IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 15, 2020, accepted August 29, 2020, date of publication September 8, 2020, date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022770

Discriminant Feature Extraction for Centrifugal Pump Fault Diagnosis

ZAHOOR AHMAD[1], AKHAND RAI[1,2], ANDREI S. MALIUK[1], AND JONG-MYON KIM[1], (Member, IEEE)

[1]School of Electrical, Electronics, and Computer Engineering, University of Ulsan, Ulsan 44610, South Korea
[2]School of Engineering and Applied Sciences, Ahmedabad University, Ahmedabad 380009, India

Corresponding author: Jong-Myon Kim (jmkim07@ulsan.ac.kr)

This work was supported in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP), and in part by the Ministry of Trade, Industry, and Energy (MOTIE), South Korea, under Grant 20181510102160.

ABSTRACT Raw statistical features can imitate the amplitude, average, energy and time, and frequency series distribution of a raw vibration signal. However, these raw statistical features are either not very sensitive to weak incipient faults or are unsuitable for more severe faults, thus affecting the fault detection and classification accuracy. To tackle this problem, this paper proposes a discriminant feature extraction method for Centrifugal Pump (CP) fault diagnosis. In order to obtain the discriminant feature pool, the proposed method is divided into three phases. In the first phase, a healthy baseline signal is selected. In the second phase, the healthy baseline signal is cross-correlated with the CP vibration signals of different classes, and a set of new features are extracted from the resulting correlation sequence. In the third phase, raw hybrid features in time, frequency, and the time-frequency domain are extracted from both the healthy baseline signals and the CP vibration signals of different classes. The correlation coefficient is calculated between the raw hybrid feature pools, which results in a new set of discriminant features. Discriminant features help the machine learning classifiers to effectively detect and classify the data into its respective classes. Furthermore, the proposed method combines all these features into a single feature vector that forms a vulnerable feature pool. The vulnerable feature pool describes the CP's vulnerability to a fault and is provided as an input to a multiclass support vector machine (MSVM) for CP fault detection and classification. The experimental results illustrate that the accuracy obtained from the proposed method shows promising improvements over the state-of-the-art conventional methods.

INDEX TERMS Centrifugal pump, cross-correlation, correlation coefficient, fault classification, mechanical faults, vulnerable feature pool.

I. INTRODUCTION

CP's have become an essential part of everyday business. It has been estimated that around 20% of the energy generated worldwide is dedicated to driving pumping systems [1]. CP's likely account for the highest percentage of pumps used in industry [2]. CP's are simple in construction, reliable in operation, and cheap in cost. However, the consequences of their unexpected failure include costly repairs, long downtime, economic losses, and reduced safety of operating staff. For this reason, health management has been attracting increasing attention to ensure the efficiency and reliability of pump operation.

The associate editor coordinating the review of this manuscript and approving it for publication was Min Xia.

A survey made by major organizations shows that with an investment of 10,000 – 20,000 dollars in health management, one can save up to 500,000 dollars per year [3]. In the past decades, several health management strategies have been developed, out of which predictive maintenance, also recognized as condition-based maintenance (CBM) is the most beneficial one. CBM suggests maintenance action based on the information collected through condition monitoring and allows the machine to have maximized running time with minimum maintenance costs [4]. As such, this paper uses the condition-based maintenance strategy for CP fault diagnosis.

Faults in the CP can be divided into mechanical faults (MF) and fluid flow related hydraulic faults (FHF) [5]. In S.M. Chittora's thesis [6], he mentioned that 39% of pump failures occur due to mechanical seal faults. Another study [7],

explains that impeller imbalance can result in MF or FHF. To avoid CP failure, the early fault detection of both mechanical seal and impeller faults is of great significance. Hence, in this study, MF, which include mechanical seal and impeller faults, are taken into consideration.

MF or failures occurring in a CP result in abnormal vibration Hence, vibration signals are commonly used for diagnosing MFs [8], [9]. However, abnormal vibration and impact signals with low amplitudes can be disguised by heavy background noise. Many signal processing approaches in the time, frequency, and time-frequency domain have been previously developed to extract appropriate fault features from raw vibration signals in order to detect MFs.

MFs in CP produce shocks resulting in variation of the vibration signal's amplitude and distribution [10]. The time-domain statistical features are utilized to capture these variations [11]–[15]. However, they are not sensitive to the changing severity of faults and result in less discriminate features. The frequency spectrum is more sensitive to faults, since an undetectable change will produce a spectrum line in the corresponding frequency spectrum. Rapur and Tiwari [16] use frequency domain features for CP fault diagnosis. However, the fault vibration signals acquired from CP are highly non-stationary and Fourier Transforms (FT) are appropriate for stationary processes only [17], [18]. Moreover, interpreting the characteristic frequency components from the spectrum requires both mathematical and experimental validation. To overcome the non-stationary behavior of CP vibration signal, time-frequency domain methods have been introduced in the past few decades. References [19]–[22], [23] utilized wavelet transform (WT) to process CP vibration signals and extract discriminant features for CP fault segregation through machine learning models. However, selecting an optimal mother wavelet function for non-stationary vibration signals is still a problem. Further, determining the proper wavelet requires many experiments and subjective judgments in the experimental processes [24], [25]. Empirical Mode Decomposition (EMD) introduced in [26], a self-adaptive signal decomposition technique can overcome the shortcomings of the wavelet transform. EMD has been effectively used to extract discriminant information from the signal for the fault diagnosis of rotating machinery such as rotors, bearings, and gears [27] and CP faults [28]. Yet, the EMD technique has limitations that can cause problems when it is used to discern the nature of extracted information for machinery fault diagnosis. Particularly, mode-mixing can pose problems for the non-orthogonal components of decomposition. Ensemble Empirical Mode Decomposition (EEMD), a variant of EMD, was proposed to solve the problem of mode-mixing [29]. However, the selection of important intrinsic mode functions (IMFs) is a major challenge in EEMD [30]. Although EMD and its variants show outstanding performance in processing complex non-stationary and nonlinear signals, they suffer from the problem of extreme interpolation, sifting stop criterion, end effects, and a limited theoretical foundation [31].

In the light of the presented literature, vibration signal preprocessing and discriminant feature extraction are of primary importance for fault diagnosis. Discriminant features help the machine learning classifiers to classify the data effectively into respective classes. Considering the shortcomings of existing signal analysis techniques, most recent fault diagnosis studies have focused on extracting hybrid feature space from the vibration signals [32]–[36]. Prosvirin *et al.* [34] preprocessed the vibration signal using EEMD and extracted hybrid features for the diagnosis of rub-impact faults with an accuracy of 99.8%. Sohaib *et al.* [35] used a hybrid feature pool with a stacked sparse autoencoder based on deep neural networks and perform diagnosis of bearing faults having multiple severities with an accuracy of 99.10%. Hybrid feature space helps to overcome the effects of common fluctuations and random impulses in the vibration signal. However, the limited understanding of deep learning-based methods makes it difficult to define features and perform vibration signal preprocessing, and these methods are also computationally expensive.

To overcome the above challenges, a correlation function based feature extraction technique is proposed in this paper. Two types of correlation functions have been considered, namely, cross-correlation and correlation coefficient. The cross-correlation represents similarities between two signals, in the time domain and bears a low computational cost [37]. The correlation operation has been widely used for Brain Computer Interface oriented signals, such as EMGs and EEGs for human brain disease classification with an accuracy of 100% [38]–[40]. On the other hand, correlation coefficient finds the relationship between variables. Shi [41] performed bearing fault diagnosis using correlation coefficient and simplified neutrosophic sets.

Despite the good discriminant information extraction capabilities and low computational cost of cross-correlation, to the best of our knowledge, preprocessing the time-domain vibration signal using cross-correlation for CP fault diagnosis has not been reported so far. This could be due to the fact that for cross-correlation, a baseline healthy signal is required to which the CP vibration signals obtained during different CP operating conditions can be compared. Further, the random fluctuations caused by multiple severity faults in the CP cannot be addressed by analyzing the signal only in the time or frequency or time-frequency domain. However, hybrid features that are extracted from the vibration signal in time, frequency, and the time-frequency domain combined together can provide enough information for CP fault classification. Yet, as far as our knowledge is concerned, CP fault diagnosis using hybrid features has been reported rarely. This is because the hybrid features are extracted separately in time, frequency, and time-frequency domain and then combined into a single feature vector forming a high-dimensional feature space. The high-dimensional feature space may produce poor diagnosis results if the features are not sufficiently separable. Thus, further preprocessing is needed to obtain highly discriminant features from the high-dimensional feature space.

To address the above problems, this study proposes a new feature extraction method that considers, (i) preprocessing the vibration signals using cross-correlation function to extract correlogram features and (ii) preprocessing the features extracted from CP vibration signals using correlation coefficient. The proposed method results in discriminant features as preprocessing the vibration signals and corresponding feature space helps in overcoming the effects of likely random impulses in the CP vibration signal. Finally, in this study, an MSVM is employed to learn the fault feature patterns and perform diagnosis, considering its computational and classification efficiencies [42]–[44].

The major contributions of this work are:

- **Healthy baseline signal selection:** Healthy baseline signal selection is important for cross-correlation, this paper presents a method for the selection of a healthy baseline signal for CP fault diagnosis.
- **Preprocessing of CP vibration signal in the time domain:** The healthy baseline signal is cross correlated with the CP vibration signal and a correlogram is obtained. After obtaining the correlograms, this work extracts a new set of discriminant features from the correlogram sequences.
- **New set of discriminant features obtained from hybrid features preprocessing:** This paper analyzes both of the health baseline and CP vibration signals in time, frequency, and the time-frequency domain and extracts raw statistical hybrid features in each domain. As a result, two hybrid feature pools were created, one for the healthy baseline signal, and the other one for the CP vibration signal obtained after a different operating condition. This work then calculates the correlation coefficient between the two hybrid feature pools, which results in a new set of discriminant features for CP fault diagnosis.
- **Vulnerable feature space creation:** This work combines all of these features into a single feature vector and forms a vulnerable feature space, which is more discriminating and helpful for CP fault segregation. The feature pool is named vulnerable feature pool because the discriminant features obtained after preprocessing the vibration signal and raw statistical features are based on similarity. The more the feature or signal is like a healthy baseline feature or signal the less vulnerable. Similarly, the less the feature or signal is similar to the healthy baseline feature or signal the more vulnerable.

i The remaining sections of this study are organized as follows. Technical background is discussed in Section II. Section III provides details about test rig and the experimental procedure. The proposed method for discriminant feature extraction based on correlations is explained in Section IV. In Section V, the proposed method is applied to the fault diagnosis of CPs and comparisons of the results with other state of the art reference methods are reported. Finally, Section VI draws conclusions and provides suggestions for future research.

