Research Article **Open Access**

Han Tran, Mohammad Noori*, Wael A. Altabey, and Xi Wu

Fault diagnosis of rotating machinery using wavelet-based feature extraction and support vector machine classifier

DOI 10.1515/hsm-2017-0003
Received April 20, 2017; accepted June 13, 2017

Abstract: Modern machine tools with high speed machining capabilities could place rotating shafts, gears, and bearings under extreme thermal, static, and impact stresses, potentially increasing their failure rates. In this research, a gearbox damage detection strategy based on discrete wavelet transform (DWT), wavelet packet transform (WPT), support vector machine (SVM), and artificial neural networks (ANN) is presented. Three case studies are conducted to compare the classification performance of SVM kernel functions and ANN. First, a fault detection analysis based on DWT and WPT is carried out to extract the damage information from the gearbox's raw vibration signal. In this step, wavelet coefficients obtained from DWT are characterized using statistical calculations. Energy characteristics of the gearbox signal are acquired using WPT and their statistical characteristics are also computed. These three sets of information extracted from wavelet transforms are utilized as the input to SVM and ANN classifiers. Secondly, the improved distance evaluation technique (IDE) is implemented to select the sensitive input features for SVM and ANN. The penalty parameter C and kernel parameter γ in SVM are also optimized using the grid-search method. Finally, the optimized features and parameters are input into SVM and ANN algorithms to detect gearbox damage. The result shows that gearbox damage detection using energy characteristics extracted from WPT (Case 2) or their statistical values as input features (Case 3) to the learning algorithms produces higher classification accuracies than using statistical values of the DWT coefficients as inputs (Case 1). In addition, RBF-SVM has the best classification performance in Case 2 and 3 while Linear-SVM has the best classification accuracy rate in Case 1 in damage detection average.

Keywords: Wavelet-based damage detection, support vector machine, gearbox damage detection

1 Introduction

Interest in the ability to monitor a structure and detect damage at the earliest possible stage is imperative throughout the fields of civil, mechanical, and aerospace engineering. This is particularly important for high speed machining technologies with its infrastructure and systems continuing to proliferate and increase in complexity. The demand for monitoring damages of structures increases as well to match throughput and ensure the safety of structures and the people within them. As infrastructure and systems continue to proliferate and increase in complexity, the demand for monitoring damages of structures increases as well to match throughput and ensure the safety of structures and the people within them. These damages are often invisible and can adversely influence the integrity of a system or structure and lead to catastrophic results. Early assessment of structural conditions is crucial to improve economical operations and reduce adverse effects from formations of damage, ensuring the structural stability of a system. Therefore, there is an ever-increasing demand for some way to identify these damages. This field is called Structural Health Monitoring (SHM).

***Corresponding Author: Mohammad Noori:** Professor of Mechanical Engineering and ASME Fellow, California Polytechnic State University, San Luis Obispo, California, USA; Visiting Affiliation at Southeast University, Nanjing, China; Email: mnoori@outlook.com
Han Tran: Graduate Student, Mechanical Engineering, California Polytechnic State University
Wael A. Altabey: International Institute for Urban Systems Engineering, Southeast University, Nanjing 210096, China; Department of Mechanical Engineering, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt
Xi Wu: Professor of Mechanical Engineering, California Polytechnic State University

1.1 Structural Health Monitoring

SHM is the discipline that manifests the ability to inspect engineering infrastructures based on the collection and analysis of real-time data accumulated from monitoring sensors installed on or within the structures. A SHM system consists of a network of sensors to collect the response data and data-mining algorithms to extract information pertaining to the structure conditions [1]. Dynamic input and response quantities are the most common measurements used for SHM. These quantities deliver sensitive information that reflects changes in a structure such as mass, stiffness, vibration response (damping), and energy dissipation. There are four levels of damage detections: 1) identifying the presence of damage in the system, 2) location of the damage, 3) identify the severity of the damage, and 4) predicting the remaining life of the structure. The work in this study is limited to level 1.

