Cross-domain bearing fault diagnosis with refined composite multiscale fuzzy entropy and the self organizing fuzzy classifier

Esther W. Gituku1 | James K. Kimotho2 | Jackson G. Njiri1

1Department of Mechatronic Engineering, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
2Department of Mechanical Engineering, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Abstract
In this article, the use of refined composite multiscale fuzzy entropy (RCMFE) for cross-domain diagnosis of bearings is introduced and verified with two publicly available datasets of varying operating conditions, a factor that challenges the diagnostic ability of trained models. For classification, the self organizing fuzzy (SOF) classifier is used. The diagnostic framework which primarily only involves extracting RCMFE feature and training the SOF classifier, is able to detect and isolate faults with over 97% accuracy when the classes are comprised of a single fault type and size. Compared to related works, the proposed approach does not require deep learning for feature extraction nor any domain adaptation technique as the RCMFE feature is robust against changing operating conditions. Furthermore, the method does not need target domain data during training. With regard to fault isolation, when the classes in the training data contain all the available fault sizes instead of a single size, the classifier can distinguish inner race faults from outer race and ball faults with an average accuracy of 96%. However, the accuracy for differentiating ball and outer race faults falls slightly to an average of 86%. Thus even for the latter arrangement which poses a tougher transfer learning problem, the proposed approach still performs very well.

KEYWORDS
bearings, cross domain diagnosis, fuzzy entropy, RCMFE, self organizing classifier

1 INTRODUCTION
Bearing faults account for between 30% and 40% of all faults encountered in rotating machinery such as electro-mechanical drives, wind turbines, industrial machinery, and so on.1,2 Thus, because early fault detection is desirable from a maintenance point of view in terms of eliminating downtime and increasing reliability, a lot of research has gone into bearing diagnosis. Although the diagnostic approaches are either model based, data based, or a hybrid of both, data-driven models are becoming more prevalent due to the ubiquity of sensors in many machines. The result of years of research directed...
to the use of sensor data in condition monitoring has been the successful development of different types of diagnostic models using a variety of features extracted from condition monitoring data.\textsuperscript{1,3-5}

Unfortunately, most data-driven classification algorithms are based on the assumption that the training and testing sets come from the same domain where domain here connotes data and its associated distribution. This assumption has been invalidated in several applications such as computer vision and natural language processing which have a long-standing use of data for classification and object recognition.\textsuperscript{6} Specifically in bearing diagnostics, the change in distribution may be brought about by the method of fault inducement, for example, artificial fault versus real fault, change in operating conditions, or change on fault size and severity. Changes in operating conditions are easily understood from a practical perspective of the operations in a machine shop. As to the distribution change brought about by training on artificial data, models are usually trained on artificial data as sufficient real-damage data (ie, damage attained during actual operation of machinery) is rare. One reason for this scarcity is safety guidelines in applications such energy and electric transport components are replaced after a determined number of operation cycles and not due to failure.\textsuperscript{7} Artificial fault data is then left as the best recourse when developing fault detection algorithms. The artificial faults are seeded by drills, engravers, and electro-discharge machines. Unfortunately, as noted by Lessmeier et al,\textsuperscript{8} actual bearing damage is complex because development of fatigue damage or damage caused by solid particles is randomly influenced. Thus, true fault evolution remains largely unknown and production of artificial faults even at different severity may not wholly reflect real fault damage. It is then feasible that the marginal distribution of artificial-fault data may be different from real-fault data. Indeed, the same authors\textsuperscript{8} show that a classifier trained with artificial fault data does not generalize well to real fault. With regard to fault size or severity, it is highly probable that models trained with data of a particular size of fault may not recognize data with differently sized faults while the latter is more likely to be the case for incoming test samples in a practical setting.

The problem of transferring the knowledge gained from models trained in one domain to similar tasks in another domain is known as the domain shift problem.\textsuperscript{9} The branch of transfer learning that deals with transfer of knowledge from the source (training) domain to the target (testing) domain is known as domain adaptation (DA).\textsuperscript{10}

DA approaches can be roughly categorized into either feature-based methods or sample-based methods.\textsuperscript{11} The former attempts to reduce distribution divergence by learning a new joint feature representation that is typically hidden and may be of reduced dimensions (subspace).\textsuperscript{6,12} The latter methods use techniques that utilize source samples that are closer to the target distribution. This may involve re-weighting the source samples to give more importance to those that bridge the distribution gap (landmarks) or projecting the rest of the source samples to the landmarks and then learning a classifier on that subspace.\textsuperscript{13,14} DA algorithms usually require complex objective functions combined from several loss functions and consequently, computationally demanding iteration is required to search for the best network architectures, for example, the number of layers and units in neural networks, the ideal weights, and the best hyper-parameters.

Regardless, several published works have reported excellent results in cross-domain fault diagnosis using DA techniques.\textsuperscript{9,15-19}

On the other hand, entropy-based features that are nonlinear parameter estimation methods have been successfully employed in bearing fault diagnosis. This is because condition monitoring data such as vibration signals present nonlinear characteristics due to inherent nonlinearities such as friction, tolerance, and stiffness.\textsuperscript{20,21} As applied in bearing fault diagnosis, entropy based features indicate the degree of randomness or complexity of a signal whereby the complexity noted here is in terms of meaningful structural richness.\textsuperscript{22} A faulty bearing’s signal is punctuated by periodic patterns in the signal length every time the fault is contacted. Thus, there is more self similarity in a faulty signal (ie, less structural complexity) than in a normal bearing’s signal which means that theoretically, the entropy of the latter signal should be higher. However, despite evidence that entropy based features especially in a multiscale approach are unaffected by operating conditions,\textsuperscript{23} there has not been much attempt to utilize entropy features for cross-domain diagnosis. Indeed many of the published works utilizing entropy based features for diagnosis assume the same distribution for the training and testing samples.\textsuperscript{20,21,23-27}

In this work, we extend the use of entropy features for cross domain diagnosis of bearing faults. We choose the refined composite multiscale fuzzy entropy (RCMFE) as the non linear feature to use due to its advantages over single scale fuzzy entropy as well as sample entropy.\textsuperscript{22,28} For classification, we use the recently proposed self organizing fuzzy (SOF) classifier whose antecedent rules check for the similarity of a test sample to class prototypes instead of using membership functions.\textsuperscript{29} The prototypes of a class represent its distribution’s multimodal nature and are learnt in a self evolving manner that is able to handle drifts and shifts in the data pattern.\textsuperscript{30} The effectiveness of RCMFE for cross domain fault diagnosis is proven using two publicly available datasets by first detecting fault then isolating the type of fault. We show that this diagnostic method is a viable tool for cross domain classification as it only requires extraction of RCMFE
features and training the SOF classifier where only the distance metric for determining similarity requires selection. Further, it holds that for any particular machine, this diagnostic model which uses training data from any one single operating condition, will generalize well to any other operating condition.