II. TECHNICAL BACKGROUND

A. REVIEW OF CROSS-CORRELATION

A cross-correlation is a mathematical operation that describes the mutual relationship that exists between two signals [37] [45]. This mutual relationship or similarity between the signals is represented by a sequence called correlogram. If a signal is correlated with itself, the resultant correlogram is called the autocorrelation. Assume that, $p(n)$ and $q(n)$ are two finite energy signals. The cross-correlation of the two signals are given by Eq.1.

$$R_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} p_{n+m} q_n & m \geq 0 \\ R_{yx}(-m) & m < 0 \end{cases} \tag{1}$$

where $m = -(N-1), ..., -2, -1, 0, 1, 2, ..., (N-1)$, The variable m denotes the time shift, also known as the lag, and $R_{xy}(m)$ represents the correlation between two signals p and q with a time shift of 'm'. For $m \geq 0$, signal $p(n)$ leads the signal $q(n)$ by 'm' positions. And for $m < 0$, $p(n)$ lags behind the signal $q(n)$ by 'm' positions. If the signals $p(n)$ and $q(n)$ have a finite number of samples N then the resultant correlogram has $2N-1$ samples.

B. REVIEW OF WAVELET PACKET TRANSFORM

The WT provides multiresolution analysis of a signal in independent frequency bands. The wavelet packet transform (WPT) is derived from the basic WT which decomposes a given signal into k levels. The WPT splits the input signal using low-pass and high-pass filters and creates 2^k nodes at each level. In this manner, the WPT overcomes the poor resolution of the WT because it provides a comprehensive time-frequency analysis of the input signal at both the high and low frequencies. Each level of the WPT provides a frequency range that is twice as wide as the proceeding level and half as wide as the preceding level. Fig. 1 shows the WPT structure having three levels. Eqs. 2 and 3 show the mathematical description for WPT coefficients.

$$c_{k+1}^{2j}(n) = c_k^j \times h(-2n), \quad 0 < j < 2^k - 1 \tag{2}$$

$$d_{k+1}^{2j+1}(n) = d_j^k(n) \times g(-2n), \quad 0 < j < 2^k - 1 \tag{3}$$

where h and g in eqs. 2 and 3 are the low pass and high pass filters associated with the mother wavelet. In WPT, the nodes (frequency parameters) are denoted by $2j$ and $2j+1$, and the levels (scale parameters) are denoted by k.

C. REVIEW OF SUPPORT VECTOR MACHINE

A SVM is a supervised machine learning model that is based on statistical learning theory. Fundamentally, the SVM is used for the classification of data into binary classes. As the SVM is a supervised classification technique, it uses a set of training data or/input data to create a hyperplane that classifies the data into two classes. Once the SVM is trained, any new set of data, called the testing data, is mapped into the same space and are characterized based on which side

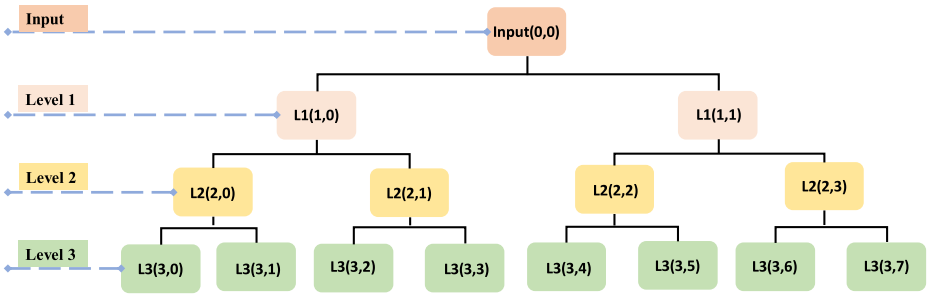

FIGURE 1. Illustration of a three-level wavelet decomposition.

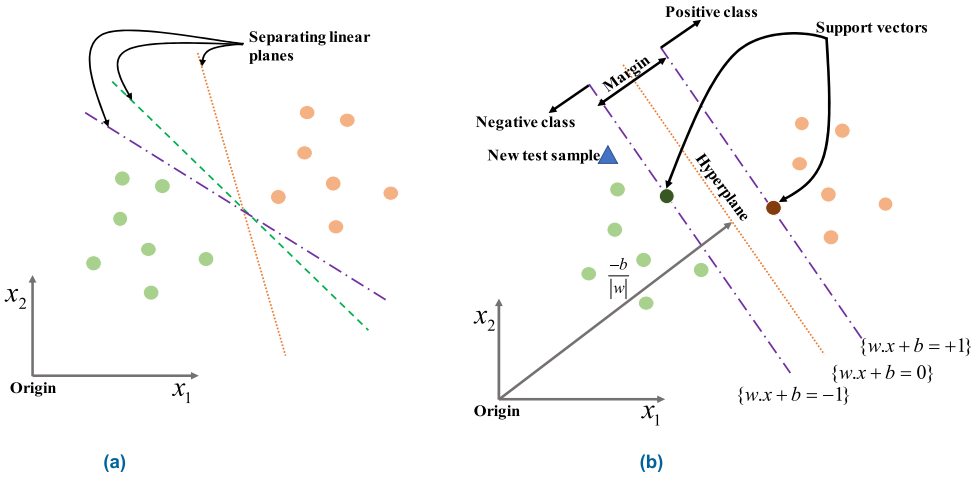

FIGURE 2. (a) Various possibilities of linear boundaries for data classification (b) SVM classifier with an optimal hyperplane.

of the margin it belongs to. Practically, there are numerous possibilities for classifying the data using a linear boundary, as shown in Fig.2a. But the Support Vector Classifier (SVC) places the boundary such that the margin is maximized and that's why the SVM is also called a large margin classifier. If the input training data $T = \{(l_1, v_1), \ldots, (l_i, v_i)\}$, where $l_i \in R^N$ is the training data and $v_i \in \{-1, +1\}$ gives the class labels, The SVC places the boundary between the classes such that the margin is maximized. The margin can be defined as the Euclidean distance between the separating hyperplanes and the data points, also called support vectors, of every class. Fig. 2b. shows an SVC classifying data into two classes. Thus, the hyperplane optimization problem can be specified as,

$$\begin{aligned} &\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{B}\xi_i \\ &\text{Suject to } \mathrm{v}_i\left[w^T.\phi(l_i) + b\right] \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0, \quad i = 1, 2, \ldots, S \end{aligned} \tag{4}$$

In Eq. 4, ξ_i is the slack variable that accounts for false classification, B is the number of training points, b is the bias, w is the weight vector, and C is the cost function which explains a tradeoff between the model complexity and learning error. If the data is not linearly distinguishable, then it is transformed to a higher dimensional feature space from the input space by using a nonlinear mapping ϕ_i, in which the data becomes linearly distinguishable. These kinds of transformations are computationally expensive. To solve this problem, a kernel function can be used for such a transformation. The kernel function is given by $K(x_i, x_j) = \phi^T(x_i).\phi(x_j)$. In the present study, a Gaussian Radial Basis Function (RBF) kernel is used.

A detailed literature review for the SVM can be found in [46]. The SVM was fundamentally developed as a binary classifier. However, in practice, there can be more than two classes of the data to be classified. Thus, the data is converted into multiple binary classes and One Against All (OAA) or Direct Acyclic Graph or One Against One techniques may be used for multiclass classification. In this work, the OAA strategy is adopted.

III. TEST RIG SETUP AND EXPERIMENTAL PROCEDURE

For experimental purposes, a test rig has been developed that consists of different parts: the CP (PMT-4008 a widely used pump in the industry) that is driven by a 5.5kW motor, a control panel with an ON/OFF switch, speed controller, flowrate controller, temperature controller, water supply controller, and display screens; pressure gauges, clear steel pipes, and two tanks (main tank and buffer tank). To maintain the net

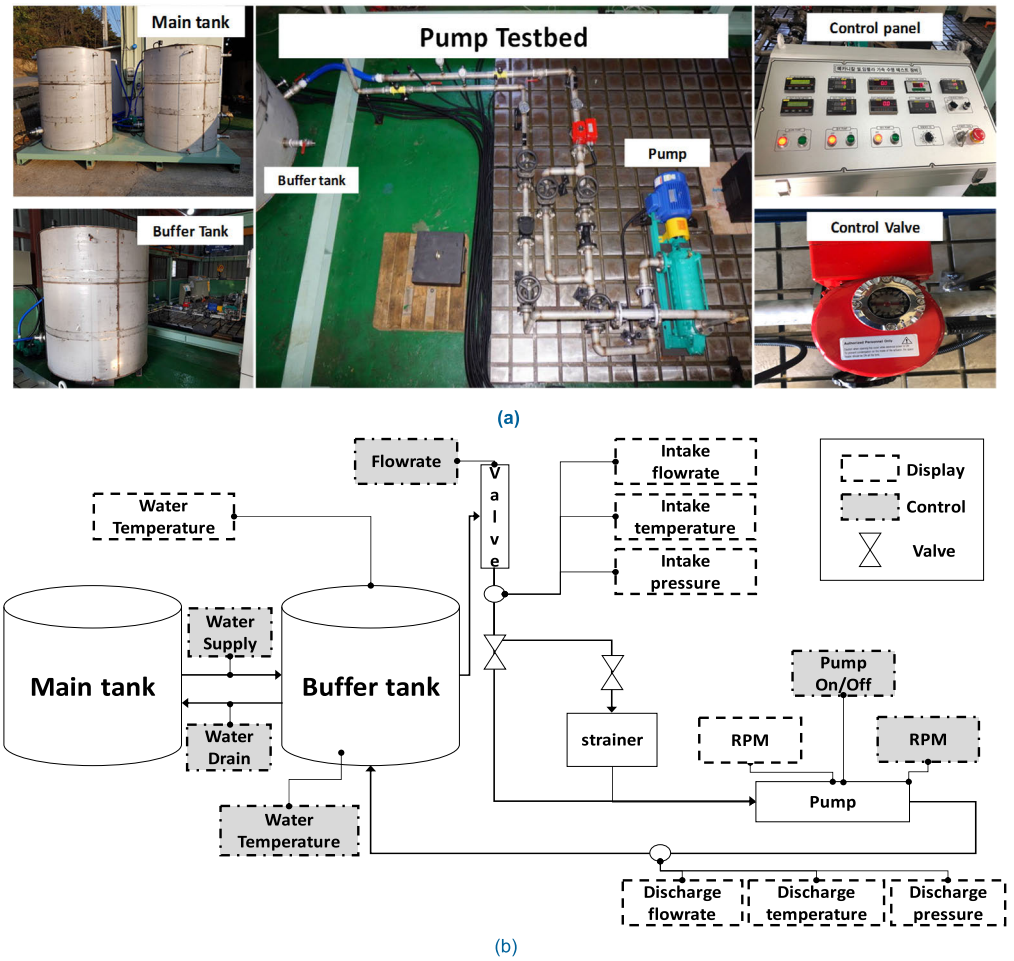

FIGURE 3. (a) Test reg setup for CP fault simulation (b) Schematics of the test rig setup.

positive suction head (NPSH) at the pump inlet for normal operation of the CP, the water tank was placed at a sufficient height. The test rig setup, as well as a schematic of the setup, are shown in Fig. 3a and 3b. After establishing the main setup, the test rig was operated for circulating water in a closed loop. The CP vibration data were collected at a constant speed of 1,733 rpm using four accelerometers, two of which were on the pump casing using adhesive, and for the other two, one was mounted close to the mechanical seal and the other was mounted near the impeller. Each sensor records the vibration of the pump using an independent channel. The recorded vibration signal was then passed to a signal monitoring unit, where the signal passes through a National Instruments 9234 device, which digitizes the acquired vibration signal. Details for the collection of data are specified in Table 1.

