Hindawi Publishing Corporation
Shock and Vibration
Volume 2016, Article ID 3843192, 11 pages
http://dx.doi.org/10.1155/2016/3843192

Research Article

Distance and Density Similarity Based Enhanced k-NN Classifier for Improving Fault Diagnosis Performance of Bearings

Sharif Uddin,[1] Md. Rashedul Islam,[2] Sheraz Ali Khan,[1] Jaeyoung Kim,[1] Jong-Myon Kim,[1] Seok-Man Sohn,[3] and Byeong-Keun Choi[4]

[1] *School of Electrical, Electronics and Computer Engineering, University of Ulsan, Ulsan, Republic of Korea*
[2] *Department of Computer Science and Engineering, University of Asia Pacific, Dhaka, Bangladesh*
[3] *Power Generation Laboratory, KEPCO Research Institute, Jeollanam-do, Republic of Korea*
[4] *Department of Energy Mechanical Engineering, Gyeongsang National University, Gyeongsangnam-do, Republic of Korea*

Correspondence should be addressed to Jong-Myon Kim; jongmyon.kim@gmail.com

Received 1 September 2016; Accepted 17 October 2016

Academic Editor: Lu Chen

An enhanced k-nearest neighbor (k-NN) classification algorithm is presented, which uses a density based similarity measure in addition to a distance based similarity measure to improve the diagnostic performance in bearing fault diagnosis. Due to its use of distance based similarity measure alone, the classification accuracy of traditional k-NN deteriorates in case of overlapping samples and outliers and is highly susceptible to the neighborhood size, k. This study addresses these limitations by proposing the use of both distance and density based measures of similarity between training and test samples. The proposed k-NN classifier is used to enhance the diagnostic performance of a bearing fault diagnosis scheme, which classifies different fault conditions based upon hybrid feature vectors extracted from acoustic emission (AE) signals. Experimental results demonstrate that the proposed scheme, which uses the enhanced k-NN classifier, yields better diagnostic performance and is more robust to variations in the neighborhood size, k.

1. Introduction

Rotary machines, in both industry and common households, use bearings to reduce friction and ensure steady and energy efficient operation. Bearings reduce the noise and vibration levels associated with a machine, which is essential for the long term health of both the machine and its operators. Although bearings are very sturdy components and have very long useful lives; nevertheless, material fatigue due to variations in operating load, currents due to electric discharge, thermal stresses due to variations in operating temperature, corrosion, and contaminants in the operating environment can cause them to fail abruptly. A bearing failure can result in the abrupt shutdown of a machine, which leads to tremendous financial losses. Bearings account for more than 50% of failures in induction motors alone [1], which makes their condition monitoring essential to preventing any abrupt failures. Thus, early and reliable detection of bearing defects is very important as these defects lead to bearing failure.

Many data driven techniques have been proposed for diagnosing faults in bearings. These techniques largely use time-frequency analysis of the fault signals for the extraction of meaningful information about underlying faults [2, 3]. Fault signals, such as stator current, vibration acceleration, and acoustic emissions, are inherently nonstationary and hence they are processed in the time-frequency domain, using the short-time Fourier transform (STFT) [4], wavelet transforms [5–10], empirical mode decomposition (EMD) [11–15], and the Hilbert-Huang transform [16–18], to extract characteristic information about different bearing defects. Acoustic emissions are characterized by their low energies and very high bandwidths. They are captured using wideband acoustic sensors and are very effective in diagnosing nascent faults [19–21]. This paper presents a data driven approach for fault diagnosis in bearings, which extracts hybrid features from the acoustic emission (AE) signals and then employs the proposed enhanced k-NN classifier to diagnose different bearing defects.

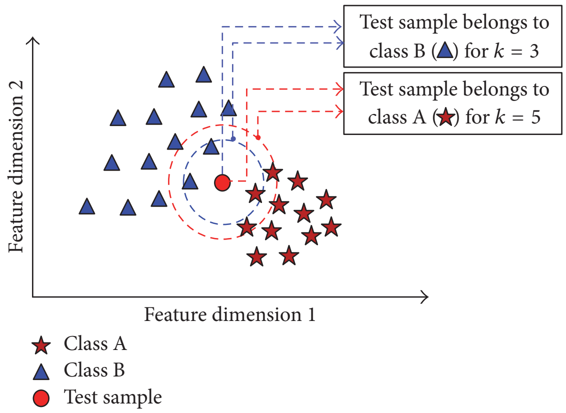

Test sample belongs to class B (▲) for $k = 3$
Test sample belongs to class A (★) for $k = 5$
Feature dimension 2
Feature dimension 1
★ Class A
▲ Class B
● Test sample

Figure 1: Limitation of classical k-NN using only the distance based similarity measure.

The hybrid feature vectors are constructed by calculating different statistical measures of the time and frequency domain AE signal and its envelope power spectrum. This rather extensive set of features is constructed to uniquely identify each fault condition; nevertheless, all features are not of equal utility in classifying a given fault correctly. Moreover, a high dimensional feature vector is bound to make the classification process computationally more expensive. Furthermore, if the feature vector contains too many redundant or irrelevant features, it may also degrade the classifier's accuracy. Hence, the dimensionality of the feature vector is reduced using feature selection methods, which prune the high dimensional feature vector by eliminating the suboptimal features and selecting only those, which would result in the highest classification accuracy. These optimal features are used to create a model of the data by training a classifier, which is then employed to classify the unknown fault signals.

Due to its simplicity and effectiveness, k-NN is usually the first choice in solving any classification problem. However, two factors can degrade its performance. First, k-NN determines the similarity between two samples using only a distance measure of similarity; the widely used distance measures are the Euclidean and Manhattan distance. Second, the classification decision and hence accuracy are sensitive to the neighborhood size, k. These problems have been highlighted in Figure 1, where the classification decision for the unknown test sample (shown as a red circle) changes with change in the neighborhood size. The test sample is labeled as "B" if $k = 3$, whereas it is labeled as "A" if $k = 5$. The limitations of traditional k-NN, due to its use of distance based similarity measure, can be overcome using the local outlier factor (LOF) [22, 23] and local correlation integral (LOCI) [24], which are measures of similarity, based on the density of data samples. Hence, in this study, hybrid similarity measures (i.e., both distance and density based) are proposed to improve the diagnostic performance of the classical k-NN and make it more resilient to the choice of neighborhood size, k.

The main contribution of this study is that an enhanced k-NN classifier is proposed, which uses hybrid measures of similarity between data samples to make it more resilient to the choice of neighborhood size, k, and to increase its diagnostic performance relative to classical k-NN. The density based similarity measure (i.e., LOF) is used to boost the decision of classical k-NN, which classifies an unknown sample based only upon its Euclidean distance from its "k" nearest neighbors using the majority rule. In the proposed k-NN, when the k nearest neighbors of an unknown sample do not belong to the same class, then the LOF is used to decide the class membership of the unknown simple.

The organization of the rest of the paper is as follows. In Section 2, the fault simulator and data acquisition setup are presented. In Section 3, the fault diagnosis scheme and the proposed enhanced k-NN classifier are discussed in detail. In Section 4, a discussion of the achieved results is provided, whereas, in Section 5, conclusions of this work are provided.