The remainder of this article is organized as follows. In Section 2, a theoretical review of RCMFE and the SOF classifier is given. In Section 3, the behavior of RCMFE for the two datasets is explored while in Section 4, the methodology to be followed is outlined. In Section 5, the results are presented and discussed while in Section 6, concluding remarks and future direction is given.

2  | THEORETICAL REVIEW

2.1  | Refined composite multiscale fuzzy entropy

Given a one-dimensional time series \( x \) with \( N \) data points \( x = \{ x_1, x_2, x_3, \ldots, x_N \} \), template vectors \( u^m_i \) whose lengths are equal to the embedding dimension \( m \) are created by

\[
u^m_i = \{ u_i, u_{i+1}, u_{i+2}, \ldots, u_{i+m-1} \} - u^0_i, \quad \text{for } i = 1, 2, 3, \ldots, N,
\]

where \( u^0_i \) is the average of the \( m \) elements in vector \( u^m_i \).

The distance between any two vectors \( u^m_i \) and \( u^m_j \) is computed using Chebyshev’s distance.

\[
d^m_{ij} = \max \{ |u^m_{i+k} - u^m_{j+k}|; 0 \leq k \leq m-1, i \neq j \}.
\]

(2)

The degree of similarity \( D^m_{ij} \) of \( u^m_i \) and \( u^m_j \) is then computed using a fuzzy function \( \mu(d^m_{ij}, n, r) \).

\[
D^m_{ij} = \mu(d^m_{ij}, n, r),
\]

(3)

where \( \mu(d^m_{ij}, n, r) \) is the exponential function with parameters \( n \) and \( r \) determining its shape/boundary.

\[
\mu(d^m_{ij}, n, r) = \exp\left(-\frac{(d^m_{ij})^n}{r}\right).
\]

(4)

A function \( \phi^m \) is then defined as

\[
\phi^m = \frac{1}{N-m} \sum_{i=1}^{N-m} \left( \frac{1}{N-m-1} \sum_{j=1, j \neq i}^{N-m} D^m_{ij} \right).
\]

(5)

The entire procedure is repeated for template vectors of length \( s = m + 1 \) such that \( \phi^s \) is similarly given by

\[
\phi^s = \frac{1}{N-s} \sum_{i=1}^{N-s} \left( \frac{1}{N-s-1} \sum_{j=1, j \neq i}^{N-s} D^s_{ij} \right).
\]

(6)

The fuzzy entropy of the series is finally obtained as the negative logarithm of the ratio of \( \phi^s \) to \( \phi^m \).

\[
\text{FuzzyEn}(x, m, n, r) = -\ln\left( \frac{\phi^s}{\phi^m} \right).
\]

(7)

The selection criteria of \( m \), \( n \), and \( r \) are given in References 23 and 28 and in this work \( m = 2 \), \( n = 2 \), and \( r = 0.15 \times \) standard deviation of the time series in accordance with the prevalent guidelines.

Machines are composed of many interacting components such as shafts, gears and bearings. Consequently, condition monitoring data are bound to contain several oscillatory modes due to the interaction and coupling between the components suggesting that analysis at a single scale is insufficient for exploring the data.26 Thus, in machine fault detection, multiscale entropy features are usually computed.20,21,23-27
As first introduced by Costa et al., a coarse grained time series $y_\tau$ is obtained for any scale $\tau$ from the original time series by calculating the arithmetic mean of $\tau$ neighboring values without overlapping.

$$y_\tau(j) = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x(i); \quad 1 \leq j \leq \frac{N}{\tau}. \quad (8)$$

This process is depicted in Figure 1. The fuzzy entropy at a scale $\tau$ is FuzzyEn$(y_\tau, m, n, r)$.

There are two main drawbacks of the aforementioned method of coarse graining. The first is that entropy thus computed lacks of symmetry in its dependence on the original series. For instance, although the measure should behave similarly for $x_3$ and $x_4$ compared to $x_1$ and $x_2$ (see Figure 1), for scale 3, $x_1$, $x_2$, and $x_3$ are separated from $x_4$, $x_5$, and $x_6$ thus breaking the symmetry. Second, because for an $N$ point time series the length of the coarse grained time series at scale factor $\tau$ is equal to $N/\tau$, this length may not be sufficient for accurate calculation of fuzzy entropy when $\tau$ is large. In addition, there may be no matching templates meaning Equation (7) is undefined. Inaccurate and undefined fuzzy entropy compromises reliability of the method. Improved methods such as the composite and refined composite coarse graining have subsequently been put forward.

Refined composite coarse graining for calculation of multiscale sample entropy was introduced in Reference 32 and extended for fuzzy entropy in Reference 22. The composite coarse graining procedure is depicted in Figure 2 where it can be seen that coarse graining at any scale results in more than one new time series. For instance at scale 2, two coarse grained series are obtained while at scale 3, three coarse grained series are created.

To compute the fuzzy entropy at a scale $\tau$, the functions $\phi_m^\tau$ and $\phi_s^\tau$ are computed for all the coarse grained time series in that scale, that is, at scale $\tau = 2$: $\phi_m^{\tau,1}$, $\phi_m^{\tau,2}$, $\phi_s^{\tau,1}$, and $\phi_s^{\tau,2}$ are computed and their averages $\bar{\phi}_m^{\tau}$ and $\bar{\phi}_s^{\tau}$ also calculated. The fuzzy entropy computed at scale $\tau$ is now referred to as the RCMFE

$$\text{RCMFE}(x, \tau, m, n, r) = -\ln \left( \frac{\bar{\phi}_s^{\tau}}{\bar{\phi}_m^{\tau}} \right). \quad (9)$$

Equation (9) is always defined unless all $\phi_m^{\tau,j}$ and all $\phi_s^{\tau,j}$ are zeros. The length of the original data series $N$ is chosen as 4000 which is sufficient for RCMFE calculation.