Data is acquired for a period of 300sec at a sampling rate of 25.6 kHz. The sampling rate was kept high because of the mechanical seal excitation frequencies, which occur in between the 2^{nd} and 3^{rd} mode of vibration. A total of 1,200 samples, each with a sample length of 25,600, were collected from the CP under different operation conditions.

TABLE 1. Specifications for data collection.

Accelerometer (622b01)	**Frequency range:** 0.42 to 10kHz **Sensitivity:** 100 mV/g (10.2mV/(m/s^2)) ± 5 %
DAQ System (NI 9234)	**Frequency range:** 0 to 13.1MHz **Generator:** 4 analog input channels 24 bits ADC resolution

In the present study, the pump was operated under normal and different simulated fault conditions. The simulated faults are:

(A) Mechanical seal fault.
  (i) Mechanical seal hole.
  (ii) Mechanical seal scratch.

(B) Impeller fault.

These faults are simulated one at a time and the signals were obtained. The measurement noises were calculated for the obtained signals in each condition with respect to healthy baseline vibration signal. The measurement noises for mechanical seal hole, mechanical seal scratch and impeller

fault vibration signals were found to be -69.10, -62.07, and -63.78 dB respectively.

A. MECHANICAL SEAL FAULT

The most common seal fault occurs because of excessive pressure. To avoid leakage from the pump during the installation, the rotating part of the mechanical seal is kept in contact with the stationary part by means of a spring or a combination of springs. These springs must be compressed by a predetermined amount of pressure. Whenever this pressure is exceeded it exerts excessive pressure on the mechanical seal faces. As a result, this may cause overheating and, in turn, the thin lubricating film of liquid in between the sealing faces converts into vapors. Dirt is one of the greatest enemies of the mechanical seal. Any trace of dirt trapped between the sealing faces can lead to holes, scratches, and can make the seal faces hard and brittle during operation because of the excessive pressure of the springs in the absence of a lubricating film. These kinds of premature seal failures are very dangerous and result in catastrophic failures of the pump. To avoid afflictions because of these premature seal failures, in this study, hole and scratch faults are seeded in the mechanical seal and vibration signals were taken.

1) MECHANICAL SEAL HOLE

A mechanical seal is made up of two parts: a rotating seal part and stationary seal part. In this study, two seals of a 38 mm inner diameter are used. In the two seals, the stationary part was defectless and in the rotating part, a hole was created, as shown in Fig. 4. the diameter of the hole was 2.8mm and the depth was 2.8 mm. This was used as a defective seal to study the weak incipient fault under a mechanical seal hole defect.

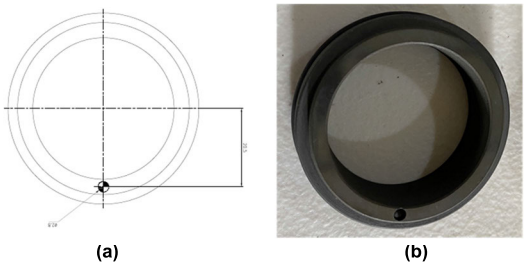

FIGURE 4. (a) Schematics of Mechanical seal having hole (b) Defective mechanical seal.

2) MECHANICAL SEAL SCRATCH

A scratch was created in the rotating part of the mechanical seal while the stationary part remained free from defects. Fig. 5 shows a mechanical seal having severe fault due to the scratch with a diameter of 2.5mm, length of 10 mm, and depth of 2.8 mm.

B. IMPELLER FAULT

Crevice corrosion is one of the common reasons for an impeller fault. Crevice corrosion results in an irregular surface with numerous holes of different sizes that overlap and look as if the surface of the impeller has been eaten away by insects. These holes of different sizes may become severe cracks due to the shear on the material and lead to fatigue which can result in catastrophic failure. In this paper, a similar fault was seeded on the impeller and vibration signals of the defective impeller were recorded.

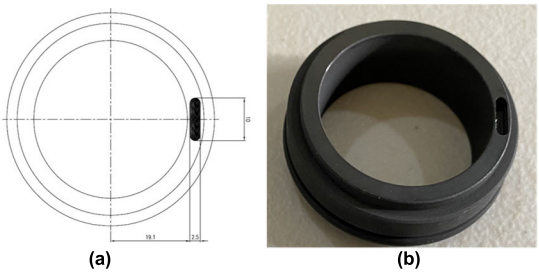

FIGURE 5. (a) Schematics of a mechanical seal with a scratch (b) Defective mechanical seal.

In this study three cast iron impellers, with 161 mm diameters, were used. Two impellers were new impellers without any defects. In the third impeller, a defect was created by removing some portion of the metal, as shown in Fig. 6. The diameters of the fault were 2.5 mm with a length of 18 mm and a depth of 2.8 mm. Fig. 7. shows the vibration signal obtained from the defective impeller while keeping all the other components in normal condition.

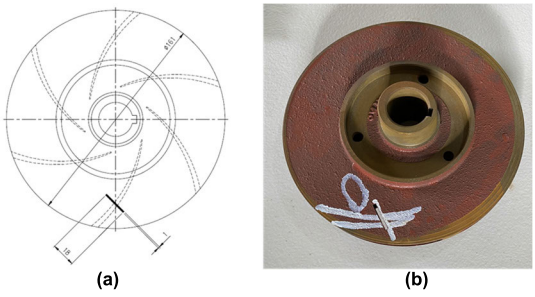

FIGURE 6. (a) Schematics of defective impeller (b) Defective impeller.

IV. PROPOSED METHOD

As explained in Section I, for discriminant feature extraction, both the vibration signal and feature preprocessing are important. Discriminant features help the subsequent classifier to classify the data effectively. This paper proposes a new method for CP fault diagnosis that preprocess both the vibration signal and raw statistical features based on the similarity. The workflow of the proposed method is presented in Fig. 8. As the proposed method is based on similarity, which means that the proposed method needs a baseline for comparison with the CP vibration signal and features of different classes. For this purpose, the proposed method can be divided into three phases. In the first phase, healthy baseline signals were selected. As the correlation measures the similarity in shape between the two signals, in the second phase, the vibration signal obtained from the CP under different operating

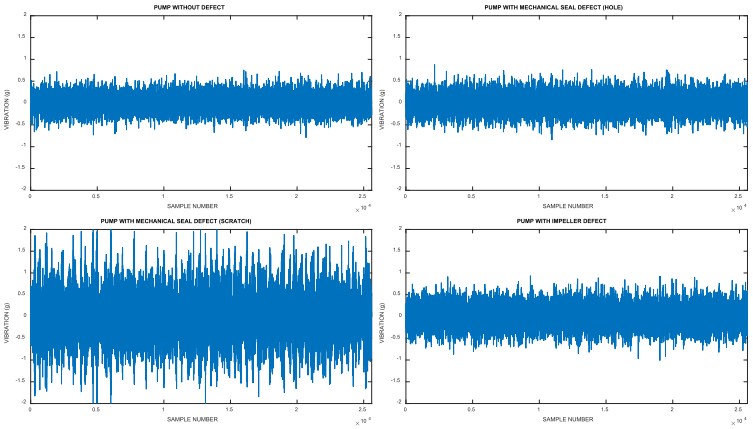

FIGURE 7. Time-domain plots of the CP operating under a normal and faulty condition (a) CP working under the normal condition (b) CP working under a defective seal (hole) (c) CP working under a defective seal (scratch) (d) CP working under a defective impeller.

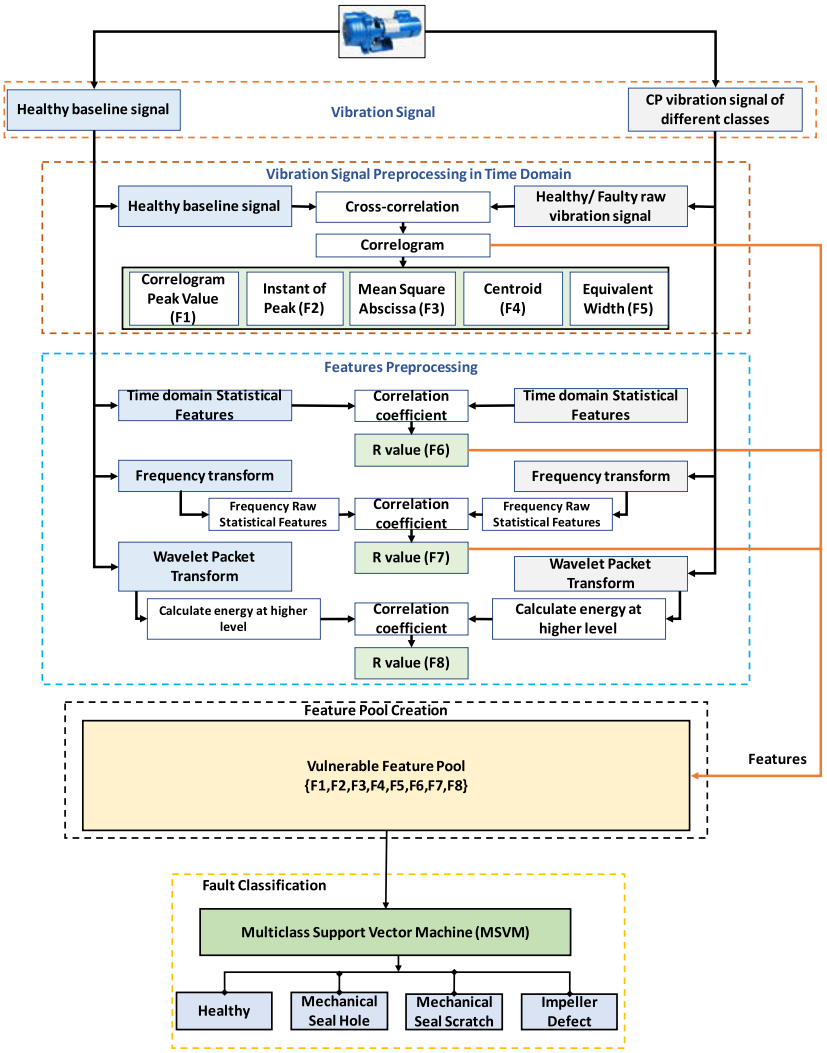

FIGURE 8. Flow diagram of the proposed method.

conditions is preprocessed by taking the cross-correlation between the healthy baseline signal and the CP vibration signal. The cross-correlation results in a correlogram signal in the time domain. The correlogram signal carries similarity information between the healthy baseline signal and the CP vibration signal. The proposed method extracts a new set of

discriminant features from the resultant correlogram. Feature preprocessing helps in overcoming the effects of likely fluctuations and random impulses in the vibration signals. In the third phase, raw statistical features are preprocessed. Raw statistical features are extracted from both the healthy baseline signal and the CP vibration signal in time, frequency, and the time-frequency domain. Raw statistical features are preprocessed by calculating the correlation coefficient between the healthy baseline signal features and the raw statistical features of the CP. The correlation coefficient results in a similarity measure. This similarity measure is a discriminant and is used as a new feature for CP fault diagnosis. After extracting the features, the proposed method combines all the features and creates a vulnerable feature pool. Furthermore, the vulnerable feature pool is provided as an input to the MSVM for CP fault segregation. The various steps involved in the proposed method are as follows:

PHASE I: SELECTION OF HEALTHY BASELINE SIGNAL

The selection of a healthy baseline signal plays a vital role in performing cross-correlation and calculating the correlation coefficient for the vibration signal and features preprocessing. Thus, in this study, a proper approach has been adopted for the selection of a healthy baseline signal. The steps involved in the healthy baseline signal selection are as follows:

1) STEP 1: FINDING PUMP PEAK EFFICIENCY POINT

The Pump peak efficiency point (PEP) is very important for the determination of the value of the flow rate and the NPSH at which the pump performs most efficiently. With a change in flowrate, the head, power consumption, and efficiency of the pump change, as well. Plotting these quantities against the flowrate gives the pump characteristic curve. The PEP is the measure that shows at which point in the characteristic curve the pump performs most efficiently. Fig. 9. shows the PMT-4008 pump characteristic curve with the PEP.