2. Fault Simulator and Data Acquisition System

The acoustic emission (AE) signals are acquired using a machinery fault simulator, which is used to simulate different fault conditions. The fault simulator uses cylindrical roller element bearings (FAG NJ206-E-TVP2), which are ingrained with cracks on its different parts. AE signals are collected for bearings at the nondrive end of the simulator using a wide-band acoustic sensor and a PCI-2 based data acquisition system, which samples the AE signals at a rate of 250 KHz [25]. The acoustic sensor is connected to the top of the bearing housing and is at an approximate distance of 21.48 mm from the bearing, as shown in Figure 2. The nondrive end shaft is connected to the drive end through a gearbox with a reduction ratio of 1.52 : 1.

The bearings are seeded with cracks of two different sizes (e.g., 3 mm and 12 mm), and these cracks are introduced on either one or two components of the bearing to study both single and compound bearing defects. The AE signals recorded for bearings with 3 mm cracks and for bearings with 12 mm cracks are grouped into separate datasets. Moreover, for each crack size, the AE signals are recorded at two different shaft speeds (e.g., 300 RPM and 350 RPM). Thus, a total of four datasets are considered, each with AE signals recorded at a different shaft speed along with different crack sizes. The types of single and compound bearing defects are shown in Figure 3; they include cracks on the roller (BFR), inner raceway (BFI), outer raceway (BFO), inner and outer raceways (BFIO), inner raceway and roller (BFIR), outer raceway and roller (BFOR), and both inner and outer raceways and the roller (BFIOR). For each shaft speed, AE signal for a healthy bearing (FFB) is also recorded.

As mentioned earlier, the AE signals are divided into 4 datasets based upon the crack size and shaft speed, as given in Table 1. For every bearing defect, 90 AE signals are recorded; each signal is of 5-second duration. Similarly, 90 AE signals are recorded for the healthy bearing. Thus, every dataset contains a total of 720 AE signals.

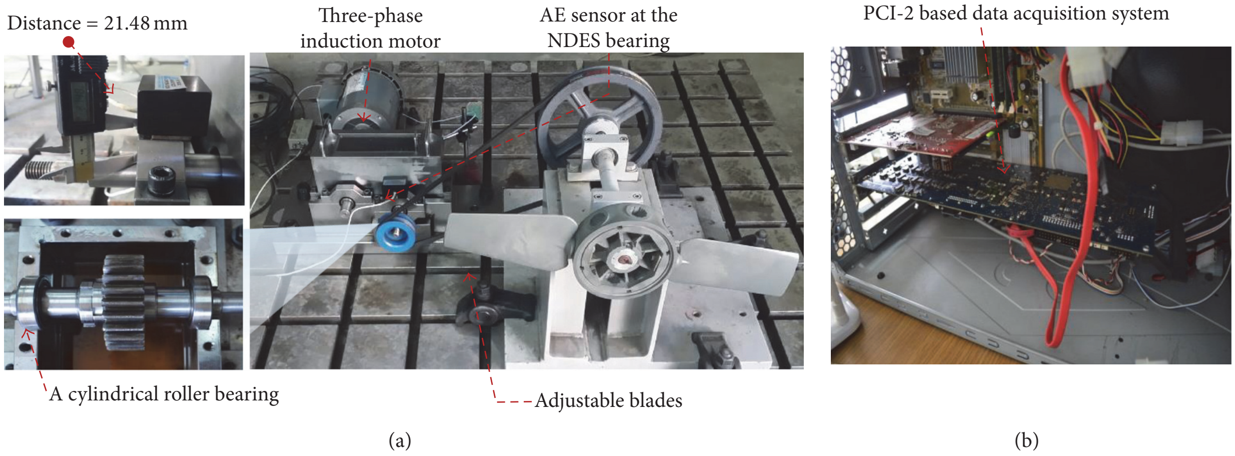

Distance = 21.48 mm
Three-phase induction motor
AE sensor at the NDES bearing
PCI-2 based data acquisition system
A cylindrical roller bearing
Adjustable blades
(a)
(b)

Figure 2: (a) The fault simulator with a three-phase induction motor, a wide-band acoustic sensor, gearbox, and (b) a PCI-2 based system for AE data acquisition.

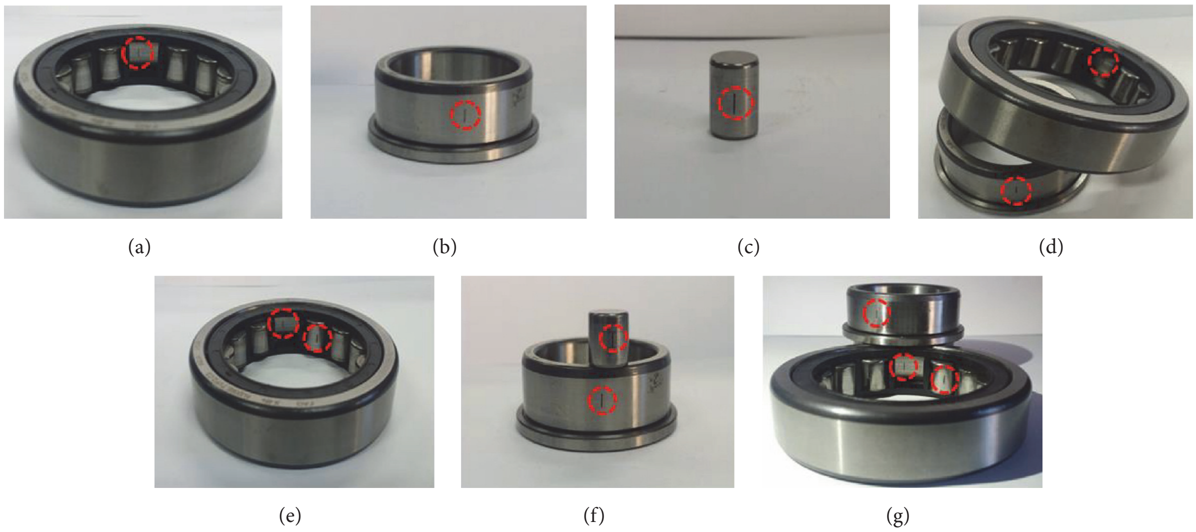


Figure 3: Different single and combined bearing faults: (a) outer raceway fault, (b) inner raceway fault, (c) roller fault, (d) outer and inner raceway faults, (e) outer raceway and roller faults, (f) inner raceway and roller faults, and (g) inner and outer raceway and roller faults.

Table 1: Datasets for the proposed bearing fault diagnosis methodology.

