lab-scale data will minimize the impact of interfering with machine operation while keeping the same fault pattern as the real plant data. In this light, we simulated four common conditions of machine operation by electric motors (water pumps) to collect their vibration patterns.

The electric motors as apparatus to replicate industrial machines are electrical motors for water pump type Panasonic GP-129JXK. This type of machine is an induction-type motor with a single phase. The source voltage is 220 V with 50 Hz frequency. The output power is 125 watts with two poles. The speed of the motor is fixed at 3000 RPM.

Five electrical motors were designed to replicate four machine conditions: a machine for a condition except for unbalance, where two machines simulate different weights of unbalance conditions. The configuration of these five machines is shown in Fig. 1. The first machine is with misalignment; the second is with the normal condition; the third is with 27 g cm unbalance; the fourth is with bearing fault; and the fifth is with 6 g cm unbalance.

The unbalance condition is given in two different mass additions to the impeller. The impeller has a diameter of 6 cm. Hence, adding 4 g of mass from 1.5 of center mass (eccentricity) will cause an unbalance of 6 g cm ($4 \times 1.5 = 6$). Consequently, adding 18 g of mass from 1.5 of center mass will cause an unbalance of 27 g cm ($18 \times 1.5 = 27$). These configurations are retained from the previous work [20, 21]. Figure 2a, b shows these unbalance conditions.

Fig. 1 Five electrical machines with different fault conditions for experiments: (1) misalignment, (2) normal, (3) unbalance 27 g cm, (4) bearing fault, and (5) unbalance 6 g cm

The misalignment condition is given by coupling the shaft with an additional metal cylinder. The metal cylinder is 1 cm in diameter and 7 cm in length. The misalignment is 3 mm offset from the center. This misalignment condition was set between the original shaft and an additional metal cylinder with a metal cylinder offset from the shaft. Figure 2c shows the misalignment condition.

The bearing fault condition was set by hitting the outer ring of the bearing with a hammer. Hence, the fault is caused by the impact of the hammer (crack-like faults). Although only the outer ring was hit by the hammer, both rings (inner and outer) showed to be in fault conditions simultaneously from the spectrum visualization. It is not possible to detect either inner or outer ring faults only in this dataset, since the goal is to detect the general pattern of bearing faults. Future research may be able to detect the specific fault of the bearing, which is one of the most faults in the machine (like dataset in [17]).

To deal with the uncertainties data, in this case, data in unstable conditions, we measured the vibration signals several minutes after the start (around 10 min by observing the vibration waveform). Uncertainties undoubtfully exist for such problems, like in vibration signals. By selecting the only robust vibration signal from the motors, it is expected that there are no uncertain input parameters that affect model outputs. Since the model is less complex (nine inputs and several output/model parameters) and less complex model leads to small uncertainty in the model outputs [22], the effect of uncertainty in the model could be neglected. Future studies could accommodate this issue, as well as take into account uncertainties using computational and experimental approaches (e.g., using the energy of mechanical systems as the loss function to machine learning methods [23]).

Table 1 summarizes our VBL-VA001 dataset. In total, there are 4000 vibration samples, with 1000 samples for each condition. This number of samples is required to train machine learning methods. The choice of four conditions

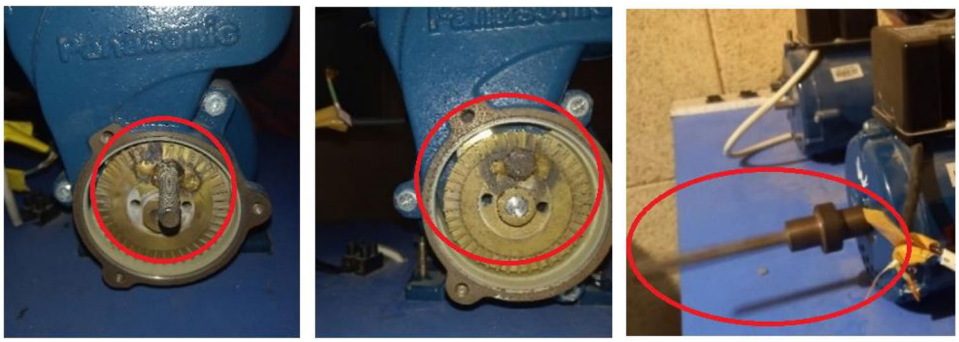

Fig. 2 Faulty condition for unbalance and misalignment: **a** adding 18 g of mass on the impeller, **b** adding 4 g of mass on the impeller, and **c** coupling shaft with an additional metal cylinder

Table 1 Data distribution of VBL-VA001

Condition		# Samples
Normal		1000
Unbalance	6 g cm	500
	27 g cm	500
Misalignment	3.0 mm	1000
Bearing fault	Outer ring	1000
Total		4000

Table 2 Comparison of VBL-VA001 with other datasets

Dataset	# Samples	Sampling freq. (kHz)	# Classes
VBL-VA001	4000	20	4
MaFaulDa	1951	50	4
CWRU	161	12 and 48	2

(normal, unbalance, misalignment, and bearing fault) is that those four conditions are the most common ones in rotating machine operations. Comparing the existing dataset (Table 2), VBL-VA001 is the largest dataset with the most number of samples. The VBL-VA001 dataset is available at https://zenodo.org/record/7006575#.Y5wlTafP2og. From the original IDE format, we provided our data in CSV format for convenience.