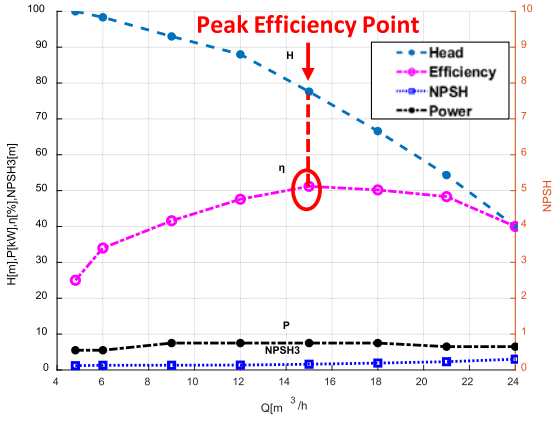

FIGURE 9. PMT-4008 pump characteristic curve.

The efficiency was calculated after measuring the head, power, and flow using Eq. 5.

$$\eta = \frac{Q\rho gH}{P} \tag{5}$$

where Q is the flowrate, g is the gravity, ρ is the density, H is the total head, and P is the power consumption.

After finding the PEP, the pump was operated at the PEP and a total of twenty PEP healthy samples of vibration signals were obtained.

2) STEP 2: HEALTHY BASELINE SIGNAL SELECTION

Selecting a healthy baseline signal randomly from the healthy samples of vibration signals collected at the pump PEP does not ensure the accuracy. After careful examination, we note that a signal among the PEP healthy samples of vibration signals obtained at the pump PEP, whose sample mean is near to the sample mean of the pump healthy vibration signals but farthermost from the sample mean of the other classes (mechanical seal hole, mechanical seal scratch, impeller fault), is selected as a healthy baseline signal. The healthy baseline signal obtained after this condition results in better accuracy while a random selection of a healthy baseline signal degrades the classification accuracy. It is to be noted that the pump healthy vibration signals are those samples which were collected from the pump under the normal operation of the pump. Fig. 10. shows the selected healthy baseline signal.

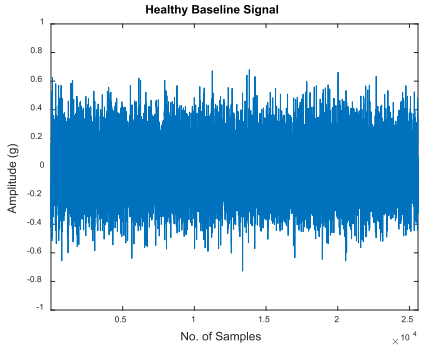

FIGURE 10. Healthy baseline signal.

PHASE II: PREPROCESSING CP VIBRATION SIGNAL IN TIME DOMAIN AND DISCRIMINANT FEATURES EXTRACTION

After selecting the healthy baseline signal, the healthy baseline signal is correlated with the CP vibration signals of different classes. The different classes of vibration signals collected from the CP are normal, mechanical seal hole fault, mechanical seal scratch fault, and impeller fault, as discussed in Section III. The cross-correlation between the healthy baseline signal and the CP vibration signals result in a correlogram which carries similarity information, as shown in Fig.11. Therefore, a new set of discriminant features are extracted from the correlogram. The features are peak correlogram value, instant of peak occurrence, correlogram

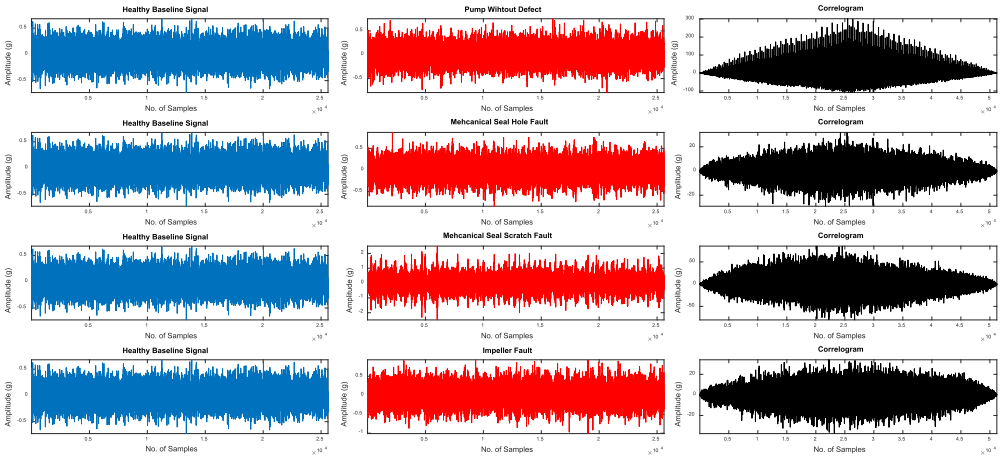

FIGURE 11. Correlogram sequences obtained after cross-correlating healthy baseline signals with CP vibration signals of different classes.

centroid, correlogram equivalent width and correlogram mean square abscissa (MSA). The peak and the time-instant of the peak can be easily extracted from the correlogram. The other three features can be computed using Eqs. 6, 7 and 8:

Let the correlogram sequence be represented by $R(m)$.

$$\text{Correlogram centroid} = \frac{\sum_{m=-M}^{M} mR(m)}{\sum_{m=-M}^{M} R(m)} \tag{6}$$

$$\text{Equivalent width} = \frac{\sum_{m=-M}^{M} mR(m)}{\text{Peak value of } R(m)} \tag{7}$$

$$\text{MSA} = \frac{\sum_{m=-M}^{M} m^2R(m)}{\sum_{m=-M}^{M} R(m)} \tag{8}$$

PHASE II: RAW STATISTICAL FEATURES PREPROCESSING FOR DISCRIMINANT FEATURES EXTRACTION

After extracting discriminant features from the correlogram sequence, in this phase, raw statistical features obtained from the vibration signal in time, frequency, and the time-frequency domain are preprocessed based on similarity and discriminant information and are extracted from the raw features in each domain. Feature preprocessing is important because it helps in overcoming the likely effect of random impulses and fluctuations in the vibration signal. The steps involved in features preprocessing are as follows.

1) STEP 1: RAW STATISTICAL FEATURES EXTRACTION IN THE TIME DOMAIN

In this step, the time-domain raw statistical features are extracted from both the healthy baseline signal and the CP vibration signals of different classes. The time domain statistical features adopted in this study are mentioned in [31]. The representative set of statistical features that are extracted from both the healthy baseline signal and CP vibration signals of different classes are the root amplitude, mean value, peak, root mean square, skewness, kurtosis value, crest factor, clearance factor, shape factor, and impulse factor. These statistical features are listed in Table. 2, along with the mathematical formulas.

2) STEP 2: TIME-DOMAIN RAW STATISTICAL FEATURES PREPROCESSING AND DISCRIMINANT INFORMATION EXTRACTION

After extracting time-domain features from both the healthy baseline and CP vibration signal of different classes, these features are combined into different vectors. These vectors are the healthy baseline features vector, CP healthy features vector, CP mechanical seal hole fault features vector, CP mechanical seal scratch fault features vector, and CP impeller fault features vector. The correlations between healthy baseline features and each CP feature vector are calculated using Eq. 9 and used as a new discriminant feature, as shown in Fig.8.

$$r_{fg} = \frac{\sum_{j=1}^{k} (f_i - \overline{f})(g_i - \overline{g})}{\sqrt{\sum_{j=1}^{k} (f_i - \overline{f})^2} \sqrt{\sum_{j=1}^{k} (g_i - \overline{g})^2}} \tag{9}$$

Eq. 9. is the Pearson correlation coefficient equation, which inspects the correlation between two variables f and g. In this equation r_{fg} is the correlation between f and g. After calculating the correlation coefficient between the healthy baseline features vector and CP feature vectors of different classes, r_{fg} is used as a new feature for CP fault diagnosis.

3) STEP 3: RAW STATISTICAL FEATURES EXTRACTION IN THE FREQUENCY DOMAIN

A fault in the CP component adds a weak fault signature to the vibration signal of the CP that can be explored using the frequency spectrum. Fault symptoms identification in the frequency spectrum helps in the detection of faulty components.

TABLE 2. Set of statistical features in time-domain.

Feature	Equation	Feature	Equation	Feature	Equation
Mean	$X_m = \frac{\sum_{n=1}^{N} x(n)}{N}$	Root amplitude	$X_{root} = \left[\frac{\sum_{n=1}^{N} \sqrt{\lvert x(n)\rvert}}{N}\right]^2$	Peak	$X_{peak} = \max \lvert x(n)\rvert$
Standard deviation	$X_{sd} = \sqrt{\frac{\sum_{n=1}^{N} (x(n)-X_m)^2}{N-1}}$	Root mean square	$X_{rms} = \sqrt{\frac{\sum_{n=1}^{N} (x(n))^2}{N}}$	Skewness	$X_{sk.} = \frac{\sum_{n=1}^{N} (x(n)-X_m)^2}{(N-1)X_{sd}^3}$
Kurtosis	$X_{kurtosis} = \frac{\sum_{n=1}^{N} (x(n)-X_m)^4}{(N-1)X_{sd}^4}$	Crest factor	$X_{crest} = \frac{X_{peak}}{X_{rms}}$	Clearance factor	$X_{clearance} = \frac{X_{peak}}{X_{root}}$
Shape factor	$X_{shape} = \frac{X_{rms}}{\frac{1}{N}\sum_{n=1}^{N} \lvert x(n)\rvert}$	Impulse factor	$X_{impulse} = \frac{X_{peak}}{\frac{1}{N}\sum_{n=1}^{N} \lvert x(n)\rvert}$		

Several experimental attempts have been made in the past for CP fault frequencies identification. However, mathematical proofs are not provided for such frequencies. Without proper mathematical equations, it is very hard to utilize the fault frequencies identification methods based on experimental results in industry, as well as in research.