Dataset	Shaft speed	Crack size	Fault types (Number of AE signals)							
Dataset 1	300 rpm	3 mm	BFI (90)	BFO (90)	BFR (90)	BFIO (90)	BFIR (90)	BFOR (90)	BFIOR (90)	FFB (90)
Dataset 2	350 rpm	3 mm	BFI (90)	BFO (90)	BFR (90)	BFIO (90)	BFIR (90)	BFOR (90)	BFIOR (90)	FFB (90)
Dataset 3	300 rpm	12 mm	BFI (90)	BFO (90)	BFR (90)	BFIO (90)	BFIR (90)	BFOR (90)	BFIOR (90)	FFB (90)
Dataset 4	350 rpm	12 mm	BFI (90)	BFO (90)	BFR (90)	BFIO (90)	BFIR (90)	BFOR (90)	BFIOR (90)	FFB (90)

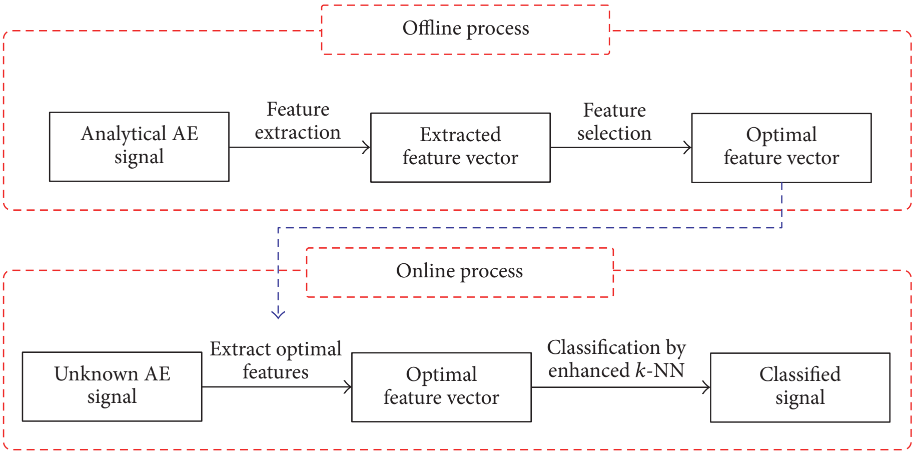


Figure 4: The proposed methodology for bearing fault diagnosis.

3. The Proposed Methodology for Bearing Fault Diagnosis

The proposed methodology for bearing fault diagnosis works in two phases, as illustrated in Figure 4. The first phase comprises an *offline process* that involves feature extraction and feature selection, which are discussed in detail in Sections 3.1 and 3.2, respectively. The offline process is used to determine the set of optimal features that would yield the highest classification accuracy. In the second phase, an *online process* is used to classify the unknown AE signals using the proposed enhanced k-NN classifier. The online process calculates only the optimal set of features for each AE signal and, using only those features, it labels the unknown AE signals.

3.1. Features Extraction. In order to accurately identify each bearing defect, a high dimensional hybrid feature vector is constructed using 22 different features of the AE signal. These features are useful in extracting maximum information about each fault [26] and include ten statistical measures of the time-domain AE signal and three statistical measures of the frequency domain AE signal. These features are listed in Table 2 along with the mathematical relationships for their calculation. Moreover, nine statistical measures, calculated over the envelope power spectrum of the AE signal, are also included in the hybrid feature vector. The features from the envelope power spectrum include the root mean square (RMS) values for each of the three defect frequencies and its first two harmonics. The defect frequencies include the ball pass frequency over inner race (BPFI), the ball pass frequency over the outer race (BPFO), and the ball spin frequency (BSF). The range of values for these defect frequencies and their harmonics is shown in Figure 5.

The range of values for the defect frequencies and their first two harmonics is calculated using (1), (2), and (3), respectively.

$$r_{\text{inner}} = 2 \times \left\{ n_{\text{sidebands}} \times \left(f_s + f_s \times e_{\text{rate}} \right) + e_{\text{rate}} \times f_i \right\}, \quad (1)$$

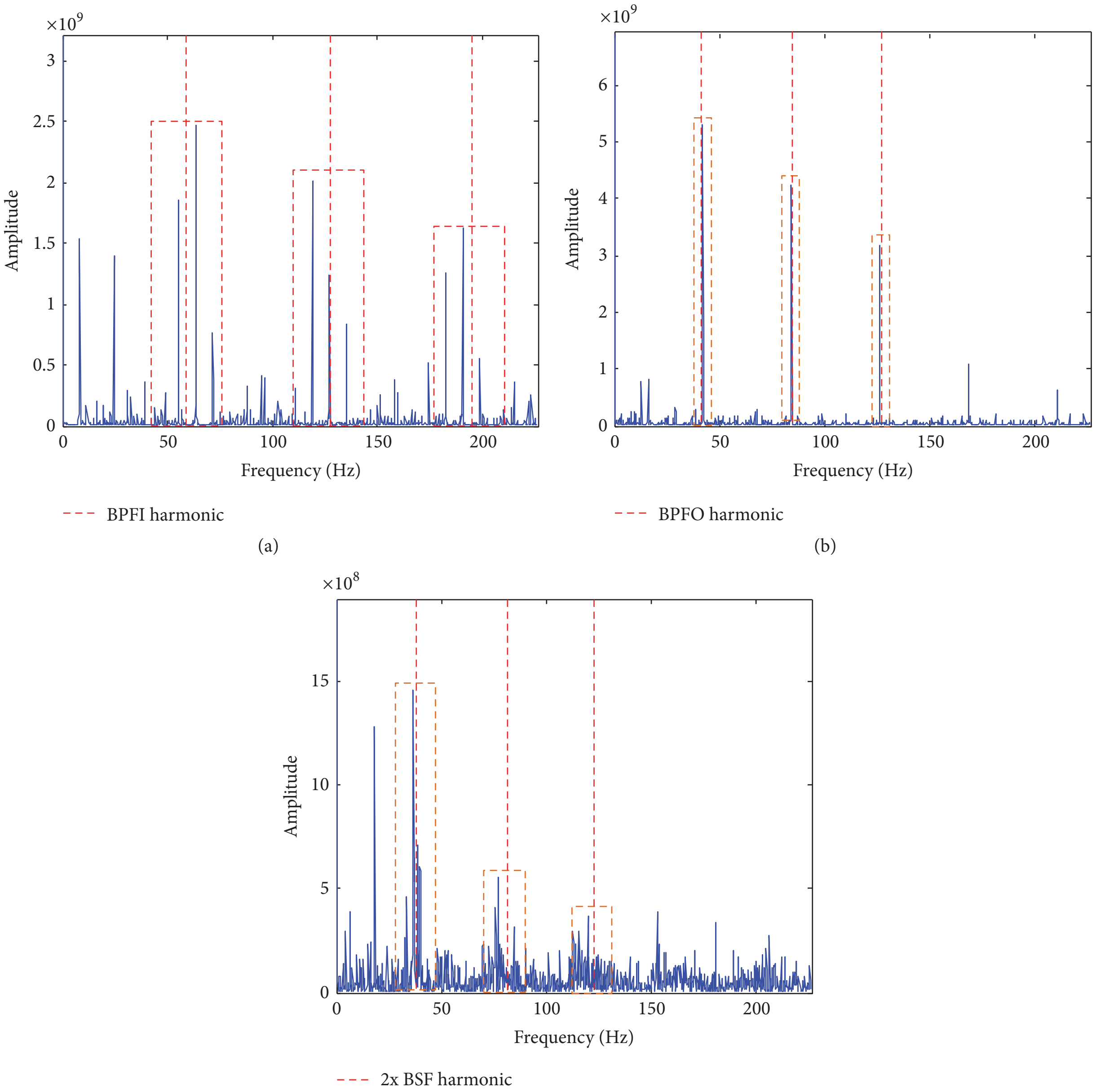
$$r_{\text{outer}} = 2 \times e_{\text{rate}} \times f_o, \quad (2)$$

$$r_{\text{roller}} = 2 \times \left\{ n_{\text{sidebands}} \times \left(f_c + f_c \times e_{\text{rate}} \right) + e_{\text{rate}} \times f_r \right\}, \quad (3)$$

Table 2: Statistical measures calculated over the time and frequency domain AE signal.