One application of SHM that has been the focus of much research in the past two decades is gearbox damage detection and identification. Gearboxes are often exposed to high stresses and compact loadings, and due to their assembly complexity and difficulty of access, they are significantly more challenging to maintain, repair, or replace than other systems. Therefore, applying SHM to gearboxes is desired. A note on nomenclature: when applied to rotating machinery and reciprocating machinery, SHM is referred to as Condition Monitoring (CM) [2]. Generally, SHM (or CM in this context) is a four-step process:

1. **Operational evaluation**. Setting limitations and justifications criteria on the "what" and "how" monitoring will be carried out.
2. **Data acquisition**. Selecting the excitation methods, sensors type, numbers and locations of sensors, data storage, and transmittal hardware.
3. **Feature selection**. Extracting the prominent (or sensitive) features to differentiate undamaged data from damaged data. This step involves reducing data condensation (*i.e.* dimension).
4. **Statistical modeling**. Establishing and implementing a statistical model to quantify the damage state of the structure based on the selected features in Step 3.

The most common classical approach in condition monitoring is through the implementation of vibration signals accumulated from the system's sensors and accelerometers. Informative characteristics of these vibration signals, also known as features, are extracted via advanced signal processing techniques. These processes are typically done in Step 2 and 3. It is desirable to classify the measured data of a structure to its states and determine whether or not the system is "healthy" or "damaged" autonomously. This is where the concept of machine learning comes into place with the basic platform of SHM. By establishing a learning algorithm into the problem, existing patterns from the extracted features can be classified. In machine learning, this is known as pattern recognition.

1.2 Machine Learning

Learning problems in pattern recognition naturally fall into two classes: supervised learning and unsupervised learning. In supervised learning, the model is provided with explicit examples that are correctly labeled by all conditions of interest during the learning process. Techniques of this type of learning depend heavily on information given by pre-determined inputs and outputs to construct classifiers that can predict outputs to new input data without knowing prior targets (*i.e.* labels). The ultimate goal of supervised learning is to let the machine generalize data into known categories. On the contrary, unsupervised learning (*i.e.* novelty detection) is when the model is provided with just input examples alone. The machine learns by finding patterns among unknown data and models the underlying structure or distribution in the data by itself. The goal of unsupervised learning is to have the algorithms congregate the input data into classes based on their intrinsic relationships, or statistical properties. In light of applications for these two types of learning in gearbox damage detection and identification alone and SHM as a whole, supervised learning can be used to identify the known damage conditions and potentially their locations while unsupervised learning can be used to identify new patterns when unknown damage conditions are introduced.

1.3 Existing Work

With rapid advancements in computer hardware and soft computing, more intelligent SHM methods for damage identification have been proposed and developed. In the past, vibration-based analysis for rotary machinery was performed by people to detect damage. Today, the same analysis can be used as feature extraction techniques for machine learning to achieve higher efficiency and accuracy in intelligent SHM. Re- search done in the last two decades has significantly expanded the capabilities of SHM methods. Some hybrid implementations of vibration-

based techniques and machine learning algorithms in condition monitoring are discussed below.

Techniques based on wavelet transform for detecting bearing-localized damage was first proposed by Li and Ma [3]. With the vibrational distribution obtained using wavelet transform, they were able to observe the frequency behaviors across the full spectrum from one instance to the other.

Samanta [4] presented a study that compared the performance of artificial neural network (ANN) and support vector machine (SVM) techniques in gear fault detection. Two cases of input feature sets pertaining to the study were considered: one set was first optimized by genetic algorithm (GA) and the other set was not. Results obtained from the study showed that SVM yielded higher classification accuracy than ANN without GA-based selection, but is equally accurate when using GA-based selection. Similarly, Saxena and Saad [5] also applied GA as a feature optimizer that determined the optimal number of "good" features for fault diagnosis. These features were then used as inputs to different ANN classifiers. They concluded that optimizing the selecting features with GA resulted in higher accuracy.

Satish and Sarma [6] demonstrated a novel and cost-effective approach for diagnosis and prognosis of bearing faults detection in induction motors. A hybrid between two different artificial intelligence techniques–ANN and fuzzy back propagation (Fuzzy BP)–were combined together to overcome their individual disadvantages.

Saimurugan *et al.* [7] demonstrated work on detecting multi-component faults in rotating machinery considering both shaft and bearing. In this work, a decision tree (DT) was used to select the best features, which were then classified by four different SVM kernel functions. It was observed that radial basis function (RBF) in support vector classification (SVC), C-SVC model, gives better classification efficiency than sigmoid function of C-SVC and nu-SVC models.