Before extracting features from the frequency spectrum, in this study, the defect frequencies will first be theoretically calculated for the CP mechanical seal fault and impeller fault.

There are three sources of frequencies in machines. These sources are generated frequencies, excitation frequencies, and electronic frequencies. In this study, the generated and excitation frequencies are considered. The generated frequencies are the frequencies generated by the CP. In this study, concerned generated frequencies are imbalance. These frequencies can be easily explored in the spectrum if the speed and geometry of the machine is known.

For an impeller fault the source frequencies are the generated frequencies. whenever an impeller fault occurs in the CP, an impeller imbalance usually appears in the vibration signature [7], [47]. Eq. 10. shows the formulation for an impeller imbalance frequency.

$$IF = n \times Z \tag{10}$$

where $n = 1, 2, 3..$ represent the number of harmonics and Z is the pump speed (in Hz). Fig.12(a) shows the frequency spectrum of the CP at the pump PEP. Fig. 12(b) shows the frequency spectrum of a CP with an impeller defect. *IF* values were calculated up to the 5th harmonic because this was relevant for impeller fault detection. As observed from Fig. 12(b) the 3rd, 4th, and 5th harmonic amplitude increased because of the impeller defect.

Excitation frequencies, also referred to as natural frequencies, are a property of the system. An excitation frequency is a single frequency present in the spectrum that represents amplified vibration of the system. When a CP is operated under a mechanical seal defect, this excitation frequency then represents the mechanical seal defect in the spectrum.

For a mechanical seal defect, the source frequency is the excitation frequency. It can be calculated from the theory of 'vibration of a circular ring', as described in [48].

Assume, that r represents the radius of the centerline of the ring and u shows the radial displacement and A is the cross-sectional area of the ring.

In circumferential direction, the unit elongation of the ring is given by $\frac{u}{r}$. The potential energy of deformation is given by

$$PE = \left(\frac{AEu^2}{2r^2}\right) 2\pi r \tag{11}$$

In Eq.11, E is the modulus of elasticity.

The kinetic energy of the vibration is given by,

$$KE = \left(\frac{\rho A}{2}\right) \left((u')^2\right) 2\pi r \tag{12}$$

In Eq.12, ρ is the material density.

By applying the conservation of energy method,

$$\frac{d(PE + KE)}{dt} = 0 \tag{13}$$

As a result, equation of motion will be,

$$(u'') + \left(\frac{E}{\rho r^2}\right) u = 0 \tag{14}$$

Eq.14. can be written as,

$$(u'') + \left(\omega_n^2\right) u = 0 \tag{15}$$

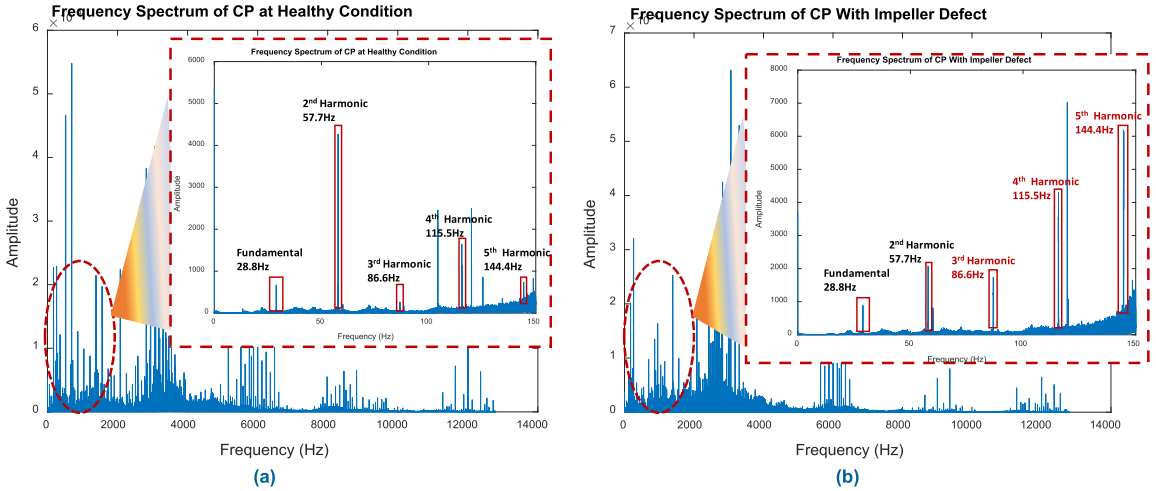

FIGURE 12. The frequency spectrum of the CP (a) Frequency spectrum of the CP in the healthy condition (b) Frequency spectrum of the CP with an impeller defect.

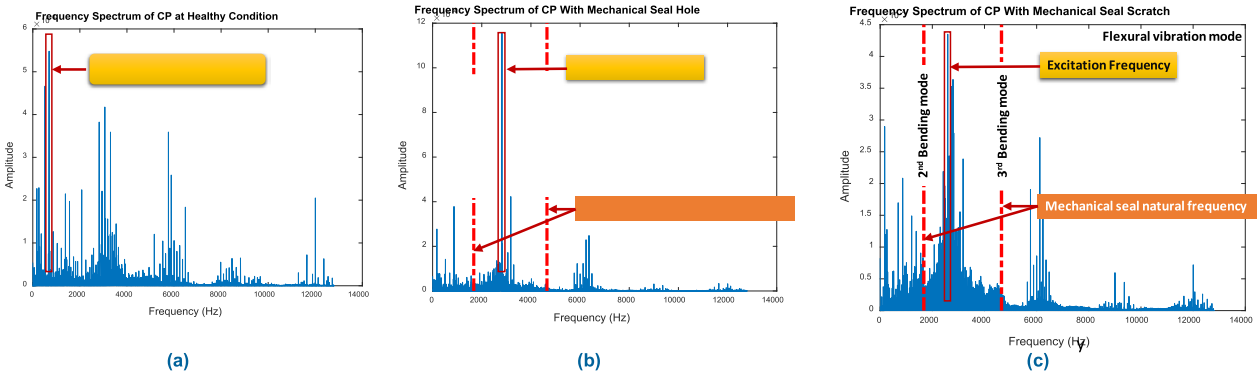

FIGURE 13. Frequency spectra of CPs (a) Frequency spectrum of a CP in the healthy condition (b) Frequency spectrum of a CP with a mechanical seal hole defect (c) Frequency spectrum of a CP with a mechanical seal hole defect (the red dotted lines show mechanical seal natural frequencies in out-of-plane bending modes).

where ω_n is the angular frequency. By solving Eq.15, we can obtain the ring fundamental frequency,

$$f_r = \left(\frac{1}{2\pi r}\right)\sqrt{\frac{E}{\rho}} \tag{16}$$

As vibration is not random, there are proper modes of vibration.

In the case of a mechanical seal, the torsional vibration mode, or in-plane bending mode, can be defined as the vibration mode in which the centerline of the ring remains undeformed and all the cross sections rotate during vibration at the same angle. The natural frequency and its corresponding modes of vibration can be determined by Eq.17.

$$f_0 = \left(\frac{2n\left(n^2-1\right)}{\pi}\right)\left(\frac{h}{d^2}\right) * \sqrt{\frac{E}{\rho\left(12n^2\right)+\left(\frac{2th^3(1+\nu)}{c}\right)}} \tag{17}$$

where n = 1,2,3,4,5 represent the modes of vibration, h is ring cross-sectional height, d is the diameter of the ring, t is the ring cross-sectional thickness, ν is Poisson's ratio and c is the torsion constant in Eq.17.

For the mechanical seal, the flexural vibration can be defined by Eq.18,

$$f_0 = \left(\frac{n\left(n^2-1\right)}{\pi d^2\sqrt{n^2+1}}\right)\sqrt{\frac{Et^2}{3\rho}} \tag{18}$$

where n = 1,2,3,4,5 represent the modes of vibration, d is the diameter of the ring, and t is the ring cross-sectional thickness in Eq.17.

The mechanical seal fundamental, in-plane and out-of-plane vibrations and their consecutive modes can be calculated using Eqs.16, 17, and 18. It is noted that the fundamental and in-plane vibrations usually have high frequencies and much lower frequencies will be found if out-of-plane vibrations of the ring are considered [48].

Next, the mechanical seal fundamental, in-plane, and out-of-plane vibrations and their consecutive modes are calculated using Eqs. 16, 17, and 18. It can be seen from Fig. 13(a), Fig. 13(b), and Fig. 13(c), that the system excitation

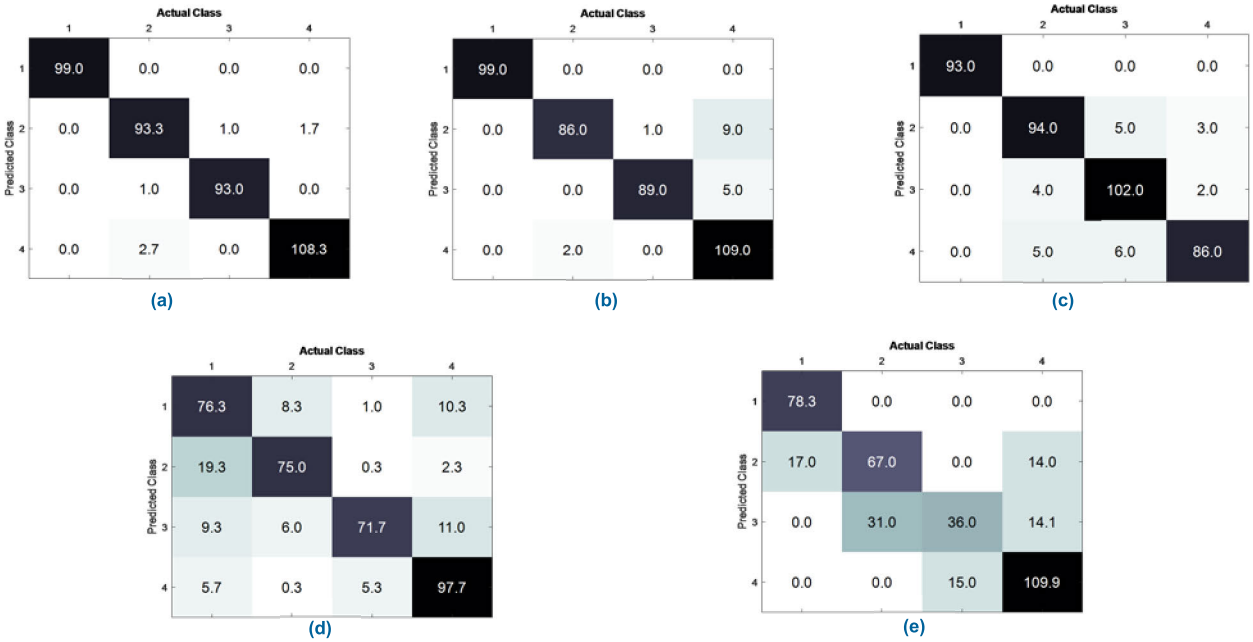

FIGURE 14. Confusion matrices (a) Proposed method (b) WPT-PCA-MSVM (c) WPT-BE-MSVM (d) J. s. Rapur *et al.* [9] (d) M. R. Nasiri *et al.* [36]. Label 1 for normal condition, 2 for mechanical seal hole, 3 for mechanical seal scratch, and 4 for impeller fault. All these results are illustrated as the average of 20 trails.