Parameter	Definition
Root mean square (RMS)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
Kurtosis value (KV)	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma}\right)^4$
Peak-to-peak value (PPV)	$\text{PPV} = \max(x_i) - \min(x_i)$
Crest factor (CF)	$\frac{\max(\lvert x_i \rvert)}{\sqrt{(1/N)\sum_{i=1}^{N} x_i^2}}$
Shape factor (SF)	$\frac{\sqrt{(1/N)\sum_{i=1}^{N} x_i^2}}{(1/N)\sum_{i=1}^{N} \lvert x_i \rvert}$
Frequency center (FC)	$\frac{1}{N}\sum_{i=1}^{N} f_i$
RMS frequency (RMSF)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} f_i^2}$
Square root of amplitude (SRA)	$\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i \rvert}\right)^2$
Skewness value (SV)	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma}\right)^3$
Impulse factor (IF)	$\frac{\max(\lvert x \rvert)}{(1/N)\sum_{i=1}^{N} \lvert x_i \rvert}$
Margin factor (MF)	$\frac{\max(\lvert x_i \rvert)}{\left((1/N)\sum_{i=1}^{N}\sqrt{\lvert x_i \rvert}\right)^2}$
Kurtosis factor (KF)	$KF = \frac{(1/N)\sum_{i=1}^{N}\left((x_i - \overline{x})/\sigma\right)^4}{\left((1/N)\sum_{i=1}^{N} x_i^2\right)^2}$
Root variance frequency (RVF)	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - f_c)^2}$

where $n_{\text{sidebands}}$ is the number of sidebands, f_s is the operating frequency, e_{rate} is the error rate, f_i is the inner defect



(a)

(b)

(c)

Figure 5: Fault frequency regions up to three harmonics at (a) inner, (b) outer, and (c) roller fault frequency.

frequency, f_o is the outer defect frequency, f_c is the cage frequency, and f_r is the roller defect frequency.

3.2. Feature Selection. Although a high dimensional hybrid feature vector is highly desirable to capture the characteristics of different types of defects, the diagnostic performance of the proposed method can be degraded by potentially irrelevant and redundant features. Moreover, a high dimensional feature vector entails an increased computational cost during feature extraction and classification, which involves the calculation of distances and densities between different samples [25–27]. Hence, the original feature vector is evaluated to determine the set of optimal features that would yield the best diagnostic performance and reduce the computational cost of the proposed method.

In this study, sequential forward selection (SFS) is used for feature selection, which is a simple and fast greedy search algorithm. It starts with an initially empty set, $S = 0$, and then iteratively selects the most significant feature from the original set with respect to the set, S. This is done by first selecting a feature from the original set and then adding it to the set, S, only if the newly selected feature maximizes the value of the objective function for the set, S. The feature is discarded and the process moves to the next feature, if the selected feature decreases the value of the objective function for the set, S. The objective function for SFS is given by (4), which is basically

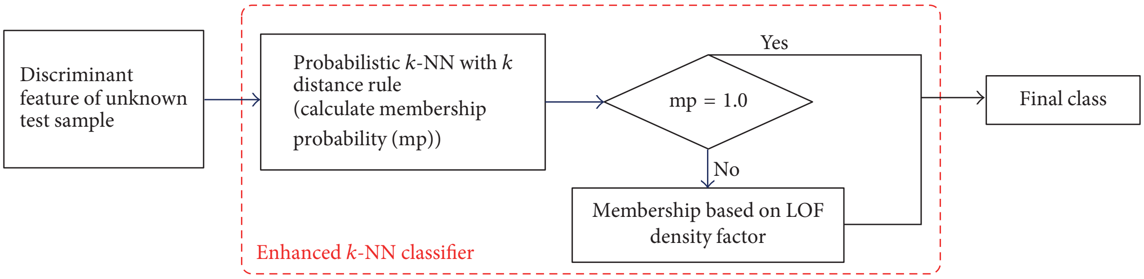

Discriminant feature of unknown test sample
Probabilistic k-NN with k distance rule (calculate membership probability (mp))
mp = 1.0
Yes
No
Final class
Membership based on LOF density factor
Enhanced k-NN classifier

Figure 6: The proposed k-NN classifier.

the ratio of interclass separability to intraclass compactness [25]. The interclass separability is given by the interclass distance $d_{\text{inter\_class}}$, whereas $d_{\text{intra\_class}}$ is the intraclass compactness. Although SFS is simple, efficient, and reasonably accurate, it has its own disadvantages. It suffers from the nesting problem; that is, a feature retained once cannot be discarded, which can result in suboptimal feature selection [28–30].

$$f_{\text{objective}} = \frac{d_{\text{inter\_class}}}{d_{\text{intra\_class}}}. \tag{4}$$

3.3. Enhanced k-NN Classification Algorithm. The traditional k-NN classifier labels an unknown test sample according to the majority of its nearest neighbors in the training set. The nearest neighbors are determined using a distance measure, which is mostly the Euclidean distance between two samples. In multiclass classification problems, where the density of each class is different, the use of a distance based measure of similarity between the test and training samples can result in misclassification and render the classification result sensitive to the choice of neighborhood size, k, as illustrated in Figure 1. This happens because traditional k-NN does not take into account variation in densities across different classes. Therefore, an enhanced k-NN classifier is proposed, which uses both distance and density based similarity measures to improve its classification accuracy. For a given test sample, first its membership probabilities for different classes are calculated. This is done through voting by its k nearest neighbors, which in turn are determined using the Euclidean distance of the test sample from all the training samples. If the membership probability for the test sample is one (i.e., all its nearest neighbors belong to a single class), then the proposed k-NN classifier admits this result and labels the test sample according to its nearest neighbors. However, if the membership probability of the test sample is less than one, (i.e., all the nearest neighbors do not belong to a single class), then the LOF based density measure is used to determine the label of the test sample. The use of LOF in conjunction with Euclidean distance makes the classification performance, of the enhanced k-NN, insensitive to the neighborhood size, k.

As shown in Figure 6, the proposed k-NN first calculates the membership probabilities for the unknown test samples using probabilistic k-NN, which uses Euclidean distance as a measure of similarity. The probabilistic k-NN does not assign any class labels to the test samples; instead it only calculates their membership probabilities for all the classes.