1.4 Objective

A great amount of research has been conducted on gearboxes damage detection and identification using machine learning algorithms such as k-NN, SVM, DT, ANN, etc. Much work carried out by some researchers investigate various ways to improve features extraction (*i.e.* reduce data dimension) while others proposed different methods on selecting the best input features for the learning algorithms. Other investigators conducted studies to compare the performance of different learning algorithms. However, not much attention has been paid on selecting the optimal penalty parameter, C, in SVM and the kernel parameter, γ when using RBF kernel. In addition, the majority of research work only incorporates one type of kernel function out of the four in SVM.

This research undertakes a serious attempt to apply SHM strategy to investigate gear faults utilizing a machine learning approach. There were three main objectives for this research study. The first objective of this thesis work was to conduct an experimental study for gearbox damage detection based on the statistical features and energy contents of two different types of wavelet transforms: discrete wavelet trans- form and wavelet packet transform in conjunction with SVM and ANN. The second objective was to investigate the performance of four different SVM kernel functions including Linear, Polynomial, RBF, and Sigmoid with the optimal parameters C and γ under various number of input features.

2 Theoretical background

2.1 Wavelet transforms

Wavelet transform was developed as an alternative to overcome limitations in the time-frequency domain [8]. Essentially, a wavelet is a finite periodic function that be- gins and ends with zero amplitude. In short, the function integrates to zero. Wavelet transform is an extended Fourier transform with adjustable windows. Applications pertaining to wavelet transform include, but are not limited to: data and image compression, noise reduction, transient detections, and pattern recognition.

Morlet and his colleague, Alex Grossmann proposed a single function called the mother wavelet, Ψ(t), defined as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \quad a, b \in \Re,\ a \neq 0 \qquad (1)$$

Where: t is the current time, a is the dilation parameter that measures the degree of compression or scale, and b is the translation parameter that determines the time location of the wavelet. The dilation and translation parameters allow the mother wavelet to manipulate its own shape through scaling and shifting along the original signal. What makes wavelets powerful is their ability to adapt the resolution trade-off between time and frequency domain simultaneously [8]. For example, the value $|a| < 1$ corresponds to smaller time-widths and higher frequencies while $|a| > 1$ associates with larger time-widths and

lower frequencies. This trade-off also obeys the Heisenberg Uncertainty Principle like other joint time-frequency domain analysis techniques.

2.1.1 Continuous Wavelet Transform (CWT)

The wavelet transform is a convolution of the wavelet functions and the continuous signal $x(t)$ as shown in Eq. (2):

$$WT(a,b) = \langle x(t), \psi_{a,b}(t) \rangle = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \psi^{\star}_{a,b} \left(\frac{t-b}{a} \right) dt \quad (2)$$

Where: $\psi(t)$ is the base wavelet and $\psi^{\star}_{a,b}$ is the complex conjugate of the base wavelet. If the base wavelet matches well with the original signal, a high wavelet coefficient value is achieved. Likewise, a lower wavelet coefficient value denotes a poor local matching. In order to be a classified wavelet, the function $x(t)$ must satisfy the following requirements:

1. **Finite energy.**

The wavelet must contain a determined energy defined by the integral of the base wavelet's magnitude squares.

$$E = \int_{-\infty}^{+\infty} |\psi(t)|^2 dt < \infty \quad (3)$$

2. **Admissibility condition.**

The Fourier Transform of $x(t)$ must vanish at zero (*i.e.* there is no zero frequency components). In simpler terms, $x(t)$ must have a zero mean. The mathematical expression is defined in Equation 3.4, where $\hat{\psi}(f)$ is the Fourier Transform of $\psi(t)$.

$$\text{Admissibility constant: } C_g = \int_0^{\infty} \frac{\left|\hat{\psi}(f)\right|^2}{f} df < \infty \quad (4)$$

In practice, a transformation is considered to be meaningful only when its corresponding inverse transformation exists [9]. This condition ensures the existence of the inverse wavelet transform.

3. **Real and zero values for complex wavelets.**

Any complex wavelet transforms must have a Fourier Transform value of real numbers and zero means for negative frequencies.

An inverse transformation of a classified wavelet can be used to reconstruct the original signal $x(t)$.

$$x(t) = \frac{1}{C_g} \int_{-\infty}^{+\infty} \int WT(a,b) \psi \left(\frac{t-b}{a} \right) \frac{1}{a^2} da db \quad (5)$$

In accordance with the criterion 2 defined above, Eq. (5) also implies that the energy of the original signal is fully preserved. Because a large number of dilation and translation values must be computed with CWT, the transformation is less computationally efficient compared against DWT and WPT.