TABLE 3. Set of statistical features in frequency-domain.

Feature	Equation	Feature	Equation	Feature	Equation
Mean Frequency	$X_{mf} = \frac{\sum_{k=1}^{K} s(k)}{K}$	Root variance frequency	$Xrvf = \sqrt{\frac{\sum_{k=1}^{K} (s(k)-Xmf)^2}{K}}$	Root mean square frequency	$X_{frms} = \sqrt{\frac{\sum_{k=1}^{K} (s(k))^2}{K}}$
Standard deviation	$\sigma_f^2 = \frac{\sum_{k=1}^{K} \left(s(k)-X_{mf}\right)^2}{K-1}$	Spectral Kurtosis	$X_{fkurtosis} = \frac{\sum_{k=1}^{K} (s(k)-X_{mf})^4}{(K)\sigma_f^4}$		

frequency appears in between the 2nd and 3rd out-of-plane modes of vibration, with almost twice the amplitude when the CP is operated under mechanical seal defect conditions.

As already mentioned, the CP faults can be explored via frequency spectrum analysis, as shown in Fig. 13 and 14. Raw statistical information can be extracted from the frequency domain, as it provides enough discriminant information. Therefore, in this phase, raw statistical frequency domain features are extracted from the healthy baseline signal and CP vibration signals of different classes. The raw statistical frequency domain features considered in this study are presented in Table. 3.

4) STEP 4: FREQUENCY DOMAIN RAW STATISTICAL FEATURES PREPROCESSING AND DISCRIMINANT INFORMATION EXTRACTION

The next step follows the extraction of frequency domain features from the both the healthy baseline and CP vibration signals of different classes. Next, these features are combined into different vectors, namely, a healthy baseline frequency features vector, CP healthy frequency features vector, CP mechanical seal hole fault frequency features vector, CP mechanical seal scratch fault frequency features vector, and CP impeller fault frequency features vector. The correlations between the healthy baseline frequency features vector and each CP frequency features vector are calculated using Eq.9 and used as a new discriminant feature, as shown in Fig. 8.

5) STEP 5: FEATURES EXTRACTION IN TIME-FREQUENCY DOMAIN

Existing methods based on the WPT for CP fault diagnosis consider the energy, standard deviation, and entropy as input features to the classification model. Among these features, wavelet packet energy is an intuitive approach for fault types classification. Each node in the WPT contains a great

deal of information about the energy variations and fault types in a specific node can be helpful in segregating fault types.

In this study, both the healthy baseline signal and CP vibration signals of different classes are decomposed up to $k = 3$ levels. This results in 2^k nodes. After decomposition, the WPT energy is calculated using Eq 19.

$$Energy = \left(\frac{\sum_{p=1}^{S} \sqrt{\left(C_k^j(p)^2 \right)}}{S} \right) \tag{19}$$

where S is the number of samples at the node, k is level of decomposition, nodes (frequency parameters) j at WPT coefficient c in Eq. 19 as explained in [35].

6) STEP 6: TIME-FREQUENCY DOMAIN FEATURES PREPROCESSING AND DISCRIMINANT INFORMATION EXTRACTION

The next step follows the extraction wavelet packet energy features from both the healthy baseline and CP vibration signals of different classes. These features are combined into different vectors, namely, a healthy baseline energy features vector, CP healthy energy features vector, CP mechanical seal hole fault energy features vector, CP mechanical seal scratch fault energy features vector, and CP impeller fault energy features vector. The correlations between healthy baseline energy features and each CP energy features vector are calculated using Eq. 9 and are used as a new discriminant feature, as shown in Fig. 8.

PHASE III: VULNERABLE FEATURE POOL CREATION

A vulnerable feature pool is formed by combining the extracted discriminant features: after preprocessing the vibration signal in the time domain, discriminant features are obtained after preprocessing raw statistical time-domain features, frequency-domain features, and after preprocessing WPT energy features. This feature pool is named vulnerable because the discriminant feature obtained after preprocessing the vibration signal and features are based on similarity. The more the feature or signal is like a healthy baseline feature or signal the less vulnerable. Similarly, the less the feature or signal is similar to the healthy baseline feature or signal the more vulnerable. The vulnerable feature pool provides detailed and particular information about the complex nonstationary and nonlinear signals obtained from the CP with multiple fault severities. The dimension of the vulnerable feature pool is $5 + 1 + 1 + 1 = 8$. After constructing the vulnerable feature pool, it is provided as an input to the MSVM for CP fault segregation. This procedure is explained in the next section.

V. RESULTS AND DISCUSSION

The effectiveness of the proposed CP fault diagnosis method is presented in this section.

A. DATA CONFIGURATION FOR TRAINING AND TESTING

An appropriate dataset configuration for testing and training is very important in order to determine the comprehensive quality of the proposed CP diagnosis method. In this study, CP MFs were simulated. In total, 3 different faults were simulated in the CP, namely, a mechanical seal hole fault, mechanical seal scratch fault, and impeller fault. The fault severities vary from weak incipient faults to severe faults in the CP. A 300 sec long signal was acquired from the CP under each normal and faulty condition. Therefore, the created dataset is comprised of 1,200 signals in total. In this study, the features extracted from the dataset consist of $CP_c \times CP_s \times CP_f$ features in total. Where CP_c is the CP conditions simulated in this paper, CP_s is the signal instances for each simulated condition, and CP_f is the total number of extracted features.

In this study, a k-fold cross-validation of k = 3 strategy was adopted for validating the proposed method during each experimental trail. In the k-fold cross-validation strategy, the whole dataset is divided randomly into k-folds, where each fold must be used once as a testing subset to a classifier trained on the remaining $k - 1$ subsets. In this study specifically, 200 randomly chosen samples from each class were used as a training subset and the remaining 100 samples from each class were used to construct the testing subset. Thus, in total, each training set contains 800 samples and each testing set contains 400 remaining samples for classification.

B. PERFORMANCE EVALUATION OF PROPOSED CP FAULT DIAGNOSIS METHOD

To evaluate the quality of the new features obtained from preprocessing the raw vibration signal and raw statistical features for CP fault diagnosis, this study compares the proposed method of feature extraction with two time-frequency domain feature extraction methods, one time-domain feature extraction method, and one hybrid feature extraction method. The first time-frequency domain method (WPT-PCA-MSVM) used a WPT for preprocessing the time-domain vibration signal and utilized Principal Component Analysis (PCA) to select the WPT bases for statistical feature extraction [5]. The second time-frequency domain method (WPT-BE-MSVM) used a WPT for preprocessing the time-domain vibration signal and utilized the best energy criteria to select the best bases for feature extraction [5]. In the third time-domain feature extraction method, statistical features are extracted from the vibration signal for CP fault diagnosis [9]. In the fourth method, hybrid features in time and frequency domain are extracted from the vibration signal for CP fault diagnosis [36]. In this study, the MSVM classifier is used to perform the comparison between the proposed method and the above existing methods. To ensure repeatability in the results and to overcome the effects of randomness, the experiments are performed 20 times with random combinations of testing and training data.

TABLE 4. Experimental results. comparisons between the proposed method and reference methods (true positive rate(TPR), average classification accuracy (ACA)).

Methods	TPR in Average (%)				ACA (%)
	Normal Condition	Mechanical seal defect (hole)	Mechanical seal defect (scratch)	Impeller fault	
Proposed	100	96.2	98.9	98.4	98.4
WPT-PCA-MSVM	100	97.72	98.8	88.6	96.3
WPT-BE-MSVM	100	91.2	90.2	94.5	94
Rapur et al. [9]	68.9	83.6	91.5	80.4	81.1
Nasiri et al [36]	82.1	68.3	70.5	79.7	75.22

The classification accuracy for each class of samples is computed using Eq. 20.

$$TPR_m = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{(N_{TP}^{j,m})}{N_{TP}^{j,m}+N_{FN}^{j,m}}\right)\times 100(\%) \qquad (20)$$

Eq.20. is the True Positive Rate index (TPR), where k represents the number of cross-validation k folds, $(N_{TP}^{j,m})$ is the number of samples in each class m that are accurately segregated as class m, $N_{FN}^{j,m}$ is the number of samples in each class m that are not accurately recognized as class m, j indicates the iterations of the k-fold cross-validation. The final TPR is calculated for each class as an average of the TPR achieved after 20 trails.

The Classification Accuracy (CA) for each trail is calculated using Eq. 21.

$$CA = \frac{1}{k}\sum_{j=1}^{k}\left(\sum_{m=1}^{L}\frac{N_{TP}^{j,m}}{N_{samples}}\right)\times 100(\%) \qquad (21)$$

In Eq.21, L represents the total number of classes and $N_{samples}$ is the total number of samples in a specific testing subset. The final classification accuracy showed in the results is the Average Classification Accuracy (ACA) obtained after 20 trails.

The experimental results are presented in Table 4, Fig. 14 and Fig 15. The results reveal that the proposed correlation analysis based feature extraction method for CP fault diagnosis outperforms the existing methods in terms of the ACA, error rate (ER), precision with macro-averaging (P_M) and recall with macro-averaging (R_M) with an ACA of 98.4%, ER of 3%, P_M of 98.4% and R_M of 98.4%. ER shows the average per-class miss classification, the P_M shows the average per-class agreement of the class labels with those of the classifiers and the R_M shows the per class efficiency of classifiers for the identification of class labels [49]. Table 4 also reflects that the average TPR values of the proposed method are over 95% for each class. Fig. 14(a) shows that the proposed method accurately identified CP faults with a negligible misclassification rate as compared to the reference methods.

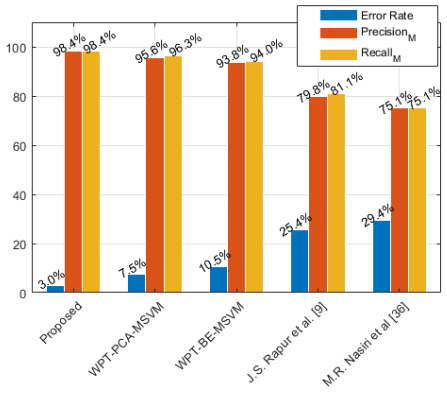

FIGURE 15. Evaluation of proposed method and reference methods for CP fault diagnosis.

The results can be explained as follows. Multiple faults with different levels of severity caused random fluctuations and impulses in the vibration signal. These random fluctuations and impulses in the vibration signal make the time or frequency domain analysis ineffective for such kinds of vibration signals. However, preprocessing the vibration signal can overcome the effect of random fluctuations and impulses in the vibration signal. On the other hand, raw statistical features obtained from the time and frequency domain are either not very sensitive to weak incipient faults or unsuitable for more severe faults and may result in less discriminate information. To overcome this issue, a proper preprocessing of raw statistical features is required, which can result in more discriminating information. This can be seen from results that are based on correlation preprocessing of the vibration signal as well as the hybrid features, which taken together results in discriminant features. Therefore, the proposed method outperforms the reference methods in terms of the ACA, ER, P_M and R_M for CP fault diagnosis. The underperforming reference methods for CP fault diagnosis result in ACAs of 96.3%, 94%, 81.1% and 75.22%, ER of 7.5%, 10.5%, 25.4% and 29.4%, P_M of 95.6%, 93.8%, 79.8% and 75.1% and R_M

of 96.3%, 94%,81.1% and 75.1% for the methods of WPT-PCA-MSVM, WPT-BE-MSVM, [9], and [36], respectively.