2.2 | Self organizing fuzzy classifier

The SOF classifier based on the AnYa antecedent was used for cross-domain classification. This antecedent differs from the traditional Zadeh-Mamdani and Tagaki-Sugeno ones in that it is not constructed from membership functions.
Instead, a sample is checked for its similarity with the prototypes $p_i^c$ that represent the multimodal nature of the training data in each class $c$.

For each class $c$ with a total of $N_c$ prototypes, $N_c$ fuzzy rules corresponding to each are formulated as

$$IF \ ((x \sim p_1^c) \ OR \ (x \sim p_2^c) \ OR \ ... \ OR \ (x \sim p_{N_c}^c)) \ THEN \ (class \ c)$$

(10)

$\sim$ is the similarity operator and is analogous to the degree of membership in alternative antecedents.

To classify a sample $x$ as belonging to one of $C$ classes, the firing strength of the $c$th class ($c = 1, 2, 3, ... , C$) is computed as

$$\lambda_c(x) = \max_{p \in \{p_1^c,...,p_{N_c}^c\}} (e^{-d(x,p)})$$

(11)

where $d(x,p)$ is the distance or similarity between $x$ and $p$ and may be any one of the common distance functions such as Euclidean, Cosine, Mahalanobis, and so on.

The final label is assigned using a “winner takes all” strategy according to

$$label \ x = \ \arg\max_{c=1,2,3,...,C} (\lambda_c(x))$$

(12)

The following section outlines how the training data are divided into the most representative local modes, each characterized by a prototype.

### 2.3 Empirical data characterization

In their work, Angelov and Gu propose a data-driven method of characterizing empirical data without imposing any restrictive assumptions, for example, presuming the data are Gaussian. Since distribution of real data is often complex and multimodal, the authors define several measures to uncover these local ensembles purely from data.

1. Cumulative proximity as the sum of square distances of a sample from all other samples. This metric provides information about the centrality of a sample.
\[ q(x_i) = \sum_{j=1}^{K} d^2(x_i, x_j); \quad j = 1, 2, 3, \ldots, K, \]  

where \( x_i, x_j \in \{x\}_K \) and \( d(x_i, x_j) \) is a function computing the distance between \( x_i \) and \( x_j \). The shorter this distance metric is, the more similar two samples are. \( x \) can be of any dimension, that is, a sample can have any number of features. \( K \) is the number of samples in the dataset.

2. Eccentricity as the normalized cumulative proximity which is useful in capturing the properties of samples at the tail ends of a distribution, that is, samples far away from the peak.

\[ \xi(x_i) = \frac{2q(x_i)}{\sum_{j=1}^{K} q(x_j)}; \quad j = 1, 2, 3, \ldots, K. \]  

The \( \frac{1}{K} \) in the denominator is used to standardize eccentricity to prevent tendency to zero as \( K \) grows large.

3. Unimodal data density as a measure describing mutual proximity of samples and is thus the inverse of eccentricity.

\[ D(x_i) = \frac{1}{\xi(x_i)} = \frac{\sum_{j=1}^{K} q(x_j)}{2Kq(x_i)} = \frac{\sum_{j=1}^{K} \sum_{j=1}^{K} d^2(x_i, x_j)}{2K \sum_{j=1}^{K} d^2(x_i, x_j)}; \quad j = 1, 2, 3, \ldots, K. \]  

From the denominator, it is seen that this measure is inversely proportional to the sum of distances between a sample and all others. Obviously, the closer a sample is to the global mean, the higher its data density, that is, more samples surround it.

4. Multimodal density. In the data, a sample may be observed repeatedly and in this case the set of unique samples is denoted by \( \{U\}_T = u_1, u_2, u_3, \ldots, u_T \) where each distinct value is only recorded once and \( T \) is the number of unique samples. The corresponding frequencies of the unique samples in the dataset are denoted by \( f_i \). Multimodal density is then defined for each unique sample \( u_i \) in \( \{U\}_T \) and is the product of the unimodal density of the unique sample and its frequency \( f_i \).

\[ D^{MM}(u_i) = f_i \ast \frac{\sum_{j=1}^{K} q(x_j)}{2K \times q(u_i)}; \quad i = 1, 2, 3, \ldots, T. \]  

The described data-centric measures are then used to find the local modes in a given dataset. In the case of classification, the modes are learned individually for each class \( c \). Let the \( K^c \) training samples of the \( c \)-th class \((c = 1, 2, 3, \ldots, C)\) be denoted by \( \{x\}^c = \{x_1^c, x_2^c, x_3^c, \ldots, x_k^c\} \). Let \( \{U\}_T^c = \{u_1^c, u_2^c, u_3^c, \ldots, u_T^c\} \) be the set of all unique samples of the \( c \)-th class and \( \{f\}_T^c = \{f_1^c, f_2^c, f_3^c, \ldots, f_T^c\} \) the corresponding frequency of occurrence set.

The local modes and their centers are found based on multimodal densities as follows.

1. Calculate the multimodal densities \( D^{MM}(u_i) \) of all unique samples \( u_i^c \) using Equation (16).

2. Find the unique sample with the highest multimodal density, remove it from \( \{U\}_T^c \) and place it as the first element \( r_1 \) in a ranking list \( \{r\} \).

3. From the remaining samples in \( \{U\}_T^c \), remove the sample with the shortest distance to \( r_1 \) and let it be the second element \( r_2 \) in the ranking matrix \( r \). Similarly, \( r_3 \) is the element with the shortest distance to \( r_2 \). This process is repeated until \( \{U\}_T^c \) is empty.