Sensor Placement

The sensor is located in the machine (electrical machine/water pump) in the position, as shown in Fig. 3. The vibration sensor is "LOG-0002-100G-DC-8GB-PC Shock and Vibration Sensor". To attach it to the machine, we used double-sided tape to mount the sensor into the machine and connected it to the PC with a USB cable. The recording process was done using enDAQ LAB Software. By this arrangement, we collected three axes of acceleration data that comply with the standard (ISO13373-1, 2002). The sensor is located on the rigid part and close to the vibration source. The first

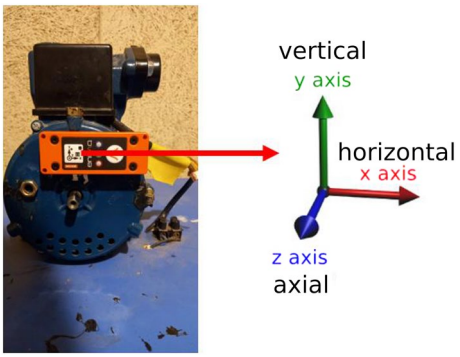

Fig. 3 Sensor placement for vibration measurements

Table 3 Excerpt of data collected by the sensor (acceleration in *g*)

Time	*x*	*y*	*z*
0.004516	− 0.102961	0.030537	0.114270
0.004566	− 0.118802	− 0.020894	0.123060
0.004616	− 0.110881	− 0.046609	0.114270
0.004666	− 0.102961	− 0.053038	0.105480
0.004716	− 0.087121	− 0.059466	0.096690
..	..	..	..
4.999968	0.039601	− 0.083574	− 0.101085

consideration (rigid surface) is to avoid resonance, while the second consideration (close to the source) is to minimize the effect of the transmission path. The data were recorded every 5 s. The example of data collected by the sensor is shown in Table 3.

Machine Learning Methods

The flow of vibration data processing is shown in Fig. 4. The vibration data are first preprocessed to convert from time-domain signals to frequency-domain signals. We performed data normalization after that since machine

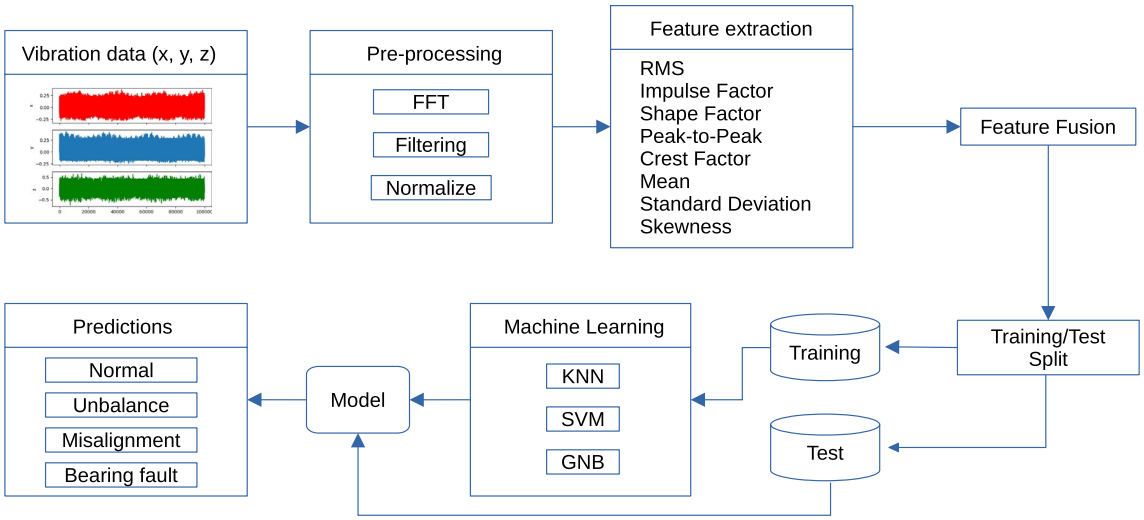

Fig. 4 Flowchart of processing the vibration data with machine learning methods; the filtering process in pre-processing removes NaN (not-a-number) values; each feature in the feature extraction process has three values (x, y, z); hence, the total feature (feature fusion) has 27-dim (9 features × 3 axes)

learning methods are sensitive to the range of input data. Then, the features are extracted. The extracted features are then used to train the machine-learning model. The trained model is then used to predict the fault condition of the machine in the test data.

Pre-processing and Feature Extractions

FFT

The main pre-processing data are FFT and filtering. FFT is used to convert the time-domain signal into a frequency-domain signal. We used the FFT package from Numpy for this purpose with the default configuration. Figure 5 shows