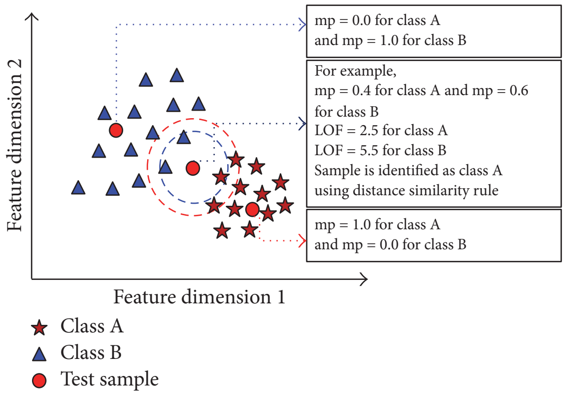


Figure 7: Classifying a test sample using the enhanced k-NN classifier.

If, for each class, the membership probability of a test sample is less than 1.0, then the output of the majority rule is ignored and the final membership of the test sample is determined using the LOF value, as shown in Figure 7.

3.4. Calculating the Local Outlier Factor (LOF). The local outlier factor (LOF) has been used for the detection of outliers or anomalous data points [22], which have relatively lower probabilities of being members of any class. An unknown sample is classified by comparing its density with that of its neighbors. Points with densities like their neighbors are classified accordingly; that is, points with lower densities are labeled according to their neighbors with lower densities, whereas points with higher densities are labeled according their neighbors with higher densities. The LOF can be calculated as follows:

(i) *First*, the calculation of the distance of every data point "q" to its kth nearest neighbor (i.e., d_k^q is calculated), for $k = 3$, is illustrated in Figure 8(a).

(ii) *Second*, for each data point "q", its reachability distance with respect to the data point "p" (i.e., $d_r^{q,p}$ is calculated) is the true distance between points "p"

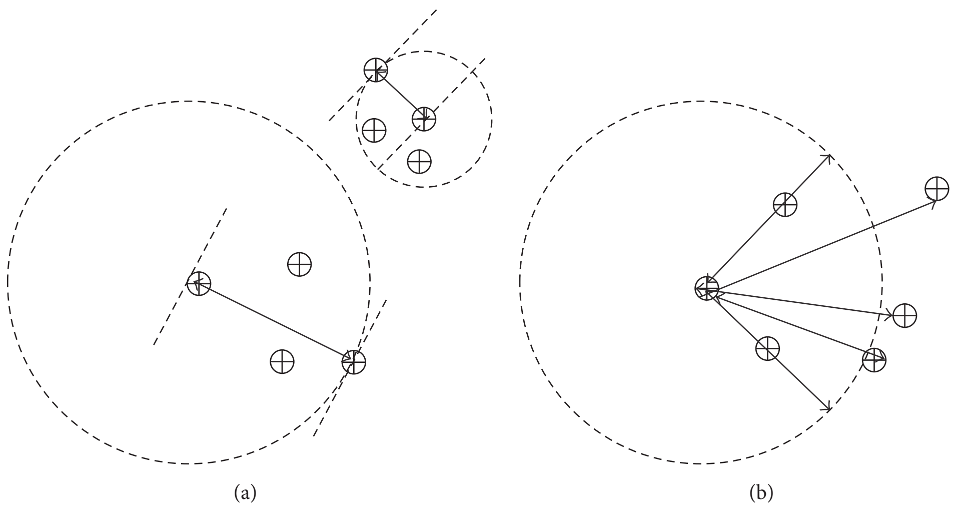


Figure 8: For $k = 3$, calculating (a) the k-distance and (b) the reachability distance.

and "q" with a minimum value of d_k^q, as illustrated in Figure 8(b). It can be calculated as follows:

$$d_r^{q,p} = \max_i \left\langle d_k^q, d_r^{q,p} \right\rangle. \quad (5)$$

(iii) *Third*, for each data point "q", its local reachability density (i.e., σ_r^q is calculated) is defined as the inverse of its average reachability distance from its "M" nearest neighbors, as given in (6). The value of "M" is set to 16, as given in Table 3:

$$\sigma_r^q = \frac{M}{\sum_M d_r^{q,p}}. \quad (6)$$

(iv) *Finally*, for each data point "q", its local outlier factor or LOF value is determined, by comparing its local reachability density to that of its "M" nearest neighbors using the following relation:

$$\text{LOF} = \frac{1}{M} \sum_{p=1}^{M} \frac{\sigma_r^p}{\sigma_r^q}. \quad (7)$$

The LOF values for all the training samples are computed using (7) during the training phase. The unknown test samples are classified based upon the similarity of their LOF values to that of their neighbors.

4. Results and Discussion

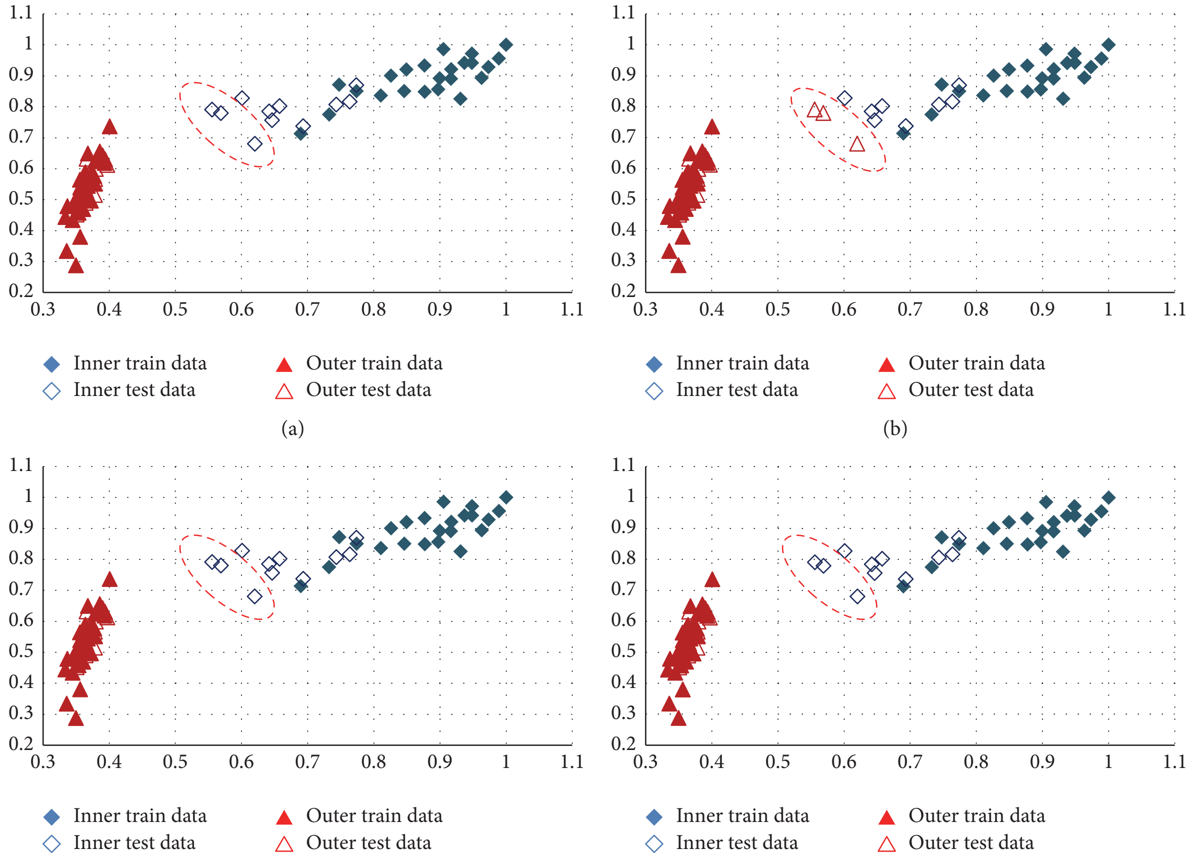
In this section, a discussion of the experimental results achieved by the proposed method for bearing fault diagnosis is provided. As mentioned earlier, four datasets are used to test the proposed method, details of which are given in Table 1. The method uses the enhanced k-NN classifier, which has been proposed to address the limitations of traditional k-NN. The enhanced k-NN classifier was used with the parameters given in Table 3.