4. Initial prototypes \( \{p\}_0^c \) are found from the local maxima of multimodal densities ranked in \( r \) according to Condition 1.

\[ \text{Condition 1: } IF \ (D^{MM}(r_1) > D^{MM}(r_{i+1})) AND \ (D^{MM}(r_i) > D^{MM}(r_{i-1})) \ THEN \ (r_i \in \{p\}_0^c). \]  

5. Once the initial prototypes are found, they are used to attract the data samples surrounding them forming data clouds which represent the local modes present in the data. The clouds are analogous to clusters except for the fact that they are nonparametric and do not conform to any particular shape. A point is assigned to the closest prototype by

\[ \text{closest prototype} = \arg\min_{p \in \{p\}_0^c} (d(x_i, p)); \quad x_i \in \{x\}^c. \]
6. The initial data clouds must then be filtered to retain only the larger and more representative clouds in order to improve the generalization ability of the model. This corresponds to increasing the level of granularity. The higher the granularity, the more information about the local modes the model is able to capture. The very first prototypes computed are at the zeroth level of granularity hence the subscript 0 in \( \{p\}_0^c \). The filtering process proceeds as follows.

a. Calculate the centers of all the data clouds \( z_i \in \{z\}_0^c \).

b. Calculate the multimodal density at the centers \( z_i \) as the product of the unimodal density weighted by the support (number of elements) of the particular cloud, \( S_i \).

\[
D^{MM}(z_i) = S_i \ast D(z_i); \quad z_i \in \{z\}_0^c.
\]

(19)

c. Calculate the average radius of local influence of each prototype at granularity level \( L \) as

\[
G_{c,L} = \sum_{x,y \in \{x\}_c^c, x \neq y, d^2(x,y) \leq G_{c,L-1}} d^2(x,y) / Q_{c,L}; \quad G_{c,0} = d^c,
\]

(20)

where \( d^c \) is the average distance between any two data samples in \( \{x\}_c^c \)

\[
\bar{d}^c = \frac{1}{K^2} \sum_{i=1}^{K} q(x_i),
\]

(21)

and \( Q_{c,L} \) are the number of pairs of data samples in \( \{x\}_c^c \) between which the distance is smaller than \( G_{c,L-1} \).

d. For each data cloud center, for example, the \( i \)th one, find elements of the set composed of the centers of its neighboring clouds \( \{z\}_{i+}^n \) through Condition 2.

Condition 2: \( IF \ (d^2(z_i, z_j) \leq G_{c,L}) \ THEN \ (z_j \in \{z\}_{i+}^n) \). \hspace{1cm} (22)

e. The prototypes at the \( L \)th granular level for class \( c \) are finally chosen as per Condition 3.

Condition 3: \( IF \ (D^{MM}(z_i) > \max_{x \in \{x\}_c^c} (D^{MM}(x))) \ THEN \ (z_i \in \{p\}_L^c) \).

(23)

In this work, the granularization level is set at 12.

With the prototypes for each of the \( C \) classes, the antecedents in Equation (10) can be created.

3 \ EXPERIMENTAL DATA ANALYSIS

In this section, RCMFE behavior of normal and faulty data from two publicly available datasets is explored.

3.1 \ Experiment 1

The first dataset used is courtesy of the Case Western Reserve University (CWRU) bearing data center.\(^{34}\) It consists of vibration data from normal/healthy bearings and as well as from faulty bearings with artificially seeded single point faults. Although faulty bearing data are sampled at both 12 and 48 kHz, only the former set from the drive end was used as the baseline healthy data were only sampled at 12 kHz.\(^{27}\) Of particular interest in this work is that the CWRU dataset was recorded for four operating conditions related to motor load and speed thus constituting four different data domains. Further, the faults are seeded at four different diameters which also introduce a shift in data distribution. There are three types of fault in the data for all the operating conditions, that is, OR, IR, and ball, all of which have data corresponding to the four fault sizes with the exception of OR data which lacks the 0.711 m dia. The organization of data is listed in Table 1. In addition, although OR data are collected in three different positions relative to the bearings’ load zone, only


| Table 1 | CWRU data description |
|---------|------------------------|
| **Tags** | **Specifications** | **Fault diameters (dia)** |
| CWRU 1  | 0 hp/1797 rpm  | IR = 0.177 m, 0.355 m, 0.5334 m and 0.711 m Ball = 0.177 m, 0.355 m 0.5334 m and 0.711 m |
| OR = 0.177 m, 0.355 m, and 0.5334 m |
| CWRU 2  | 1 hp/1772 rpm    |                       |
| CWRU 3  | 2 hp/1750 rpm    |                       |
| CWRU 4  | 3 hp/1730 rpm    |                       |

Measurements at the 6 o’clock position were used since there was no mechanism to convert the dynamometer torque into radial load borne by the bearings, and thus the only effective radial load was gravity.\(^{35}\)

Figure 3 shows some of the raw vibration signals of normal and faulty bearings from the CWRU dataset. Although the waveforms show some general differences, the raw data are not sufficiently distinguishable.

The RCMFE of healthy and faulty bearings is shown in Figure 4. Especially when viewed at higher scales, the RCMFE of faulty bearings is lower than that of healthy ones which fluctuate around a constant value. This observation is expected as a fault produces periodic impact in the vibration signal every time it is encountered thus increasing the regularity of the data or inversely, reducing its structural complexity and hence the diminishing RCMFE.

On the other hand, a healthy bearing’s RCMFE fluctuates around a constant value across the higher scales indicating both a randomness of the data but also a richness in terms of the information contained about the machine. RCMFE behaves consistently across operating conditions indicating its robustness as a feature where conditions are changing.

RCMFE values are more volatile when compared for different fault sizes but the behavior is still consistent scale-wise as seen in Figure 5. Because in practice a test sample may have any diameter of fault size within some range, the classifier’s performance will be boosted by having data from all the fault sizes available present in the training set.