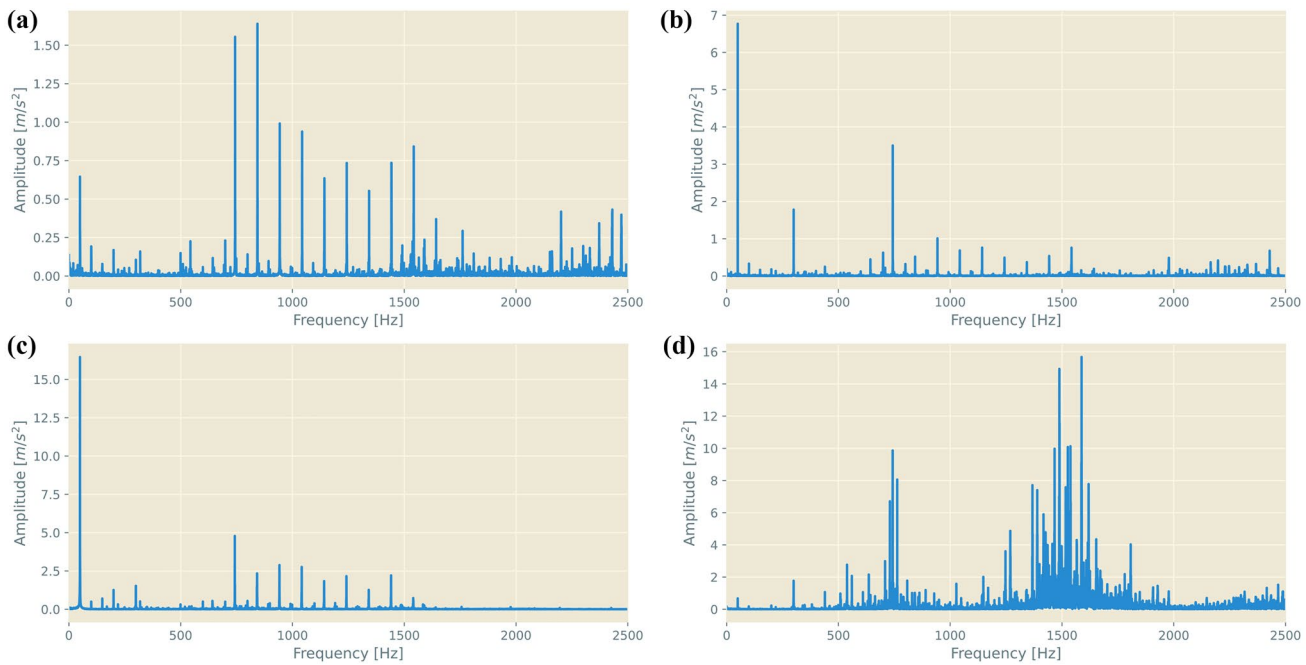

Fig. 5 Spectrum of vibration signal in each machine condition: **a** normal, **b** unbalance, **c** misalignment, and **d** bearing fault

the results of FFT for each machine condition (shown for the z-axis only). It can be seen that the amplitude of the normal condition is lower than the faulty condition. In that figure, we did not normalize the amplitude of the signal to show the difference between each machine condition. However, the limit of frequency is set to 2500 Hz for clarity of comparison. From the original data provided in acceleration (g) units, we showed the unit in m/s^2.

Normalization

Since machine learning methods are sensitive to the range of input data, we normalized the data after merging all samples. We used the following formula to normalize the data:

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}, \tag{1}$$

where x_{norm} is the normalized data, x is the original data, x_{min} is the minimum value of the data, and x_{max} is the maximum value of the data. The normalized data are then used to extract the features.

Extracted Features

The vibration data after the FFT process are simplified using nine feature extraction methods, as shown in Table 4. The goal of the feature extraction is for dimensional reduction as well as finding a correlation between the specific patterns of machine conditions. The features are statistics extracted from the frequency domain signal (x, y, z) and then merged into one feature vector. Statistical descriptors are known to be useful for SVM-based machine condition classification [24]. In this research, the feature vector has 27 dimensions (9 features × 3 axes) compared to 44 dimensions in the aforementioned literature.

Finally, we performed filtering by removing outlier data using Pandas' 'dropna' method. This step is performed after the feature extraction process and before feeding extracted features to machine learning methods. These outlier data are recorded by the measurements process that returns NaN values. It is not clear why the NaN values are recorded, but it is safe to remove them, since they could trigger an error when using the machine learning tool.

Table 4 Feature extraction methods

Feature	Formula
Mean ($\bar{x}$)	$\frac{1}{N}\sum_{i=1}^{N} x_i$
Standard deviation (std)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$
Root mean square (RMS)	$\sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}(x_i)^2}$
Peak to peak (PP)	$x_{\text{max}} - x_{\text{min}}$
Impulse factor (IF)	$\frac{x_{\text{max}}}{\bar{x}}$
Skewness (S)	$\frac{1}{N}\sum_{i=1}^{N}\frac{\left(x_i - \bar{x}\right)^3}{\sigma^3}$
Kurtosis (K)	$\frac{1}{N}\sum_{i=1}^{N}\frac{\left(x_i - \bar{x}\right)^4}{\sigma^4}$
Crest factor (C)	$\frac{\lvert x_{\text{max}}\rvert}{\text{RMS}}$
Shape factor (SF)	$\frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_i}$

Classifiers

Machine learning is a field of artificial intelligence that involves training computers to perform tasks without explicit programming. It is based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. At the heart of machine learning is a classifier, the method to classify inputs into outputs. There are a lot of methods (classifiers) developed for machine learning and counting. In this study, we evaluated three machine learning methods, namely support vector machine (SVM), K-nearest neighbors (KNN), and Gaussian naive Bayes (GNB).