Table 3: Values of various parameters for the enhanced k-NN classifier.

Property	Value
Neighborhood size for k-NN	3, 5, 7, and 9
Neighborhood size for local reachability density	16
Neighborhood size for LOF	12
Outlier threshold	$>2\sigma$

To demonstrate the effectiveness of the proposed k-NN classifier, the classification of inner race fault samples from dataset 1 is illustrated in Figure 9, using both the traditional and proposed k-NN classifiers with neighborhood sizes of 3 and 7 (i.e., $k = 3$ and $k = 7$). The samples shown inside the red ellipse are to be classified; their true label is "*inner_race_fault*" (i.e., these samples belong to the inner race fault class). However, the classification result of the traditional k-NN classifier varies with the value of k (i.e., for $k = 3$); it correctly classifies these samples as inner race fault samples, whereas, for $k = 7$, it classifies these as outer race fault samples, which is incorrect. It happens because traditional k-NN uses the majority rule to decide the class label for an unknown test sample. In this particular case, among the nearest three neighbors of these unknown test samples, two are inner race fault and one is outer race fault. Hence, for the case of $k = 3$, they are correctly classified as inner race fault samples. However, among the nearest seven neighbors of these unknown test samples, four are outer race fault and three are inner race fault. Hence, for the case of $k = 7$, they are incorrectly classified as outer race fault samples. In contrast, the proposed k-NN always classifies these samples as inner race fault samples, irrespective of the size of neighborhood (i.e., the value of k).

The proposed k-NN classifier correctly classifies these unknown test samples because it uses the LOF, which is a density based similarity measure. LOF is used only when the nearest neighbors of a given test sample do not belong to



(a)

(b)

(c)

(d)

FIGURE 9: Classification of inner race fault samples from dataset 1 using the traditional k-NN classifier (a) with $k = 3$ and (b) with $k = 7$ and using the proposed k-NN classifier (c) with $k = 3$ and (d) with $k = 7$.

the same class (i.e., the vote is not unanimous). Therefore, the class membership probabilities for the unknown test samples are determined. In this particular case, for $k = 3$, the probability that a given test sample is a member of the inner race fault is 66.7%, and the probability that it belongs to the outer race fault is 33.33%. Since both class membership probabilities are less than one, the proposed k-NN classifier employs the LOF values of the unknown test samples and their neighbors to determine the final class labels. This is demonstrated in Figure 10, which shows the LOF values for the test samples and their nearest neighbors. The LOF values of the test samples for outer race fault class are 5.09, 5.069, and 4.979, whereas, for the inner race fault class, their LOF values are 3.33, 3.399, and 3.192, respectively. If the LOF values of these test samples for both the outer and inner race fault classes are compared to the LOF values of their nearest training samples, it can be observed that the LOF values of the test samples for inner race fault are similar to the LOF values of training samples from the inner race fault class. Hence, it can be argued that these test samples are outliers to the outer race fault class and inliers to or members of the inner race fault class.

Similarly, when $k = 7$, the probability that a given test sample is a member of the inner race fault is 42.86%, and the probability that it belongs to the outer race fault is 57.14%. Here again, the class membership probabilities are less than one, and, thus, the proposed k-NN classifier employs the LOF values of the unknown test samples and their neighbors to determine the final class labels. Using the LOF values of the test samples and their nearest training samples, the test samples are classified as members of the inner race fault class.

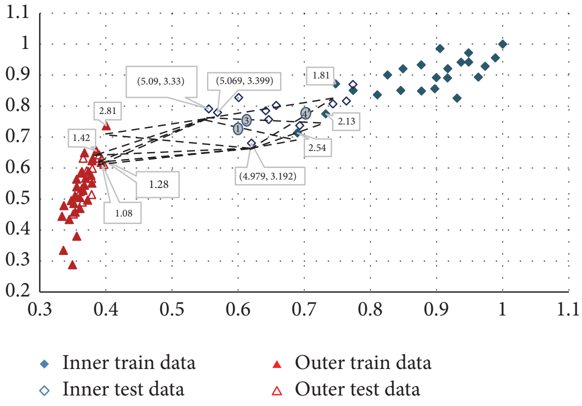


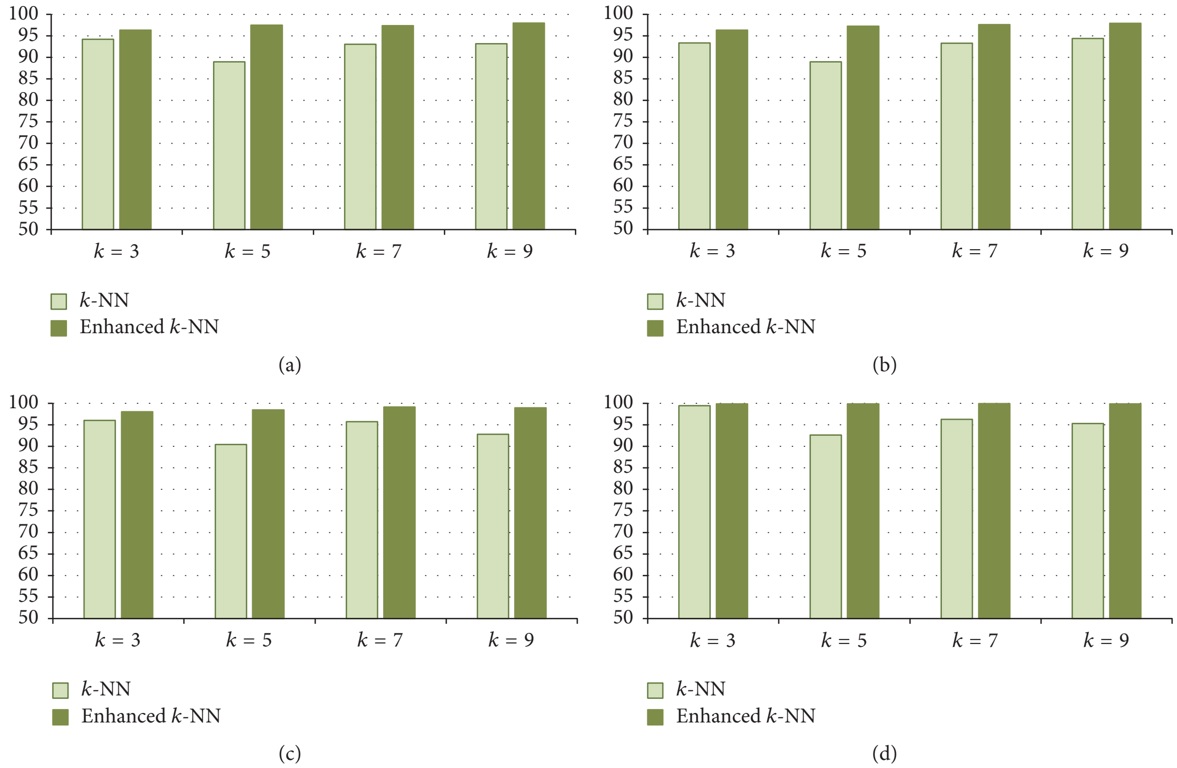
FIGURE 10: The classification of unknown test samples using LOF based density similarity measure.