2.1.2 Discrete Wavelet Transform (DWT)

Because CWT is a redundant transform in signal processing, a discrete version of the CWT is introduced to enhance computational practicality. This discretization of the dilation parameter a and translation parameter b is called the discrete wavelet transform (DWT). The DWT analyzes the signal with a length 2^N at different frequency bands where each band correlates with a different frequency resolution. An input signal in the time domain is decomposed into approximate and detail coefficients via high-pass and low-pass filters; at each pass the length of a sampling signal is reduced by a factor of two. Detail coefficients are values obtained from the high-pass filter, and approximation coefficients are values achieved through the low-pass filter. This procedure is iterated only in the low-pass filter at each level of decomposition.

Of the various forms of wavelet discretization, the dyadic discretization with values$a = 2^j$, $b = 2^j k$, and $j, k \in Z$ is the simplest, most efficient form, and most widely used where j and k are positive integers that control the wavelet dilation and translation respectively. With a special choice of the base wavelet $\psi(t) \in L^2(\Re)$, the dyadic discretization constructs a corresponding orthogonal wavelet such that:

$$\psi_{j,k}(t) = \frac{1}{2^j} \psi \left(\frac{t - 2^j k}{2^j} \right) \quad (6)$$

A function is considered to be orthogonal when all of its components vector are perpendicular to one another, in which case the product of all their basic functions is zero. By using the dyadic discretization in Eq. (6), the inverse DWT of a function $x(t)$ can be expressed as:

$$x(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{j,k} \psi(t) \quad (7)$$

Where: $C_{j,k}$ is the wavelet coefficient defined by the original signal and the dual function $\bar{\psi}(t)$ and $\psi(t)$.

$$C_{j,k} = \int_{-\infty}^{+\infty} x(t)\bar{\psi}_{j,k}(t)\,dt \tag{8}$$

Information achieved from DWT can be expressed by the signal's approximations and details defined. The signal details and approximations at level j and level J are defined in Eq. (9) and Eq. (10) respectively [1, 6].

$$D_j = \sum_{k\in Z} C_{j,k}\psi_{j,k}(t) \tag{9}$$

$$A_J = \sum_{k\in Z} D_j \tag{10}$$

Because the details and approximations that constitute the wavelet through DWT are less than the original signal, there is no redundancy. This striking aspect is what makes DWT less computationally expensive compared to CWT. Consequently, the binary structure of the signal and reconstruction is shown in Eq. (11) and Eq. (12).

$$A_{J-1} = A_J + D_J \tag{11}$$

$$x(t) = A_J + \sum_{j\leq J} D_j \tag{12}$$

For a decomposition level of j, there are $(j+1)$ sets of wavelet coefficients where one set contains the approximations (lowest frequency band) and j sets include the detail coefficients (higher frequency bands).

2.1.3 Wavelet Packet Transform (WPT)

WPT was introduced to overcome the limitations of DWT. WPT is similar to DWT but with further decomposed details and approximations in the high-frequency region. This makes WPT an attractive tool for many applications such as code theory, image processing, and signal processing. Since the recursive procedure of filter-decimation is used to compute details and approximations of a signal in both high-pass and low-pass filters, the wavelet packet tree structure is a complete binary tree where all frequency subbands in WPT are of the same width. The wavelet packet is defined as:

$$\begin{cases} \psi_{2p}^{(j)}(t) = \sqrt{2}\sum_n h[n]\,\psi_p^{(j)}(2t-n) \\ \psi_{2p+1}^{(j)}(t) = \sqrt{2}\sum_n g[n]\,\psi_p^{(j)}(2t-n) \end{cases} \tag{13}$$

with $p = 0, 1, 2, \ldots$ and $n = 0, 1, 2, \ldots, m$

Where: $\{h[n]\}_{n\in Z}$ and $\{g[n]\}_{n\in Z}$ are the high-pass and low-pass filter coefficients respectively, and n is the number of wavelet coefficients, p is the number of nodes in the binary tree at the j^{th} level [9]. The properties in Eq. (13) describe the nodes in the horizontal direction, *i.e.* subbands; the parent node is divided into two orthogonal nodes. Once the wavelet packet basic is determined, the result from WPT decomposition yields:

$$\begin{cases} d_{2p}^{(j+1)}[n] = d_p^{(j)} \star h[2n] \\ d_{2p+1}^{(j+1)}[n] = d_p^{(j)} \star g[2n] \end{cases} \tag{14}$$