Support Vector Machine

Support vector machines (SVMs) are a type of supervised learning algorithm that can be used for classification (support vector classification, SVC) or regression (support vector regression, SVR). The goal of an SVM is to find the hyperplane in a high-dimensional space that maximally separates the two classes. An SVM model is trained by finding the hyperplane that has the greatest distance (called the margin) between the two classes. Data points that are closest to the hyperplane are called support vectors and have the greatest influence on the position of the hyperplane. In this study, we used SVC to classify the machine condition.

One of the key strengths of SVMs is their ability to use kernels, which are functions that can transform the data into a higher dimensional space in which it may be more linearly separable. This allows SVMs to model complex relationships in the data and can lead to improved performance. This study used the radial basis function (RBF) kernel, which is the default kernel in the scikit-learn implementation of SVM [25]. We optimized the regularization hyper-parameter C constant (in a range [0, 100]) with onefold and fivefold

cross-validation. The best C value is selected based on the highest accuracy score during the training phase.

K-Nearest Neighbors

K-Nearest neighbors (KNN) is a machine learning algorithm that is used for classification and regression. It works by finding the K nearest data points to a given data point and using those points to make a prediction. KNN is a simple and effective algorithm that is easy to implement and works well on a variety of datasets. However, it can be computationally expensive, especially for large datasets, and it can be sensitive to the choice of K. In this study, we optimized the number of neighbors K (in a range [1, 100]) with onefold and fivefold cross-validation as SVM.

Gaussian Naive Bayes

Gaussian naive Bayes (GNB) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is a simple and effective algorithm that is easy to implement and works well on a variety of datasets. The algorithm works using the training data to estimate the probability of each class, as well as the probability of each feature given a class. When given a new data point, the algorithm uses these probabilities to predict the class that is most likely to be associated with the data point. As in previous classifiers, we optimized the main hyper-parameter in GNB, that is `var_smoothing` (in a range $[10^{-1}, 10^{-100}]$) with onefold and fivefold cross-validation.

The methods described above (feature extraction and the classifiers) are implemented in Python (tested on Python 3.7.4) using scikit-learn [25], Numpy [26], and Pandas libraries with simple procedural implementations (no object-oriented programming and avoid a large loop). The source code is available at https://github.com/bagustris/VBL-VA001.

Results and Discussion

We split the results and their discussions into two parts: extracted features and classification results. The former aims to show the differences among nine features at different axis vibration data for each machine condition. The latter shows the overall accuracy results of the three classifiers.

Distinct Feature on Different Machine Condition

Figures 6, 7 and 8 show the plot of nine feature values for different machine conditions on the x-, y-, and z-axes, respectively. It can be shown that, in general, our proposed features can discriminate a different condition of each

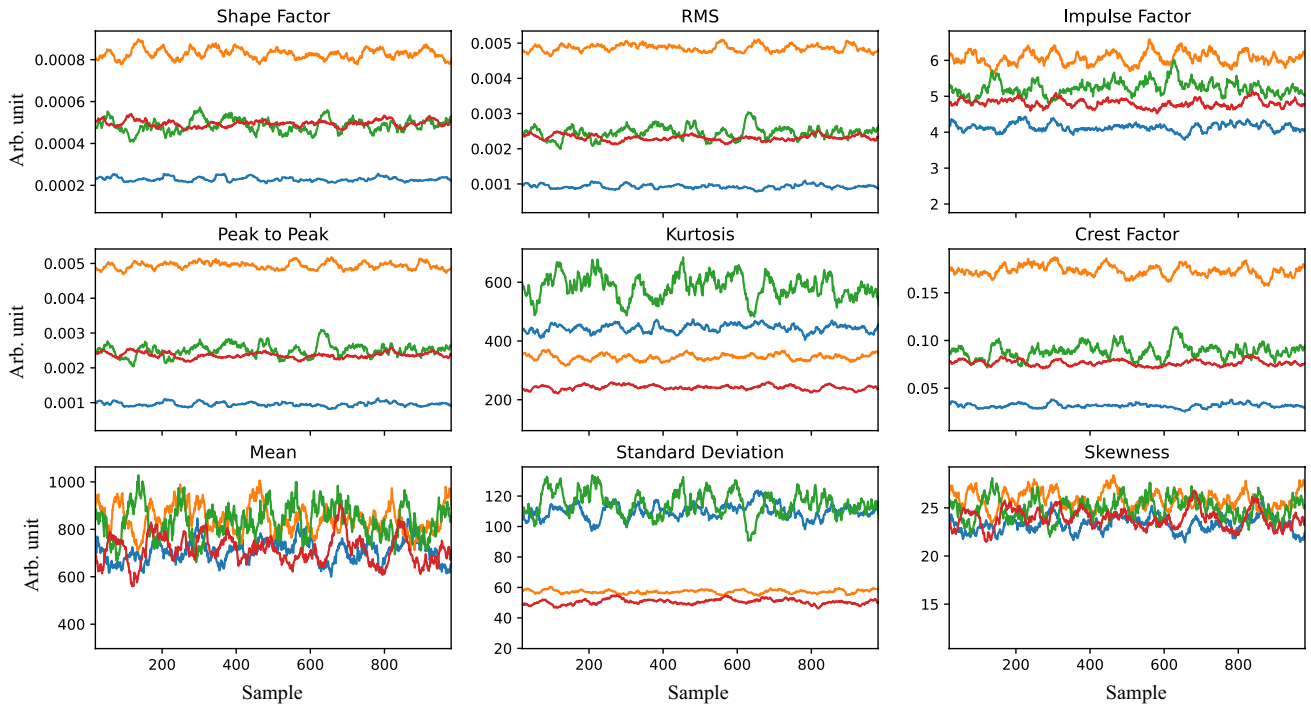

Fig. 6 Plots of different feature values on the x-axis and the corresponding machine condition; blue: normal, orange: misalignment, green: unbalance, and red: bearing fault (color figure online)