Table 4: Diagnostic performance of the two classifiers for different fault types and datasets.

k = ?	Model	BFI	BFO	BFR	BFIO	BFOR	BFIR	BFIOR	FFB	Avg. classification accuracy (%)
					Dataset 1: 300 rpm, 3 mm crack					
k = 3	k-NN	94.65	79.50	96.42	98.53	95.67	97.08	91.67	100.00	94.19
	Enhanced k-NN	95.70	87.24	98.75	99.58	96.25	99.92	93.08	100.00	96.32
k = 5	k-NN	72.52	75.61	92.35	91.35	89.25	96.08	94.53	100.00	88.96
	Enhanced k-NN	96.25	92.35	99.45	98.25	99.50	97.92	96.25	100.00	97.50
k = 7	k-NN	98.76	78.50	97.75	96.25	90.75	94.92	87.42	99.83	93.02
	Enhanced k-NN	99.05	94.56	98.65	98.96	98.42	97.56	91.56	99.89	97.33
k = 9	k-NN	89.76	80.56	91.53	97.45	95.26	98.12	92.54	100.00	93.15
	Enhanced k-NN	98.68	94.56	98.16	98.56	99.92	99.92	93.89	100.00	97.96
					Dataset 2: 350 rpm, 3 mm crack					
k = 3	k-NN	93.56	78.52	94.25	97.86	94.58	97.12	92.35	98.68	93.37
	Enhanced k-NN	95.59	88.45	99.02	98.95	96.12	99.43	93.56	99.26	96.30
k = 5	k-NN	73.48	76.53	90.89	90.75	87.62	96.89	95.42	100.00	88.95
	Enhanced k-NN	96.45	93.25	98.45	98.75	98.69	96.25	95.86	100.00	97.21
k = 7	k-NN	98.48	81.52	96.53	95.48	91.48	93.28	89.45	99.89	93.26
	Enhanced k-NN	99.75	95.48	98.15	98.75	97.98	98.06	92.56	100.00	97.59
k = 9	k-NN	94.58	84.83	94.25	96.53	94.56	97.85	92.56	100.00	94.40
	Enhanced k-NN	98.65	95.69	98.75	98.45	99.05	98.45	94.12	100.00	97.90
					Dataset 3: 300 rpm, 12 mm crack					
k = 3	k-NN	97.05	86.52	98.45	98.53	96.98	97.08	93.58	100.00	96.02
	Enhanced k-NN	97.45	93.56	100.00	99.58	98.46	99.92	95.26	100.00	98.03
k = 5	k-NN	78.96	81.45	91.75	92.86	88.46	95.63	94.53	100.00	90.46
	Enhanced k-NN	98.96	95.86	99.26	99.43	99.50	98.76	96.25	100.00	98.50
k = 7	k-NN	98.76	86.46	97.75	96.25	95.36	96.46	94.85	99.83	95.72
	Enhanced k-NN	100.00	98.12	100.00	99.79	99.86	98.75	96.53	100.00	99.13
k = 9	k-NN	89.76	81.65	91.47	96.85	94.62	97.45	90.67	100.00	92.81
	Enhanced k-NN	100.00	98.45	100.00	100.00	99.45	99.92	93.89	100.00	98.96
					Dataset 4: 350 rpm, 12 mm crack					
k = 3	k-NN	99.16	99.56	99.48	100.00	100.00	98.54	98.65	100.00	99.42
	Enhanced k-NN	99.86	100.00	100.00	100.00	100.00	99.75	99.46	100.00	99.88
k = 5	k-NN	89.45	80.54	93.25	92.45	90.74	95.86	98.56	100.00	92.61
	Enhanced k-NN	100.00	100.00	99.86	100.00	99.94	99.75	99.46	100.00	99.88
k = 7	k-NN	98.76	89.45	97.75	96.25	94.52	97.63	96.25	99.83	96.31
	Enhanced k-NN	99.80	100.00	100.00	100.00	100.00	99.75	99.70	100.00	99.91
k = 9	k-NN	97.84	84.25	96.57	94.56	95.26	98.12	95.68	100.00	95.29
	Enhanced k-NN	100.00	100.00	99.86	100.00	99.94	99.75	99.46	100.00	99.88

Likewise, for other datasets and fault types, this is how the proposed k-NN classifier improves the classification accuracy of traditional k-NN. It is clearly evident in Figure 11, which compares the performance of these two classifiers in terms of average classification accuracy, and Table 4, which lists the classification accuracies for each dataset and individual fault type. Moreover, it can also be observed that the accuracy of the proposed k-NN is not affected by the neighborhood size, k, whereas the accuracy of traditional k-NN varies with variations in the neighborhood size, k. It achieves a maximum accuracy for $k = 3$.

The size of the optimal neighborhood, which maximizes the classification accuracy of traditional k-NN, has to be determined on a case to case basis. There are no general rules that work equally well in all situations and for all classes, which can be challenging as it makes the whole process computationally expensive and inflexible. The robustness of the proposed k-NN to variations in the neighborhood size, k, makes it more flexible and efficient to use. It delivers better and steadier performance. Moreover, in multiclass problems like the one considered in this study, where the densities of different classes vary, traditional k-NN performs poorly as it



(a)

(b)

(c)

(d)

Figure 11: Performance comparison of traditional k-NN and the proposed enhanced k-NN in terms of average classification accuracy: (a) dataset 1, (b) dataset 2, (c) dataset 3, and (d) dataset 4.

does not consider variations in density. The proposed k-NN takes into account variations in density of different classes and uses the LOF to decide the class membership of test samples in such cases.