Figure 5 also seems to indicate that although IR fault is readily distinguished from the other two faults, OR and ball fault data are less obviously differentiable. We therefore propose performing the diagnosis in two stages. The first stage...
will be concerned with fault isolation and thus a binary classifier will be trained on healthy and faulty data since from Figure 4 the two groups are readily differentiable. The training data will be extracted from a single operating condition and tested against features from all the other conditions as the RCMFE seems to be unaffected by operating conditions. In the second stage where fault isolation is performed, several classifiers will be built: the first will be a multiclass classifier trained on data from the three faults. From Figure 5, it is expected that this classifier will have high recall for IR fault, that is, this fault will be rarely confused for the other two classes and vice versa. The implication is that if a fault is identified as IR, the verdict has a high probability of being correct. However, since OR and ball faults are likely to be confused, once
a sample is classified as either of the two, it will be passed through a binary classifier trained exclusively on ball and OR data in an attempt to raise accuracy.

3.2 | Experiment 2

The second dataset consists of bearing diagnosis vibration data collected under different time varying speed conditions courtesy of the University of Ottawa. The operating conditions in this dataset are time-varying throughout the recording of the vibration signal. The variations are as follows: increasing speed (Cond 1), decreasing speed (Cond 2), increasing then decreasing speed (Cond 3), and decreasing then increasing speed (Cond 4). The speed ranges spanned are given in the dataset’s documentation. The data are categorized into healthy, OR, and IR classes. For each operating setting and class, three trials are conducted to increase authenticity resulting in a total of 36 files.

Figure 6 shows the first few milliseconds of some the raw waveforms of healthy data as well as IR and OR fault data for the four operating conditions.

RCMFE values are shown in the plots of Figure 7 with healthy bearing values fluctuating around a generally constant value at the higher scales while those of faulty data are monotonically decreasing.

Although Figure 7 implies good separability of the classes, we still choose the two stage approach where fault detection is first performed followed by fault isolation.
4 | METHODOLOGY

From the insight gathered from Figures 4, 5, and 7, we proposed the diagnostic process as follows:

1. Each data file was divided into 20 examples containing 4000 consecutive nonoverlapping points.
2. Each example was denoised using a bior 3.7 wavelet of level 3 and RCMFE computed for 20 scales with an embedding factor of 2. The 20 RCMFE values made up the feature vector.
3. Training data were organized according to the four operating conditions, that is, for source domain 1, the training data were obtained from CWRU 1/Cond 1 for the CWRU and Ottawa datasets, respectively. Similarly, training data for source domain 2 were drawn from CWRU 2/Cond 2, source domain 3 data were drawn from CWRU 3/Cond 3, and finally source domain 4's data were drawn from CWRU 4/Cond 4.
4. The files used from the CWRU dataset were as follows. For the fault detection stage, healthy data vs 0.177 m dia IR, OR and ball data were used. For the initial stage of fault isolation where a classifier is trained on the three fault classes, the training data consisted of 0.177 m dia, 0.355 m dia, and 0.711 m dia IR, 0.177 m dia, 0.355 m dia, and 0.533 m dia OR, and 0.177 m dia, 0.355 m dia, and 0.711 m dia ball fault data. The data were mixed in the described manner to ensure the largest fault data available for each class was used for training. Thus, because the 0.711 m dia is unavailable for OR from the dataset repository, one fault class (0.533 m dia) was skipped from the IR and ball classes to ensure class imbalance did not occur. For IR vs Ball classifier, 0.177 m dia, 0.355 m dia, 0.533 m dia, and 0.711 m dia were used for both the IR and ball classes. In the ball vs OR classifier, 0.177 m dia, 0.355 m dia, and 0.711 m dia were used for the ball class, while 0.177 m dia, 0.355 m dia, and 0.533 m dia were used for the OR class.
5. The files used from the Ottawa dataset were as follows. For the first stage, that is, fault detection, the classifiers were trained with H-A-1 vs I-A-1/O-A-1 for source domain 1, H-B-2 vs I-B-2/O-B-2 for source domain 2, H-C-3 vs I-C-3/O-C-3 for source domain 3, and H-D-1 vs I-D-1/O-D-1 for source domain 4. In the second stage of fault isolation, the classifiers were trained with I-A-1 vs O-A-1 for source domain 1, I-B-2 vs O-B-2 for source domain 2, I-C-3 vs O-C-3 for source domain 3, and I-D-1 vs O-D-1 for source domain 4. The suffixes 1, 2, and 3 in this dataset indicate the trial number. The middle letter is indicative of the operating condition, that is, A for Cond 1, B for Cond 2, C for Cond 3, and D for Cond 4.
6. During the training phase, all classes consisted of an equal number of samples to avoid class imbalance.
7. The testing/target data for each of the source domains consisted of every other sample not in the training data, that is, all remaining examples from the rest of the conditions were combined into the test set.
8. For each of the four source domains, SOF classifiers were trained and used to categorize the test samples.

The work was carried out in the Matlab environment on a 16 gb RAM, core i7 computer.

5 | RESULTS AND DISCUSSION

The CWRU and Ottawa datasets were used to verify the proposed use of RCMFE for cross domain classification and the results closely follow the insights drawn from Figures 4, 5, and 7.

Table 2 shows that in the CWRU dataset, the healthy and faulty data are perfectly separable using RCMFE with a 100% on all the test data regardless of which source domain was used in training. The column headers indicate the source domain of the training data while test data include data from all other conditions not used for training.

For fault isolation, a preliminary classifier was trained on all the three fault classes, that is, IR, OR, and ball (Table 3). Just as indicated by Figure 5 the IR fault is perfectly separable from the other two faults. Further more, IR is never confused for OR and vice versa. However, since some ball faults are categorized as IR faults as seen in the last rows of the confusion matrices in Table 3, it would be worthwhile to train a binary classifier on IR and ball fault data. The accuracy of the IR/ball classifier is shown in (Table 4) where it is seen that the two faults exhibit good separability.

The entries of the matrices of Table 3 concerning OR and ball faults confirm that the two are frequently confused for each other as intimated by Figure 5. Thus, samples categorized as either fault should be passed through a binary classifier trained on ball and OR data only. Table 5 shows the accuracy of such classifiers.