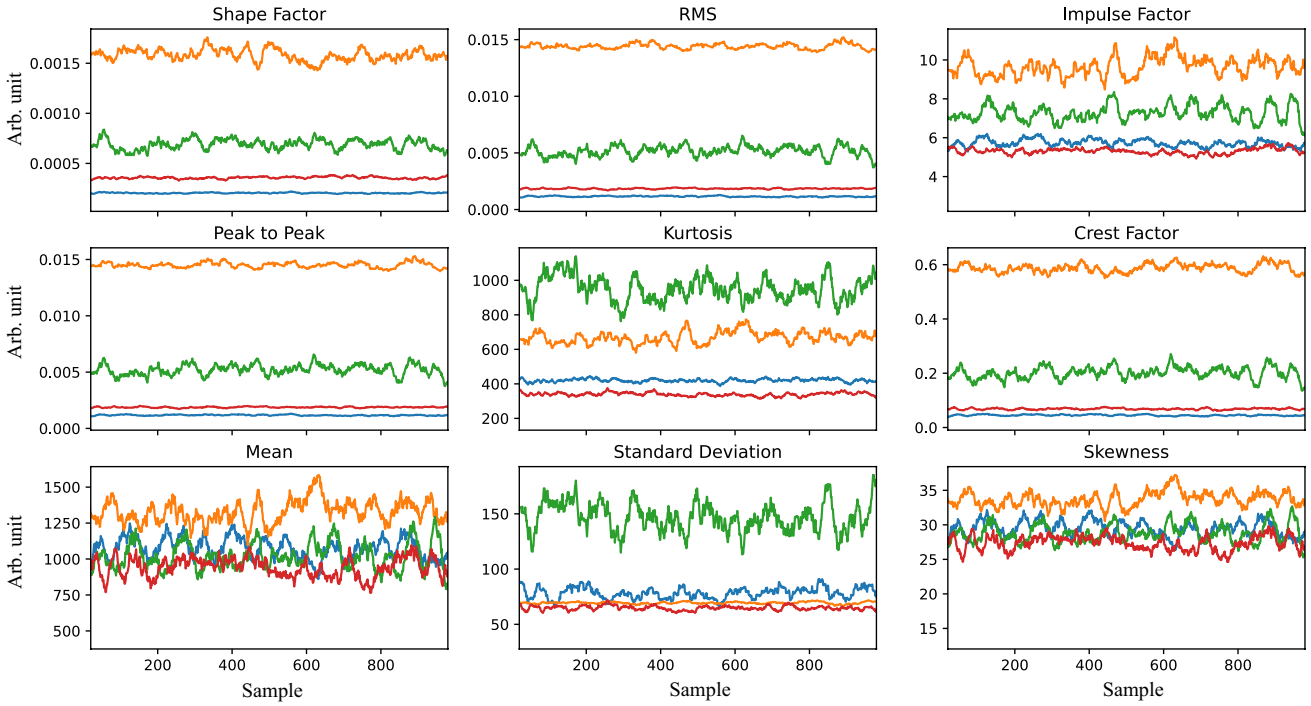

Fig. 7 Plots of different feature values on the *y*-axis and the corresponding machine condition; blue: normal, orange: misalignment, green: unbalance, and red: bearing fault (color figure online)