5. Conclusion

In this paper, an enhanced k-nearest neighbor (k-NN) classification algorithm was presented, which employs both density and distance based similarity measures to improve the diagnostic performance in bearing fault diagnosis. The density based similarity measure, LOF, was used to boost the classification performance of traditional k-NN, which deteriorates in case of overlapping samples, outliers, and multiple classes that show different feature distributions. Moreover, the distance based similarity measure makes the classification performance of traditional k-NN highly susceptible to the neighborhood size, k. These limitations were addressed through the use of both distance and density based similarity metrics, between the training and test samples. Using the enhanced k-NN classifier, the diagnostic performance of the proposed bearing fault diagnosis scheme was significantly improved, and the results were more robust to variations in the neighborhood size, k.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the development of a basic fusion technology in electric power industry (Ministry of Trade, Industry & Energy, 201301010170D); funded in part by The Leading Human Resource Training Program of Regional Neo Industry through the National Research Foundation of Korea (NRF), Ministry of Science, ICT, and Future Planning (NRF-2016H1D5A1910564); funded in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry, & Energy (MOTIE) of the Republic of Korea (no. 20162220100050); funded in part by the Business for Startup R&D funded by the Korea Small and Medium Business Administration in 2016 (Grants S2381631 and C0395147); and funded in part by the "Leaders INdustry-university Cooperation" Project supported by the Ministry of Education (MOE).

References

[1] E. Cabal-Yepez, M. Valtierra-Rodriguez, R. J. Romero-Troncoso et al., "FPGA-based entropy neural processor for online detection of multiple combined faults on induction motors," *Mechanical Systems and Signal Processing*, vol. 30, pp. 123–130, 2012.

[2] H. Berriri, M. W. Naouar, and I. Slama-Belkhodja, "Easy and fast sensor fault detection and isolation algorithm for electrical drives," *IEEE Transactions on Power Electronics*, vol. 27, no. 2, pp. 490–499, 2012.

[3] E. Cabal-Yepez, A. G. Garcia-Ramirez, R. J. Romero-Troncoso, A. Garcia-Perez, and R. A. Osornio-Rios, "Reconfigurable monitoring system for time-frequency analysis on industrial equipment through STFT and DWT," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 2, pp. 760–771, 2013.

[4] S. Nandi, T. C. Ilamparithi, S. B. Lee, and D. Hyun, "Detection of eccentricity faults in induction machines based on nameplate parameters," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 5, pp. 1673–1683, 2011.

[5] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: a review with applications," *Signal Processing*, vol. 96, pp. 1–15, 2014.

[6] J. Seshadrinath, B. Singh, and B. K. Panigrahi, "Investigation of vibration signatures for multiple fault diagnosis in variable frequency drives using complex wavelets," *IEEE Transactions on Power Electronics*, vol. 29, no. 2, pp. 936–945, 2014.

[7] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Fault diagnosis of rolling element bearing using cyclic autocorrelation and wavelet transform," *Neurocomputing*, vol. 110, pp. 9–17, 2013.

[8] P. Konar and P. Chattopadhyay, "Bearing fault detection of induction motor using wavelet and Support Vector Machines (SVMs)," *Applied Soft Computing Journal*, vol. 11, no. 6, pp. 4203–4211, 2011.

[9] J. Rafiee, M. A. Rafiee, and P. W. Tse, "Application of mother wavelet functions for automatic gear and bearing fault diagnosis," *Expert Systems with Applications*, vol. 37, no. 6, pp. 4568–4579, 2010.

[10] P. H. Nguyen and J.-M. Kim, "Multifault diagnosis of rolling element bearings using a wavelet kurtogram and vector median-based feature analysis," *Shock and Vibration*, vol. 2015, Article ID 320508, 14 pages, 2015.

[11] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 108–126, 2013.

[12] J. Zheng, J. Cheng, and Y. Yang, "Generalized empirical mode decomposition and its applications to rolling element bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 40, no. 1, pp. 136–153, 2013.

[13] X. Zhang and J. Zhou, "Multi-fault diagnosis for rolling element bearings based on ensemble empirical mode decomposition and optimized support vector machines," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 127–140, 2013.

[14] G. F. Bin, J. J. Gao, X. J. Li, and B. S. Dhillon, "Early fault diagnosis of rotating machinery based on wavelet packets—empirical mode decomposition feature extraction and neural network," *Mechanical Systems and Signal Processing*, vol. 27, no. 1, pp. 696–711, 2012.

[15] M. Amarnath and I. R. Krishna, "Empirical mode decomposition of acoustic signals for diagnosis of faults in gears and rolling element bearings," *IET Science, Measurement and Technology*, vol. 6, no. 4, pp. 279–287, 2012.

[16] J. Yan and L. Lu, "Improved Hilbert-Huang transform based weak signal detection methodology and its application on incipient fault diagnosis and ECG signal analysis," *Signal Processing*, vol. 98, pp. 74–87, 2014.

[17] G. Cheng, Y.-L. Cheng, L.-H. Shen, J.-B. Qiu, and S. Zhang, "Gear fault identification based on Hilbert-Huang transform and SOM neural network," *Measurement*, vol. 46, no. 3, pp. 1137–1146, 2013.

[18] S. Osman and W. Wang, "An enhanced Hilbert-Huang transform technique for bearing condition monitoring," *Measurement Science and Technology*, vol. 24, no. 8, Article ID 085004, 2013.

[19] A. Widodo, B.-S. Yang, E. Y. Kim, A. C. C. Tan, and J. Mathew, "Fault diagnosis of low speed bearing based on acoustic emission signal and multi-class relevance vector machine," *Nondestructive Testing and Evaluation*, vol. 24, no. 4, pp. 313–328, 2009.

[20] D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, "Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4137–4145, 2013.

[21] S. A. Niknam, V. Songmene, and Y. H. J. Au, "The use of acoustic emission information to distinguish between dry and lubricated rolling element bearings in low-speed rotating machines," *International Journal of Advanced Manufacturing Technology*, vol. 69, no. 9–12, pp. 2679–2689, 2013.

[22] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, 2000.

[23] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 190–237, 2014.

[24] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: fast outlier detection using the local correlation integral," in *Proceedings of the 19th International Conference on Data Engineering*, pp. 315–326, March 2003.

[25] M. Kang, J. Kim, B.-K. Choi, and J.-M. Kim, "Envelope analysis with a genetic algorithm-based adaptive filter bank for bearing fault detection," *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. EL65–EL70, 2015.

[26] R. Islam, S. A. Khan, and J.-M. Kim, "Discriminant feature distribution analysis-based hybrid feature selection for online bearing fault diagnosis in induction motors," *Journal of Sensors*, vol. 2016, Article ID 7145715, 16 pages, 2016.

[27] I.-K. Jeong, M. Kang, J. Kim, J.-M. Kim, J.-M. Ha, and B.-K. Choi, "Enhanced DET-based fault signature analysis for reliable diagnosis of single and multiple-combined bearing defects," *Shock and Vibration*, vol. 2015, Article ID 814650, 10 pages, 2015.

[28] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25–41, 2000.

[29] D. P. Muni, N. R. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 1, pp. 106–117, 2006.

[30] M. Last, A. Kandel, and O. Maimon, "Information-theoretic algorithm for feature selection," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 799–811, 2001.