In the end, the accuracy of label assignment to a sample is then an average of accuracies of all the classifiers the samples pass through.

| TABLE 2 | CWRU dataset: Confusion matrices for the fault detection stage |
|---------|----------------------------------------------------------|
| Predicted Label | CWRU 1 | CWRU 2 | CWRU 3 | CWRU 4 |
| Actual Label | H | F | H | F | H | F | H | F |
| H | 60 | 0 | 60 | 0 | 60 | 0 | 60 | 0 |
| F | 0 | 380 | 0 | 380 | 0 | 380 | 0 | 380 |
| Accuracy | 100% | 100% | 100% | 100% |

| TABLE 3 | CWRU dataset: Confusion matrices for the preliminary round in fault isolation |
|---------|----------------------------------------------------------|
| Predicted Label | CWRU 1 | CWRU 2 | CWRU 3 | CWRU 4 |
| Actual Label | IR | OR | Ball | IR | OR | Ball | IR | OR | Ball | IR | OR | Ball |
| IR | 130 | 0 | 0 | 130 | 0 | 0 | 130 | 0 | 0 | 130 | 0 | 0 |
| OR | 0 | 63 | 27 | 0 | 81 | 9 | 0 | 77 | 13 | 0 | 71 | 19 |
| Ball | 10 | 41 | 69 | 9 | 47 | 74 | 12 | 29 | 89 | 6 | 26 | 98 |
| Accuracy | 74.8% | 81.4% | 84.5% | 85.4% |

| TABLE 4 | CWRU dataset: Confusion matrices for binary classifiers trained with IR and Ball data only |
|---------|----------------------------------------------------------|
| Predicted Label | CWRU 1 | CWRU 2 | CWRU 3 | CWRU 4 |
| Actual Label | Ball | IR | Ball | IR | Ball | IR | Ball | IR | Ball |
| Ball | 120 | 0 | 120 | 0 | 120 | 0 | 112 | 8 |
| IR | 7 | 113 | 6 | 114 | 11 | 109 | 3 | 117 |
| Accuracy | 97% | 97.5% | 95.4% | 95.4% |
From Figure 5, it was expected that differentiating ball and OR faults would be more challenging than either IR versus ball or IR versus OR. In the figure, RCMFE values for OR and ball data from different fault sizes divide themselves into a high or low range. Further, the fault size that ends up in the high range differs for the two faults, for example, in CWRU 1, the 0.355 m dia fault is in the higher range for OR fault but in the low range for ball fault. On the other hand, the 0.177 m dia and 0.533 m dia sizes, which are in the higher range for ball fault, are in the lower range with OR fault. For the rest of the operating conditions, the 0.355 m dia is always in the high range for OR fault while the other two sizes remain in the lower range. For the ball fault, in CWRU 2, the 0.177 m dia and 0.711 m dia sizes are in the higher range and the rest in the lower range. For CWRU 3 and CWRU 4 ball faults, the 0.177 m dia and 0.533 m dia sizes are in the higher range but the 0.355 m dia size steadily climbs from the low range to the middle. The 0.711 m dia remains in the low range. By observing this behavior across the operating conditions, it is probable that the segmentation of fault sizes in the ball and OR fault may be a peculiarity of the CWRU setup and is thus possible that the phenomenon may lessen in another machine.

The RCMFE features result in high accuracy in the two stages for the Ottawa dataset as seen in the confusion matrices of Tables 6 and 7. As observed in the CWRU dataset, healthy and fault data are easily distinguishable as well as IR and OR data.

Thus, RCMFE features are consistent in detecting fault and in at least differentiating IR and OR faults as verified with the CWRU and Ottawa datasets. The features also perform well in distinguishing IR and ball faults and only suffer a drop in performance when differentiating ball and OR fault.

### 5.1 Comparison with related works

In published works that have tackled cross-domain diagnosis with the CWRU dataset, the domains are constituted by operating conditions as is the case in this article. However, despite the differences in the methods applied to feature
extraction and DA, a key similarity in the literature is that each fault size is considered in a separate class with the exception of the 0.711 m diameter fault which is mostly omitted. Thus, there are usually a total of 10 classes, with the remaining three fault sizes and three fault types contributing nine classes and the healthy class bringing the number of categories to 10. As a result of preliminary investigations on the dataset as shown in Figures 4 and 5, it is seen that the RCMFE features are very robust against operating conditions and the challenge in diagnosis would be presented by the varying fault sizes even in the same operating condition/domain. Therefore, the results presented in section 5 focused on the case where each class was constituted by a single fault type but with at least three of all the fault sizes available for the particular fault type. This is because in the OR fault data the 0.711 m diameter fault is missing from the data repository and appropriate adjustments had to be made to avoid class imbalance. In order to compare the method proposed in this article with other works, data were similarly prepared such that a class consisted of a single fault type and single fault size. Table 8 shows some of the results for various fault sizes for the latter arrangement. Once again, the condition shown in the column header is the training domain while the test data were drawn from the other three domains/operating conditions.

The results are very promising as noted by the high accuracy of classification. The good separability of the classes when only a single fault size is considered was expected as depicted in Figure 4.

In Table 9, a comparison with other works that have used CWRU data is given. The comparison is made in terms of the method of feature generation, the DA technique, the total number of classes considered, the fault size per class, average accuracy, and whether the target domain data are required during training. The subscript * in the proposed approach indicates the arrangement of one fault size per class. Total direct comparability is not possible due to some minor differences. For instance, Li et al.19 only presented results considering CWRU 1 and CWRU 4 as the source and target domains while in the rest of the entries, each condition was considered as the source in its own turn. Also, in the reported literature, the 0.711 m diameter fault was entirely omitted from analysis. In the approach reported here, diagnosis was divided into two stages, that is, fault detection followed by isolation while in the rest of the entries in Table 9, the diagnosis process was performed in one stage. Therefore the accuracy of the proposed method is an average of accuracy from the two stages. The number of classes are listed as 3 for this article because separate classifiers were built for each of the fault sizes thus the classes only consisted of the three fault types for a particular size.