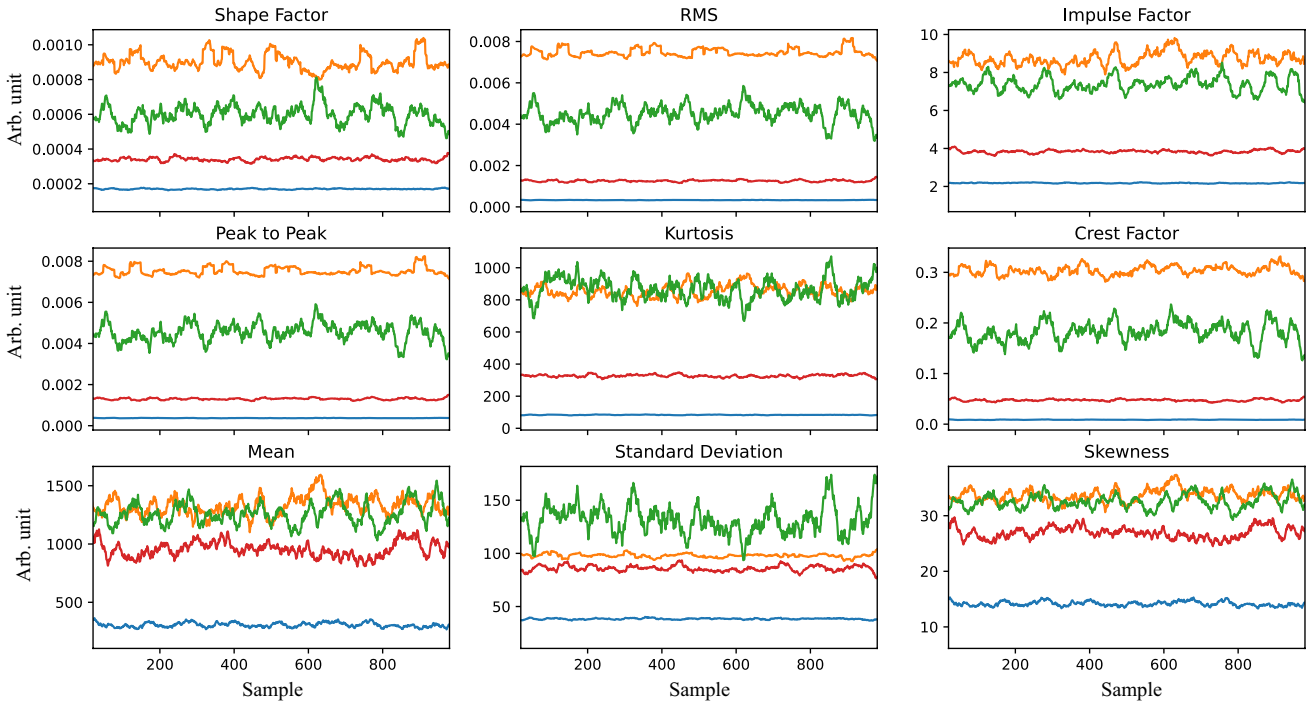

Fig. 8 Plots of different feature values on the *z*-axis and the corresponding machine condition; blue: normal, orange: misalignment, green: unbalance, and red: bearing fault (color figure online)

machine. The most distinct features were observed in the *y*-axis where shape factor, RMS, impulse factor, peak-to-peak, kurtosis, crest factor, and standard deviation of each machine condition are separable. Only mean and skewness features are confused among the machine conditions. By this observation, we can expect that the machine learning methods will be able to classify the machine condition with high accuracy.

Since the size of features is small, we can include all features from all axes. Also, the visualization of feature values among machine conditions shows distinct characteristics of each machine condition. Therefore, we also assumed no need to perform feature selection. However, all of those assumptions need verification by machine learning methods.

Classification Results

The proposed balanced dataset allows us to evaluate the machine learning methods with a single metric, weighted accuracy (WA, same as unbalanced accuracy or overall accuracy). In the first step, we optimize the main hyper-parameter values for each classifier (Fig. 9), report accuracy results on onefold test and fivefold cross-validation (Table 5), and show the confusion matrix on onefold test data (Fig. 10).

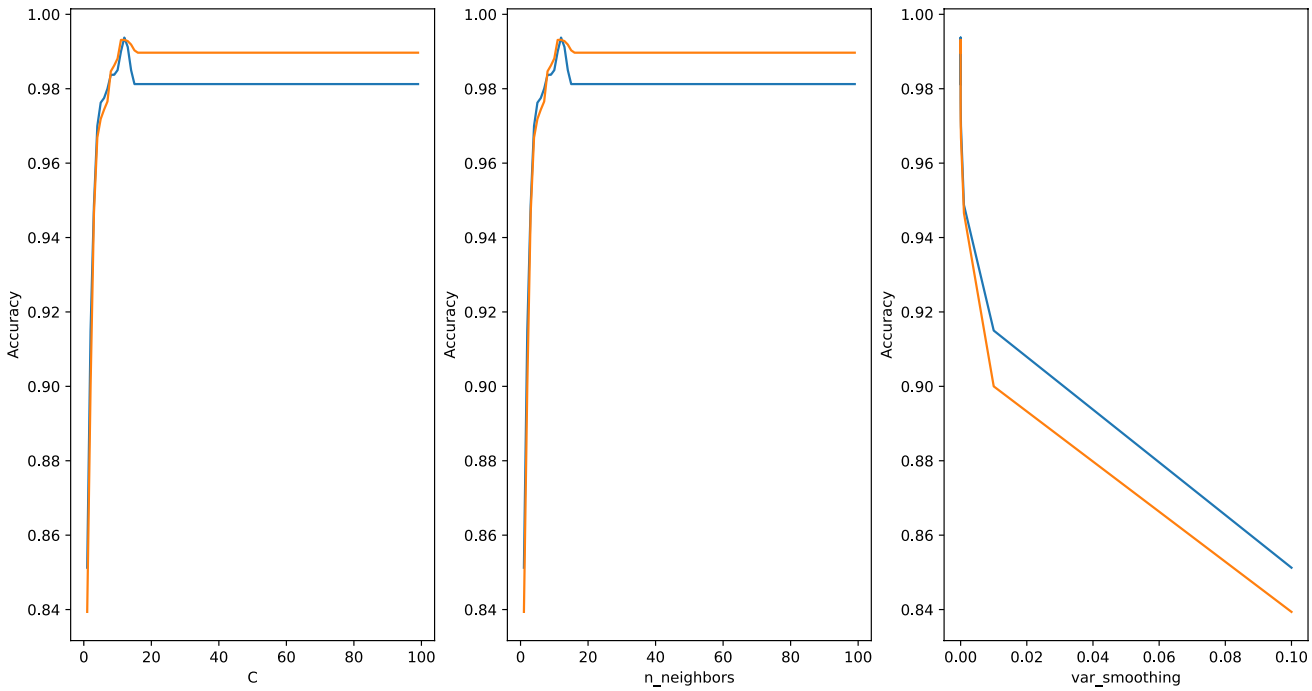

Fig. 9 Hyper-parameter optimization results of SVM, KNN, and GNB; orange: training data; blue: test data. For onefold (shown above), the optimal hyper-parameters are 69 (C for SVM), 5 (K for KNN), and 10^{-11} (var_smoothing for GNB) (color figure online)