Where: $d_p^{(j)}$ is the wavelet packet coefficients at the j^{th} level and subband p, and $\star$ is the convolution operator. Consequently, $d_{2p}^{(j+1)}$ and $d_{2p+1}^{(j+1)}$ are the wavelet coefficients of the children nodes of $d_p^{(j)}$. The reconstruction of the parent coefficient $d_p^{(j)}$ can be defined as a summation of its children coefficients as shown in Eq. (15):

$$d_p^{(j)}[n] = d_{2p}^{(j+1)}[n] \star h[n] + d_{2p+1}^{(j+1)} \star g[n] \tag{15}$$

In addition to using WPT coefficients in this work, the energies of these decomposed component signals are also exploited to enhance the effectiveness of gearbox damage detection. High-resonance frequencies are often located in high-frequency regions and are characterized by the energy concentration at that frequency band. In the case of a WPT, the energy content in each frequency subband (or node) of a signal $x(t)$ is:

$$E_p^{(j)}[n] = \int \left|d_p^{(j)}(t)\right|^2 dt = \sum_{n=1}^{m} d_p^{(j)}[n]^2 \tag{16}$$

Consequently, the total energy of a signal is a summation of all the energy components of all the subbands is:

$$E_{x(t)} = \sum_{j=1}^{2j} d_p^{(j)} \tag{17}$$

2.2 Support Vector Machines (SVM)

An alternate collection of discriminative classifiers in supervised machine learning are the Support Vector Machines (SVMs)–another members of the feedforward networks learning algorithms. SVMs were inspired by the statistical learning theory that was invented by Vladimir Vapnik and Alexey Chervonenkis in 1963, and was further developed by Vapnik and his colleges in the late 90s [10]. Unlike other classical learning algorithms (*e.g.* decision trees/forest and ANNS) whose principle is to minimize the error on the training data set, *i.e.* Empirical Risk Minimization (ERM), SVM is designed based on Structural Risk Minimization (SRM). The SRM principle is a trade-off between

the quality of the approximation and the complexity of the approximation function [11].

A SVM is a binary learning machine algorithm with sophisticated properties that can construct a hyperplane as the decision surface where positive and negative examples are separated with a maximum margin [12]. SVM uses a technique known as the kernel trick to transform the input data into higher dimension space where the optimal boundary can be drawn to separate outputs. Because of this, SVM can deal with a large number of input features. This capability allows SVMs to better identify the relationships between the data points from different classes, thus exhibiting outstanding generalization performance in conjunction with a high level of accuracy. Another advantage of SVM is that fewer training data points are required compared to other conventional classifiers (*e.g.* decision tree(s) or ANN) [13].

SVMs have been implemented in various fields for pattern recognition, regression estimation, and density estimation [11]. The state-of-the-art applications delivered by SVM learning algorithms spread across multiple fields such as image processing, text categorization, bioinformatics, etc. In fact, SVMs are arguably the most popular classifiers used in the SHM regime. Due to the robust performance of SVM and its popularity, it is chosen as the main classifier for this research. Detail background explanation on SVM can be found in Vapnik [11]. For completion, the fundamentals of SVMs is discussed as below.

Consider the training sample data input $\{(x_i, y_i)\}_{i=1}^{N}$, $y_i \in \{-1, 1\}$, $x_i \in \Re^n$ where x_i is the feature of the i training sample and y_i is the corresponding class label. The class label $y_i = -1$ and $y_i = 1$ represent Class 1 and 2 respectively. The data points x_i that lie on a two-dimensional hyperplane of a linearly separable must satisfy the inequalities:

$$\left(w^T.x_i + b\right) y_i \geq 1 \quad \text{for} \quad \forall i,\ y_i \in \{-1, 1\} \tag{18}$$

The normal distance projected from the origin to the optimal hyperplane is defined by $\frac{b}{\|w\|}$. When the data points are nonlinearly separable, a new set of non-negative numbers are incorporated into the definition of the optimal hyperplane such that

$$\left(w^T.x_i + b\right) y_i \geq 1 - \xi_i \tag{19}$$
$$\text{for} \quad i = 1, 2, \ldots, N\ \xi_i \in \Re_+$$