As is seen from Table 9, in order to handle cross-domain diagnosis, most of the researchers have adopted deep learning for feature extraction as indicated by the use of CNN. Also, MMD that is an optimization problem solved by iteration is commonly used to reduce distribution discrepancy. With the exception of Zhang et al.16 the other works require that all or part (Li et al9) of target domain data be present during training. This means the learning process must be repeated for

**Table 8** CWRU dataset: Confusion matrices fault isolation considering a single fault size

| Actual Label | Predicted Label | CWRU 1 (0.177 m dia) | CWRU 2 (0.355 m dia) | CWRU 3 (0.533 m dia) | CWRU 4 (0.355 m dia) |
|--------------|----------------|----------------------|----------------------|----------------------|----------------------|
| IR           | IR             | 30                   | 0                    | 0                    | 0                    |
| OR           | IR             | 30                   | 0                    | 0                    | 0                    |
| Ball         | Ball           | 0                    | 30                   | 0                    | 0                    |
| Ball         | Ball           | 0                    | 0                    | 30                   | 0                    |
| Ball         | Ball           | 0                    | 0                    | 0                    | 30                   |

**Table 9** Comparison with other works using the CWRU dataset

| Resource     | Feature Generation | DA Method                   | No. of Classes | Fault sizes per class | Avg. Accuracy | Target Required |
|--------------|--------------------|-----------------------------|----------------|-----------------------|---------------|-----------------|
| Li et al19   | CNN                | Multilayer & Multikernel MMD| 10             | 1                     | >90%          | Yes             |
| Zhang et al16| CNN                | Adaptive batch normalization| 10             | 1                     | >90%          | No              |
| Li et al9    | CNN and GNN        | Multikernel MMD             | 10             | 1                     | >90           | Yes             |
| Tong et al18 | PCA                | MMD                         | 10             | 1                     | 100%          | Yes             |
| Wang et al15 | CNN                | Adversarial learning        | 10             | 1                     | >96%          | Yes             |
| *Proposed approach | RCMFE            | None                        | 3              | 1                     | >97%          | No              |

Abbreviations: CNN, convolutional neural networks; GNN, generative neural networks; MMD, maximum mean discrepancy.
each new domain of test data. In contrast, the RCMFE features approach proposed in this article does not require deep learning for feature extraction and neither does it require any DA technique. Furthermore, no target domain data are required during training which means that the learning process need not be repeated every time data from a new domain is tested. For the generally equivalent performance, RCMFE features allow for cross-domain analysis while avoiding the iterative solving of complex objective functions.

In practical implementation, categorizing each fault size in an individual class does not benefit generalization. This is because, rather than an unseen test sample having a fault size exactly equal to one of those used for training, it is more probable that the fault will be in the range of the fault sizes used for training. Thus, in this work, we opted for the harder transfer learning problem where classes were based on the type of fault only and in the training data several fault sizes were included. The more challenging nature of the latter arrangement is clear by comparing Figures 4 and 5.

6 CONCLUSION

In this article, the use of RCMFE features for cross domain classification has been introduced and its effectiveness verified using the CWRU dataset where data were collected for different operating conditions and the Ottawa dataset where data were collected under time-varying operating conditions. Because it is impractical to know and have data from all possible operating conditions, the proposed diagnostic procedure only uses training data from a single condition and has been proven to perform well against test data from unseen operating conditions. The approach is straightforward and computationally desirable since primarily, it only involves extracting RCMFE features followed by classification. It takes 2.4 seconds to compute fuzzy entropy at 20 scales for a 4000 point time series, and classification using the SOF classifier takes between 0.2 and 0.5 second. This method compares favorably with deep learning DA approaches which require nontrivial manipulation of data before training the classifiers, for example, reducing the difference in distributions of the source and target or learning a subspace wherein features are transferrable across domains. In contrast, our approach only requires extracting RCMFE from data. The classification task is divided into two with fault detection being followed by fault isolation. Performance is excellent in the former stage but drops slightly in fault isolation due to the fact that ball and OR faults are not very distinguishable when data from all available fault sizes are used for training. By considering several fault sizes simultaneously for training, our work is unlike other published works which consider each fault type and size as a separate class. We argue that although the latter arrangement results in high classification accuracy using deep learning approaches as well as the proposed approach, it is not practical as it is unlikely that a test sample will have a fault size exactly equal to one of those used in training. Rather, it is more probable that the unseen size will be in the range of the sizes used for training.

In the case of the CWRU dataset where ball and OR faults are less distinguishable when several fault sizes are combined in training, an avenue of further investigation is opened. Another future direction is to consider using the data from all recording channels available for diagnosis. For instance, CWRU data are collected from the fan end, drive end, and in some cases an accelerometer was placed on the base of the machine making at least two channels available for all the files in the repository. RCMFE has already been used in a multichannel approach in other fields and its suitability for cross domain bearing diagnosis should be explored.

ACKNOWLEDGMENTS

We wish to thank the Mwalimu Nyerere African Union Scholarship Scheme 2018 and the Jomo Kenyatta University of Agriculture and Technology for their financial facilitation that enabled this work.

PEER REVIEW INFORMATION

Engineering Reports thanks the anonymous reviewers for their contribution to the peer review of this work.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

AUTHOR CONTRIBUTIONS

Esther W. Gituku: Conceptualization-Equal, Formal analysis-Lead, Investigation-Lead, Methodology-Lead, Software-Lead, Writing-original draft-Lead, Writing-review and editing-Lead. James K. Kimotho: Conceptualization-Equal, Data curation-Lead, Project administration-Equal, Supervision-Lead, Writing-review and editing-Supporting.
Jackson G. Githu: Conceptualization-Supporting, Project administration-Equal, Supervision-Supporting, Writing-review and editing-Supporting.

DATA AVAILABILITY STATEMENT
The two datasets used in this work are publicly available online at https://csegroups.case.edu/bearingdatacenter/pages/download-data-file for the CWRU data, and https://data.mendeley.com/datasets/v43hmbwxpm/1 for the Ottawa dataset.