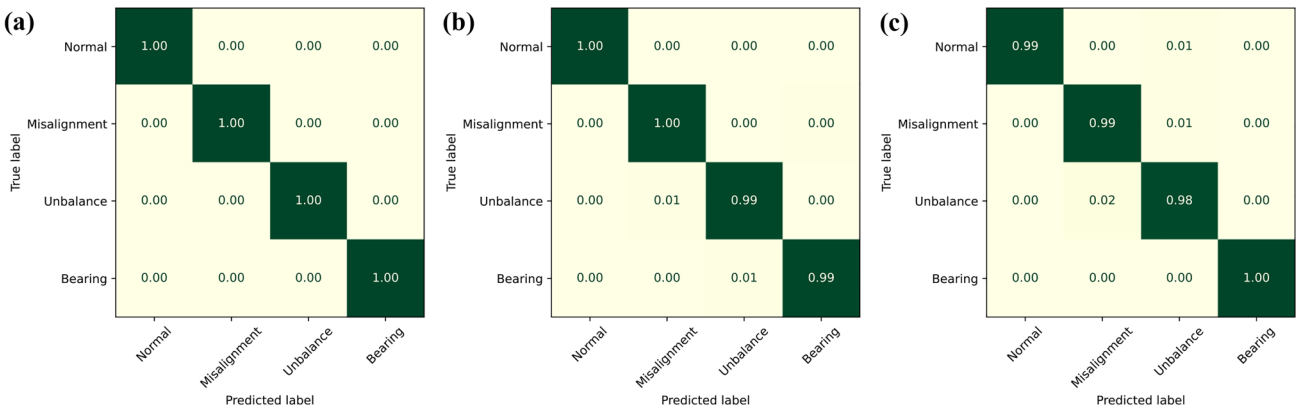

Fig. 10 Confusion matrix on onefold test data for SVM (**a**), KNN (**b**), and GNB (**c**)

Table 5 Overall accuracy (%) on onefold and fivefold test data for SVM, KNN, and GNB

Classifier	Onefold	Fivefold
SVM	100	99.75
KNN	99.625	99.525
GNB	99.5	99.375

Table 5 shows the accuracy of onefold test and fivefold cross-validation. It is clear in both cases the trend is the same. The highest performing classifier is SVM, followed by KNN and GNB. Given a very high accuracy on both fixed split and cross-validation (100% and 99.75%), our model showed great potency for real-case deployment beyond lab-scale experiments.

Figure 9 shows the results of the hyper-parameter optimization of onefold validation on three different machine learning methods: SVM, KNN, and GNB. We obtained the best hyper-parameters for this fixed split (onefold) with $C = 69, K = 5$, and `var_smoothing` $= 11$, for SVM, KNN, and GNB, respectively. For the fivefold (figure is not shown), the best hyper-parameters are $C = 93, K = 1$, and `var_smoothing` $= 13$. The best hyper-parameter for fivefold is achieved by determining the maximum average of fivefold tests on the given hyper-parameter values search range.

Figure 10 shows the confusion matrix for onefold test. Perfect accuracy was obtained by the SVM method in this single-set test. In all classes, the recall was 100% (shown as 1.00). The accuracies obtained by SVM in fivefold are [99.875%, 99.635%, 99.5%, 99.875%, 100%]. The worst case is 4 of 800 test data are incorrectly predicted by SVM (accuracy of 99.5%). In a real implementation, the number of data for measurements can be added to mitigate this error. For instance, a single sample only contains five-second vibration data. More data for the same machine can be collected to increase the confidence level of the classifier. A recall of 0.01 corresponds to 2 samples, and a recall of 0.02 corresponds to 4 (from 200) samples in each class (for KNN and GNB).

The computation time for the classification process is also short for SVM, KNN, and GNB. Given the input features in a CSV file (not the original vibration data), it only takes a minute to obtain the model's accuracy. The only process that takes longer time is the extraction process which takes about 10 min. The computation process was done on a PC with Intel Core i9-10850K CPU and 64GB RAM.

Conclusion

In this paper, we present a new dataset for vibration analysis (machine condition classification) recorded from electric pumps in a laboratory environment. The dataset is balanced in four classes of machine conditions and contains 1000 samples for each condition. The evaluated conditions are normal, unbalance, misalignment, and bearing fault. We provided three classifiers as a baseline: SVM, KNN, and GNB. The inputs for the classifiers are nine statistical functions derived from the spectrum of vibration signals, which show the distinct pattern for each machine condition. The results show that SVM has the best performance in this dataset, which achieve overall accuracy of 99.75% in fivefold cross-validation. The high accuracy obtained by the SVM shows the potential use of the proposed dataset for machine learning research. Future research can be explored to obtain perfect accuracy and improve the classifiers' robustness and generalization, perhaps beyond the lab-scale environment.

Acknowledgements The authors would like to thank enDAQ for providing calibrated vibration sensor, LOG-0002-100G-DC-8GB-PC Shock and Vibration Sensor, and data acquisition system (enDAQ LAB) used in this research. Parts of this research were supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan, under Project No. JPNP20006.