We now focus on WPT-PCA-MSVM, which preprocesses the signal using WPT. The signal is decomposed up to 2 levels having 4 bases. The number of bases is the dimension of the data. PCA assumes that the larger the variation in the data the more important the dynamics attached to it. Therefore, the principal components are ordered according to decreasing covariance. The first two WPT bases are selected as these bases correspond to more than 70% of data covariance and the rest of the bases are discarded. After selecting the WPT bases, statistical features are extracted from them and using a wrapper model, the best statistical features are selected among the statistical features pool. After applying this model for feature extraction and classification to our experimental setup, we obtain ACAs of 96.3%, P_M of 95.6% and R_M of 96.3% which is less than the ACA of 98.4%, P_M of 98.4% and R_M of 98.4% for our proposed method as can be seen from Table 4 and Fig 15. This is because of the information loss due to PCA. Also, the variance in the WPT depends on the level of signal decomposition and the optimal mother wavelet used for wavelet transformation. On the other hand, the proposed method utilizes all the bases of the WPT, along with new discriminant features obtained from hybrid features preprocessing and vibration signal preprocessing in the time domain. However, it is observed that the WPT-PCA-MSVM is sensitive to weak faults, such as a mechanical seal hole defect. As can be seen in Table 4, and Fig. 14(a) and (b) the average TPR of this reference method is slightly higher than that of the proposed method in the case of a mechanical seal hole defect.

Next, we discuss WPT-BE-MSVM, which preprocess the signal using WPTs. The signal is decomposed up to 2 levels having 4 bases. The energy is calculated for each of the bases and bases are arranged in descending order according to energy. After arranging the WPT bases, statistical features are extracted from each of the bases and classification is performed. Based on the individual base classification accuracy, the best energy bases are selected among the 4 WPT bases. The statistical features are extracted from the best energy bases and by using the wrapper model, the best statistical features are selected among the statistical features pool. After applying this model for feature extraction and classification to our experimental setup we obtain ACAs of 94%, P_M of 93.8%, and R_M of 94%, which is less than the ACA of 98.4%, P_M of 98.4% and R_M of 98.4% provided by our proposed method. This is because the quality of the WPT strongly depends on the choice of the mother wavelet. Also, it is observed that discarding bases of the WPT on energy criteria can result in the loss of information about certain faults. On the other hand, the proposed method utilizes all the bases of the WPT along with new discriminant features obtained from hybrid features preprocessing and vibration signal preprocessing in the time domain. However, it is observed that the WPT-BE-MSVM model is sensitive to faults which causes severe vibrations, such as an impeller defect. As can be seen in Table 4 and Fig. 14(a) and (c), the average TPR of the reference method is slightly less than that of the proposed method in the case of impeller defect.

Rapur and Tiwari [9], extract time-domain statistical features for CP fault diagnosis using SVMs. After extracting the same time domain features and repeating the analysis with our experimental data, we obtained an ACA of 81.1%, P_M of 79.8% and R_M of 81.1, which is less than the ACA of 98.4%, P_M of 98.4% and R_M of 98.4% achieved by our proposed method. This is because the raw vibration signal obtained for the CP contains random fluctuations caused by the fault. The features obtained from the raw vibration signal without preprocessing the raw vibration signal result in less discriminant information, which can be seen in Table 4 and Fig 14 (a) and (d).

Regarding Nasiri *et al.* [36], for cavitation detection in CP, he extracted two features in time and two features in the frequency domain from the vibration signal. The hybrid feature pool is used as an input to the neural network for the detection of cavitation in the CP. To make the comparison fair we used the MSVM for classification instead of a neural network. After using his feature extraction method, we obtained ACA of 75.22%, P_M of 75.1% and R_M of 75.1% which is less than the ACA of 98.4%, P_M of 98.4% and R_M of 98.4% obtained by our proposed method. This is because, the dimension of the feature vector is very small, and the features are not sufficiently discriminant. The miss classification of [36] can be seen from Fig 14 (e). Although he used frequency domain features, the frequency harmonics of CP MFs show only a minute change in the amplitude from the healthy signal, as shown in Fig. 12. This method cannot be used for classification when it uses only two statistical features. Furthermore, the ER of the proposed method is less than that of the reference methods as can be seen from Fig. 15.

The MFs in the CP can result in to two main kinds of failure phenomena's namely hard failure and soft failure. Hard failure can be detected by simple analysis of the CP however soft failure causes CP performance degradation, but the CP remains in function [50]. MFs such as mechanical seal hole, scratch and impeller crack may result in soft failures of the CP. Soft failure in CP are not obvious enough to be diagnosed without a well-designed diagnosis technique. Our proposed method will help the decision makers to address these soft failures. Thus, after detecting and classifying the fault by our proposed method, the decision maker will only replace that specific component in the CP and bring the CP back to operation. Through this way the CP will remain healthy, take less time and minimum cost for maintenance. It is highly recommended for the decision makers to follow the guidelines provided by the pump manufacturers for replacing components in the CP. After replacing the components in CP, the net positive suction head and discharge head of the CP must be compared with the one provided by the manufacturers.

Overall, the proposed method of discriminant feature extraction is very useful for the diagnosis of CP faults with varying severities. The effectiveness is explained through the proposed method's main idea: preprocessing the vibration

signal and raw hybrid statistical features based on their correlation/or similarity. The proposed method is very easy to implement with less computational complexity. These properties of the proposed method make it worthy for industrial purposes.

VI. CONCLUSION

This paper proposed a new discriminant feature extraction method for CP fault diagnosis based on correlations. The proposed method is divided into three phases. In the first phase, healthy baseline signals were selected. In the second phase, vibration signals are preprocessed in the time domain and a set of new features are extracted from the preprocessed vibration signal. In the third phase, raw statistical features are preprocessed by calculating the correlation coefficients between the healthy baseline signal features and the CP raw statistical features. Each correlation coefficient results in a similarity measure. This similarity measure is discriminant and is used as a new feature for CP fault diagnosis. After extracting the features, the proposed method combines all of the features into a single feature vector for each class and creates a vulnerable feature pool. In the experimental part of this study, the vulnerable feature pool was used to segregate different CP MF's using the MSVM classifier. The experimental results shows that the proposed method of the new discriminant feature extraction outperforms state-of-the-art reference methods with an ACA of 98.4%. however, the proposed method is not tested for the classification of CP hydraulic faults. For future work, the proposed method will be tested for the classification of CP hydraulic faults.

REFERENCES

[1] R. J. Hart, "Pumps and their systems-a changing industry," in *Proc. 19th Int. Pump Users Symp.*, 2002, pp. 141–144.

[2] J. Kléma, O. Flek, J. Kout, and L. Nováková, "Intelligent diagnosis and learning in centrifugal pumps," in *Emerging Solutions for Future Manufacturing Systems* (IFIP International Federation for Information Processing), vol. 159. Friesach, Austria: Springer, 2005, pp. 513–522, doi: 10.1007/0-387-22829-2_56.

[3] H. Saranga and J. Knezevic, "Reliability prediction for condition-based maintained systems," *Reliab. Eng. Syst. Saf.*, vol. 71, no. 2, pp. 219–224, 2001, doi: 10.1016/S0951-8320(00)00094-6.

[4] R. B. Randall, Ed., *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications*, 1st ed. Hoboken, NJ, USA: Wiley, 2011.

[5] J. S. Rapur and R. Tiwari, "Experimental fault diagnosis for known and unseen operating conditions of centrifugal pumps using MSVM and WPT based analyses," *Measurement*, vol. 147, Dec. 2019, Art. no. 106809, doi: 10.1016/j.measurement.2019.07.037.

[6] S. M. Chittora, "Monitoring of mechanical seals in process pumps," KTH Roy. Inst. Technol., Stockholm, Sweden, Tech. Rep. 2018:686, 2018.

[7] K. McKee, G. Forbes M. I. Mazhar, R. Entwistle, and I. Howard, "A review of major centrifugal pump failure modes with application to the water supply and sewerage industries," in *Proc. ICOMS Asset Manage. Conf.*, Gold Coast, QLD, Australia, 2011, pp. 1–12.

[8] E. Ebrahimi and M. Javidan, "Vibration-based classification of centrifugal pumps using support vector machine and discrete wavelet transform," *J. Vibroeng.*, vol. 19, no. 4, pp. 2586–2597, Jun. 2017, doi: 10.21595/jve.2017.18120.

[9] J. S. Rapur and R. Tiwari, "Experimental time-domain vibration-based fault diagnosis of centrifugal pumps using support vector machine," *ASCE-ASME J Risk Uncertainty Eng. Syst. Part B, Mech. Eng.*, vol. 3, no. 4, pp. 1–7, Dec. 2017, doi: 10.1115/1.4035440.

[10] M. A. Abu-Zeid and S. M. Abdel-Rahman, "Bearing problems' effects on the dynamic performance of pumping stations," *Alexandria Eng. J.*, vol. 52, no. 3, pp. 241–248, Sep. 2013, doi: 10.1016/j.aej.2013.02.002.

[11] M. Delgado, G. Cirrincione, A. Garcia, J. A. Ortega, and H. Henao, "Accurate bearing faults classification based on statistical-time features, curvilinear component analysis and neural networks," in *Proc. IECON-38th Annu. Conf. IEEE Ind. Electron. Soc.*, Oct. 2012, pp. 3854–3861, doi: 10.1109/IECON.2012.6389596.

[12] B. Samanta and K. R. Al-Balushi, "Artificial neural network based fault diagnostics of rolling element bearings using time-domain features," *Mech. Syst. Signal Process.*, vol. 17, no. 2, pp. 317–328, Mar. 2003, doi: 10.1006/mssp.2001.1462.

[13] S. Wang, J. Xiang, Y. Zhong, and H. Tang, "A data indicator-based deep belief networks to detect multiple faults in axial piston pumps," *Mech. Syst. Signal Process.*, vol. 112, pp. 154–170, Nov. 2018, doi: 10.1016/j.ymssp.2018.04.038.

[14] L. Hong and J. S. Dhupia, "A time domain approach to diagnose gearbox fault based on measured vibration signals," *J. Sound Vibrat.*, vol. 333, no. 7, pp. 2164–2180, Mar. 2014, doi: 10.1016/j.jsv.2013.11.033.

[15] N. R. Sakthivel, V. Sugumaran, and S. Babudevasenapati, "Vibration based fault diagnosis of monoblock centrifugal pump using decision tree," *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4040–4049, Jun. 2010, doi: 10.1016/j.eswa.2009.10.002.