ORCID
Esther W. Gituku https://orcid.org/0000-0002-2512-1242

REFERENCES
1. Zhang S, Zhang S, Wang B, Habetler TG. Machine learning and deep learning algorithms for bearing fault diagnostics – a comprehensive review; 2019. arXiv e-prints:1901.08247.
2. IEEE. Report of large motor reliability survey of industrial and commercial installations. IEEE Trans Ind Appl. 1985;21:853-864.
3. Chang C-W, Lee H-W, Liu C-H. A review of artificial intelligence algorithms used for smart machine tools. Inventions. 2018;3(41):1-28.
4. Gupta P, Pradhan M. Fault detection analysis in rolling element bearing: a review. Mater Today Proc. 2017;4:2085-2094.
5. Yu W, Li Y, Xu M, Wenhua H. A review of early fault diagnosis approaches and their applications in rotating machinery. Entropy. 2019;21:409–1–409–26.
6. Aljundi R, Emone R, Muselet D, Sebban M. Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA; 2015.
7. Zubizarreta U.L. Hybrid modelling in condition monitoring [PhD thesis]. Lulea University of Technology; 2016.
8. Lessmeier C, Kimotho JK, Zimmer D, Sextro W. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: a benchmark data set for data-driven classification. Paper presented at: Proceedings of the European Conference of the Prognostics and Health Management Society, Bilbao, Spain; 2016.
9. Li X, Zhang W, Ding Q. Cross-domain fault diagnosis of rolling element bearings using deep generative neural networks. IEEE Trans Ind Electron. 2019;66:5525-5534.
10. Csurka G. Domain Adaptation for Visual Applications: A Comprehensive Survey. New York, NY: Springer; 2017:1-35.
11. Jing M, Li J, Zhao J, Lu K. Learning Distribution-Matched Landmarks for Unsupervised Domain Adaptation. New York, NY: Springer; 2018:491-508.
12. Chen M, Xu ZE, Weinberger KQ. Marginalized denoising autoencoders for domain adaptation. Paper presented at: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland; 2012.
13. Lu W, Liang B, Cheng Y, Meng D, Yang J, Zhang T. Deep model based domain adaptation for fault diagnosis. IEEE Trans Ind Electron. 2017;64(3):2296-2305.
14. Fernando B, Habrard A, Sebban M, Tuyltelaars T. Subspace alignment for domain adaptation. Paper presented at: Proceedings of the IEEE Conference on Computer Vision (ICCV), Sydney, Australia; 2013.
15. Wang Q, Michau G, Fink O. Domain adaptive transfer learning for fault diagnosis; 2019. arXiv e-prints:1905.06004.
16. Zhang W, Peng G, Li C, Chen Y, Zhang Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors. 2017;17:1-21.
17. Lu W, Liang B, Meng D, Yang J, Zhang T. Deep model based domain adaptation for fault diagnosis. Mech Syst Signal Process. 2017;64:2296-2305.
18. Tong Z, Li W, Zhang B, Zhang M. Bearing fault diagnosis based on domain adaptation using transferable features under different working conditions. Shock Vib. 2018;2018:1-12.
19. Li X, Zhang W, Ding Q, Sun J-Q. Multi-layer domain adaptation method for rolling bearing fault diagnosis. Signal Process. 2018;157:180-197.
20. Zhu K, Li H. A rolling element bearing fault diagnosis approach based on hierarchical fuzzy entropy and support vector machine. J Mech Eng Sci. 2015;230:2314-2322.
21. Li Y, Wang X, Liu Z, Liang X, Si S. The entropy algorithm and its variants in the fault diagnosis of rotating machinery: a review. IEEE Access. 2018;6:66723-66741.
22. Azami H, Fernandez A, Escudero J. Refined multiscale fuzzy entropy based on standard deviation for biomedical signal analysis. Med Biol Eng Comput. 2017;55:2037-2052.
23. Zheng J, Pan H, Chen J. Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines. Mech Syst Signal Process. 2017;85:746-759.
24. Zhu K, Li H. A roller bearing fault diagnosis method based on hierarchical entropy and support vector machine with particle swarm optimization algorithm. Measurement. 2014;47:669-675.
25. Wu S-D, Wu C-W, Lin S-G, Wang C-C, Lee K-Y. Time series analysis using composite multiscale entropy. Entropy. 2013;15:1069-1084.
26. Zhang L, Xiong G, Liu H, Guo W. Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference. Expert Syst Appl. 2010;37:6077-6085.
27. Li Y, Wang X, Si S, Huang S. Entropy based fault classification using the case western reserve university data: a benchmark study. *IEEE Trans Reliab*. 2019;69:754-767.
28. Chen W, Wang Z, Xie H, Yu W. Characterization of surface emg signal based on fuzzy entropy. *IEEE Trans Neural Syst Rehabil Eng*. 2007;15:266-272.
29. Gu X, Angelov P. Self-organising fuzzy logic classifier. *Inf Sci*. 2018;447:36-51.
30. Angelov P, Gu X. *Empirical Approach to Machine Learning*. Cham: Springer; 2019.
31. Costa M, Goldberger AL, Peng C. Multiscale entropy analysis of complex physiologic time series. *Phys Rev Lett*. 2002;89:068102–1-068102–4.
32. Wu S, Wu C, Lin S, Lee K, Peng C. Analysis of complex time series using refined composite multiscale entropy. *Phys Lett A*. 2014;378:1369-1374.
33. Angelov P, Yager R. Self-organising fuzzy logic classifier. *Int J Gen Syst*. 2011;41:163-185.
34. Case Western Reserve University Western Bearing Data Center Seeded fault test data. data retrieved from case Western Reserve University Bearing Data Center Website; 2018, https://csegroups.case.edu/bearingdatacenter/pages/.
35. Smith WA, Randall RB. Rolling element bearing diagnostics using the case western reserve university data: a benchmark study. *Mech Syst Signal Process*. 2015;64–65:100-131.
36. Huang H, Baddour N. Bearing vibration data collected under time-varying rotational speed conditions. *Data Brief*. 2018;21:1745-1749.
37. Azami H, Escudero J. Refined composite multivariate generalized multiscale fuzzy entropy: a tool for complexity analysis of multichannel signals. *Phys A Stat Mech Appl*. 2017;465:261-276.

How to cite this article: Gituku EW, Kimotho JK, Njiri JG. Cross-domain bearing fault diagnosis with refined composite multiscale fuzzy entropy and the self organizing fuzzy classifier. *Engineering Reports*. 2021;3:e12307. https://doi.org/10.1002/eng2.12307