Declaration

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Girdhar P (2004) Practical machinery vibration analysis and predictive maintenance. Elsevier, Oxford
2. Bagavathiappan S, Lahiri BB, Saravanan T, Philip J, Jayakumar T (2013) Infrared thermography for condition monitoring—a review. Infrared Phys Technol 60(April):35–55. https://doi.org/10.1016/j.infrared.2013.03.006
3. Delgado-Arredondo PA, Morinigo-Sotelo D, Osornio-Rios RA, Avina-Cervantes JG, Rostro-Gonzalez H, Romero-Troncoso RdJ (2017) Methodology for fault detection in induction motors via sound and vibration signals. Mech Syst Signal Process 83:568–589. https://doi.org/10.1016/j.ymssp.2016.06.032
4. Glowacz A (2018) Acoustic based fault diagnosis of three-phase induction motor. Appl Acoust 137:82–89. https://doi.org/10.1016/j.apacoust.2018.03.010
5. Atmaja BT, Arifianto D (2009) Machinery fault diagnosis using independent component analysis and instantaneous frequency. In: Proceeding international conference on instrumentation, communication information technology and biomedical engineering. ITB, Bandung. https://doi.org/10.1109/ICICI-BME.2009.5417257. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5417257
6. Moya MDCC (2007) Model for the selection of predictive maintenance techniques. INFOR Inf Syst Oper Res 45(2):83–94. https://doi.org/10.3138/infor.45.2.83
7. Shozo Tanaka (2015) Life cycle maintenance. JR EAST Tech Rev 22(54):29–44
8. Ypma A (2001) Learning methods for machine vibration analysis and health monitoring. Ph.D. thesis, TU Delft
9. Yang Q (2019) Highly accurate machine fault diagnosis using deep transfer learning. IEEE Trans Ind Inform 15(4):2446–2455
10. Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX (2019) Deep learning and its applications to machine health monitoring. Mech

Syst Signal Process 115:213–237. https://doi.org/10.1016/j.ymssp.2018.05.050
11. Neupane D, Kim Y, Seok J (2021) Bearing fault detection using scalogram and switchable normalization-based CNN (SN-CNN). IEEE Access 9:88151–88166. https://doi.org/10.1109/ACCESS.2021.3089698
12. Sokolovsky A, Hare D, Mehnen J (2021) Cost-effective vibration analysis through data-backed pipeline optimisation. Sensors 21(19):1–12. https://doi.org/10.3390/s21196678. arXiv:2108.07017
13. Nath AG, Sharma A, Udmale SS, Singh SK (2021) An early classification approach for improving structural rotor fault diagnosis. IEEE Trans Instrum Meas. https://doi.org/10.1109/TIM.2020.3043959
14. Marins MA, Ribeiro FML, Netto SL, da Silva EAB (2018) Improved similarity-based modeling for the classification of rotating-machine failures. J Frankl Inst 355(4):1913–1930. https://doi.org/10.1016/j.jfranklin.2017.07.038
15. Ribeiro F, Marins M, Netto S, Silva E (2017) Rotating machinery fault diagnosis using similarity-based models. In: XXXV Simpósio Bras. Telecomunicações e Process. Sinais-SBrT2017, pp 277–281. https://doi.org/10.14209/sbrt.2017.133
16. Ribeiro FML (2022) MaFaulDa—Machinery Fault Database [Online]. https://www02.smt.ufrj.br/texttildelowoffshore/mfs/page_01.html. Accessed 2 Nov 2022
17. Case Western Reserve University (CWRU) Bearing Fault Dataset. https://engineering.case.edu/bearingdatacenter. Accessed 16 Dec 2022
18. Toh G, Park J (2020) Review of vibration-based structural health monitoring using deep learning. Appl Sci. https://doi.org/10.3390/app10051680
19. Scalabrini Sampaio G, Vallim Filho ARdA, Santos da Silva L, Augusto da Silva L (2019) Prediction of motor failure time using an artificial neural network. Sensors 19(19):4342. https://doi.org/10.3390/s19194342
20. Taufan I (2018) Transfer path analysis Sebagai Fitur Untuk Deteksi Kerusakan Pada Sistem Pompa Sentrifugal-Beam. Technical report, Institut Teknologi Sepuluh Nopember
21. Ihsannur H (2022) Deteksi Kerusakan Pompa Berdasarkan Sinyal Vibrasi Menggunakan Machine Learning. Technical report, Institut Teknologi Sepuluh Nopember
22. Vu-Bac N, Lahmer T, Zhuang X, Nguyen-Thoi T, Rabczuk T (2016) A software framework for probabilistic sensitivity analysis for computationally expensive models. Adv Eng Softw 100:19–31. https://doi.org/10.1016/j.advengsoft.2016.06.005
23. Samaniego E, Anitescu C, Goswami S, Nguyen-Thanh VM, Guo H, Hamdia K, Zhuang X, Rabczuk T (2020) An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications. Comput Methods Appl Mech Eng 362:112790 https://doi.org/10.1016/j.cma.2019.112790. arXiv:1908.10407
24. Ebrahimi E, Javidan M (2017) Vibration-based classification of centrifugal pumps using support vector machine and discrete wavelet transform. J Vibroeng 19(4):2586–2597. https://doi.org/10.21595/jve.2017.18120
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
26. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2. arXiv:2006.10256