[16] J. S. Rapur and R. Tiwari, "Automation of multi-fault diagnosing of centrifugal pumps using multi-class support vector machine with vibration and motor current signals in frequency domain," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 40, no. 6, pp. 1–21, Jun. 2018, doi: 10.1007/s40430-018-1202-9.

[17] H. Sun, S. Yuan, and Y. Luo, "Characterization of cavitation and seal damage during pump operation by vibration and motor current signal spectra," *Proc. Inst. Mech. Eng., Part A, J. Power Energy*, vol. 233, no. 1, pp. 132–147, Feb. 2019, doi: 10.1177/0957650918769761.

[18] S. Li, N. Chu, P. Yan, D. Wu, and J. Antoni, "Cyclostationary approach to detect flow-induced effects on vibration signals from centrifugal pumps," *Mech. Syst. Signal Process.*, vol. 114, pp. 275–289, Jan. 2019, doi: 10.1016/j.ymssp.2018.05.027.

[19] M. A. S. ALTobi, G. Bevan, P. Wallace, D. Harrison, and K. P. Ramachandran, "Fault diagnosis of a centrifugal pump using MLP-GABP and SVM with CWT," *Eng. Sci. Technol., Int. J.*, vol. 22, no. 3, pp. 854–861, Jun. 2019, doi: 10.1016/j.jestch.2019.01.005.

[20] V. Muralidharan, V. Sugumaran, and V. Indira, "Fault diagnosis of monoblock centrifugal pump using SVM," *Eng. Sci. Technol., Int. J.*, vol. 17, no. 3, pp. 152–157, Sep. 2014, doi: 10.1016/j.jestch.2014.04.005.

[21] X. Tang, Z. Zhang, Q. Huang, and Y. Gong, "Fault location and fault type recognition of power system based on wavelet transform," in *Proc. IEEE Innov. Smart Grid Technol.-Asia (ISGT Asia)*, May 2019, pp. 689–692, doi: 10.1109/ISGT-Asia.2019.8881101.

[22] Y. Jin, C. Shan, Y. Wu, Y. Xia, Y. Zhang, and L. Zeng, "Fault diagnosis of hydraulic seal wear and internal leakage using wavelets and wavelet neural network," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 4, pp. 1026–1034, Apr. 2019, doi: 10.1109/TIM.2018.2863418.

[23] V. Muralidharan and V. Sugumaran, "Feature extraction using wavelets and classification through decision tree algorithm for fault diagnosis of mono-block centrifugal pump," *Measurement*, vol. 46, no. 1, pp. 353–359, Jan. 2013, doi: 10.1016/j.measurement.2012.07.007.

[24] N. Bessous, S. E. Zouzou, W. Bentrah, S. Sbaa, and M. Sahraoui, "Diagnosis of bearing defects in induction motors using discrete wavelet transform," *Int. J. Syst. Assurance Eng. Manage.*, vol. 9, no. 2, pp. 335–343, Apr. 2018, doi: 10.1007/s13198-016-0459-6.

[25] B. Bessam, A. Menacer, M. Boumehraz, and H. Cherif, "Wavelet transform and neural network techniques for inter-turn short circuit diagnosis and location in induction motor," *Int. J. Syst. Assurance Eng. Manage.*, vol. 8, no. S1, pp. 478–488, Jan. 2017, doi: 10.1007/s13198-015-0400-4.

[26] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998, doi: 10.1098/rspa.1998.0193.

[27] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mech. Syst. Signal Process.*, vol. 35, nos. 1–2, pp. 108–126, Feb. 2013, doi: 10.1016/j.ymssp.2012.09.015.

[28] S. Alabied, U. Haba, A. Daraz, F. Gu, and A. D. Ball, "Empirical mode decomposition of motor current signatures for centrifugal pump diagnostics," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2018, pp. 1–6, doi: 10.23919/IConAC.2018.8749109.
[29] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, Jan. 2009, doi: 10.1142/S1793536909000047.
[30] A. E. Prosvirin, M. M. M. Islam, and J.-M. Kim, "An improved algorithm for selecting IMF components in ensemble empirical mode decomposition for domain of rub-impact fault diagnosis," *IEEE Access*, vol. 7, pp. 121728–121741, 2019, doi: 10.1109/access.2019.2938367.
[31] Y. Lei, *Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery*. Oxford, U.K.: Butterworth-Heinemann, 2016.
[32] H. Helmi and A. Forouzantabar, "Rolling bearing fault detection of electric motor using time domain and frequency domain features extraction and ANFIS," *IET Electr. Power Appl.*, vol. 13, no. 5, pp. 662–669, May 2019, doi: 10.1049/iet-epa.2018.5274.
[33] T. Xiaodong, "Fault prognosis feature extraction and selection for bearings based on statistical indicator optimization," in *Proc. Prognostics Syst. Health Manage. Conf. (PHM-Qingdao)*, Oct. 2019, pp. 1–6, doi: 10.1109/PHM-Qingdao46334.2019.8942829.
[34] A. Prosvirin, M. Islam, J. Kim, and J.-M. Kim, "Rub-impact fault diagnosis using an effective IMF selection technique in ensemble empirical mode decomposition and hybrid feature models," *Sensors*, vol. 18, no. 7, p. 2040, Jun. 2018, doi: 10.3390/s18072040.
[35] M. Sohaib, C.-H. Kim, and J.-M. Kim, "A hybrid feature model and Deep-Learning-Based bearing fault diagnosis," *Sensors*, vol. 17, no. 12, p. 2876, Dec. 2017, doi: 10.3390/s17122876.
[36] M. R. Nasiri, M. J. Mahjoob, and H. Vahid-Alizadeh, "Vibration signature analysis for detecting cavitation in centrifugal pumps using neural networks," in *Proc. IEEE Int. Conf. Mechatronics*, Apr. 2011, pp. 632–635, doi: 10.1109/ICMECH.2011.5971192.
[37] D. G. Proakis and J. G. Manolakis, *Digital Signal Processing: Principles, Devices and Applications*, 3rd ed. Newdelhi, India: Prentice-Hall, 1990.
[38] S. Chandaka, A. Chatterjee, and S. Munshi, "Cross-correlation aided support vector machine classifier for classification of EEG signals," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1329–1336, Mar. 2009, doi: 10.1016/j.eswa.2007.11.017.
[39] R. Bose, K. Samanta, and S. Chatterjee, "Cross-correlation based feature extraction from EMG signals for classification of neuro-muscular diseases," in *Proc. Int. Conf. Intell. Control Power Instrum. (ICICPI)*, Oct. 2016, pp. 241–245, doi: 10.1109/ICICPI.2016.7859710.
[40] P. N. Paranjape, M. M. Dhabu, P. S. Deshpande, and A. M. Kekre, "Cross-correlation aided ensemble of classifiers for BCI oriented EEG study," *IEEE Access*, vol. 7, pp. 11985–11996, 2019, doi: 10.1109/ACCESS.2019.2892492.
[41] L. Shi, "Correlation coefficient of simplified neutrosophic sets for bearing fault diagnosis," *Shock Vibrat.*, vol. 2016, pp. 1–11, Dec. 2016, doi: 10.1155/2016/5414361.
[42] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, Aug. 2007, doi: 10.1016/j.ymssp.2006.12.007.
[43] A. Azadeh, A. Nourmohammadzadeh, A. Kazem, V. Ebrahimipour, Z. Saberi, and M. Saberi, "A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization," *Appl. Soft Comput.*, vol. 13, no. 3, pp. 1478–1485, Mar. 2013, doi: 10.1016/j.asoc.2012.06.020.
[44] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection," *Eng. Appl. Artif. Intell.*, vol. 16, nos. 7–8, pp. 657–665, Oct. 2003, doi: 10.1016/j.engappai.2003.09.006.
[45] S. H. Oppenheim, A. V. Willsky, and A. S. Nawab, *Signal & Systems*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.
[46] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.
[47] S. Yedidiah, *Centrifugal Pump User's Guidebook: Problems and Solutions*. Cham, Switzerland: Springer, 2012.
[48] D. H. Young, W. Weaver, and S. P. Timoshenko, *Vibration Problems in Engineering*, vol. 207, no. 2. Hoboken, NJ, USA: Wiley, 1929.
[49] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[50] D. Li, G. Hu, and C. J. Spanos, "A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis," *Energy Buildings*, vol. 128, pp. 519–529, Sep. 2016, doi: 10.1016/j.enbuild.2016.07.014.

ZAHOOR AHMAD received the B.S. degree in computer engineering from the Comsats Institute of Information Technology (CIIT), now called Comsats University Islamabad (CUI), Attock, Pakistan, in 2016, and the M.S. degree in electronics and information engineering from Korea Aerospace University (KAU), Goyang, South Korea. He is currently pursuing the Ph.D. degree in computer engineering with the University of Ulsan, South Korea.

He has been a Graduate Research Assistant with the Smart Health, Safety, and Environment (HSE) Laboratory, University of Ulsan, since 2019. He worked on the development of Advance algorithms for UAVs path planning. His current research interests include artificial intelligence, signal processing, fault diagnosis, vibration-based condition monitoring of industrial machinery, and fault feature extraction.

Mr. Ahmad received the Institutional Highest Prize Gold Medal in B.S. from CUI, Attock, Pakistan. He was awarded with a fully funded scholarship for M.S. degree from Korean Government project. His recent awards and honors include the National Research Foundation (NRF) of Korea Research Fellowship scholarship and the University of Ulsan President Excellence Scholarship for pursuing Ph.D. Program.

AKHAND RAI received the B.Tech. degree in mechanical engineering from the College of Engineering Roorkee, Roorkee, Uttarakhand, India, in 2010, the M.Tech. degree in mechanical engineering from the Indian Institute of Technology–Banaras Hindu University, Varanasi, Uttar Pradesh, India, in 2012, and the Ph.D. degree in mechanical engineering from the Indian Institute of Technology (IIT) Roorkee, Roorkee, in 2018. He worked as a Postdoctoral Professional Researcher with the School of Electrical, Electronics, and Computer Engineering, University of Ulsan, Ulsan, South Korea. He is currently working as an Assistant Professor with the School of Engineering and Applied Sciences, Ahmedabad University, India. His primary research interests include fault diagnosis and prognosis of rotating machines.

ANDREI S. MALIUK received the B.S. and M.S. degrees in electrical engineering from Moscow Polytechnic University, Moscow, Russia, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree in computer engineering with the University of Ulsan, South Korea.

Since 2019, he has been working as a Graduate Research Assistant with the Smart Health, Safety, and Environment (HSE) Laboratory, University of Ulsan. His research interests include fault diagnosis and condition monitoring of complex engineering systems using artificial intelligence, machine learning, and signal processing.

JONG-MYON KIM (Member, IEEE) received the B.S. degree in electrical engineering from Myongji University, Yongin, South Korea, in 1995, the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2000, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2005.

He is currently a Professor with the Department of IT Convergence, University of Ulsan, Ulsan, South Korea. His research interests include fault diagnosis and condition monitoring, multimedia-specific processor architecture, parallel processing, and embedded systems.

Dr. Kim is a member of the IEEE Industrial Electronics Society.