Where: ξ_i denotes the slack variables that measures the distance from the hyperplanes to their correlated data points that fall inside the margin. SVM classifies patterns by mapping the input vectors into higher dimensional space via a mapping function and separates them with hyperplanes. Let φ be the mapping function, the constrained-optimization problems for nonlinearly separable case is defined as:

$$\text{Minimize}\,(w, \xi): \quad \Phi(w, \xi) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \tag{20}$$
$$\text{With subject to}: \quad \left(w^T.\Phi(x_i) + b\right) y_i \geq 1 - \xi_i$$
$$\text{for} \quad i = 1, 2, \ldots, N \quad \xi_i \in \Re_+$$

In order to obtain the optimums values of the weight vector and bias in the constrained-optimization problem, the method of Lagrange multipliers is exploited; this application is known as the dual problem. A dual problem is when primal variables (*i.e.* dual variables) are solved to minimize the Lagrangian then with the same dual variables, the maximum of a Lagrangian can also be found with respect to the Lagrange multipliers. The equivalent Lagrange function is:

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i \left[y_i \left(w^T \Phi(x_i) + b \right) - 1 \right] \tag{21}$$
$$\alpha_i \in \Re_+$$

Going back to the constrained-optimization problem using the Lagrange function, it can be expressed as $\min_{w,b,\alpha} J(w, b, \alpha)$. This expression is the first step to solving the dual problem. In classical Calculus, finding the minima or maxima of a function is to differentiate that function and set the results equal to zero. Differentiating the Lagrange function $J(w, b, \alpha)$ with respect to w and α yield two conditions:

$$\frac{\partial J(w, b, \alpha)}{\partial w} = 0 \quad \text{and} \quad \frac{\partial J(w, b, \alpha)}{\partial \alpha} = 0 \tag{22}$$

The partial derivatives in Eq. (22) result in:

$$w = \sum_{i=1}^{N} \alpha_i y_i \Phi(x_i) \tag{23}$$

And

$$\sum_{i=1}^{N} \alpha_i, y_i = 0 \tag{24}$$

respectively. For any constraints that do not satisfy Eq. (23) and Eq. (24), their Lagrange multipliers must equal to zero. The next step is to obtain the Lagrange multipliers that maximize the objective function of the primal problem. This can be done by first expanding Eq. (21) and substituting the equalities in Eq. (23) and Eq. (24) into the equation, it follows:

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i y_i w^T \Phi(x_i) \tag{25}$$

$$-b\sum_{i=1}^{N}\alpha_i+\sum_{i=1}^{N}\alpha_i=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\alpha_i$$

The term w^Tw can be expressed as a function of α by substituting Eq. (23):

$$w^Tw=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_j\Phi(x_i)^T\Phi\left(x_j\right) \tag{26}$$

The Lagrange multipliers α_i can be solved using the optimization problem of $Q(\alpha)$. The constrained-optimization's dual problems for nonlinearly separable (soft-margin SVM) is:

$$\begin{aligned}\text{Minimize }(\alpha):\quad Q(\alpha)&=\sum_{i=1}^{N}\alpha_i \\ &\quad-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_j\Phi(x_i)^T\Phi\left(x_j\right)\end{aligned} \tag{27}$$

$$\text{With subject to}:\quad \sum_{i=1}^{N}\alpha_iy_i\quad 0\le\alpha_i\le C$$

$$\text{for}\quad i=j=1,2,\ldots,N$$

Once α_i values are determined, they can be used to compute the weight vectors w. Replacing the results of these two parameters back to the separating hyperplane's constraints, the decision function, $f(x)$, can be obtained:

$$\begin{aligned}f(x)&=\text{sgn}\left(w^T.\Phi(x)+b\right) \\ &=\text{sgn}\left(\sum_{i=1}^{N}\alpha_iy_i\Phi(x_i)\,\Phi\left(x_j\right)+b\right)\end{aligned} \tag{28}$$

Where: sgn is the sign function that extracts the sign of a real number. In SVM classification, the sign function determines the class label of an input sample (*i.e.* the test samples).

2.2.1 Kernel functions

The mapping functions φ always come in the form of $\Phi(x_i)^T\Phi\left(x_j\right)$, which represents an inner product that is a symmetric positive defined kernel function given by Mercer's theorem denoted as (x,x_i) [14]. Four major kernel functions in SVM classification are listed in Table 1.

Table 1: Sensor physical specification Kernel functions in SVM. Note: γ, r, and d are the kernel parameters.

Kernel function $K(X, X_i)$	Definition
Linear	x_i^Tx
Polynomial	$\left(\gamma x_i^Tx+r\right)^d,\ \gamma>0$
Radial Basis Function (RBF)	$\exp\left(-\gamma\left\|x_i-x\right\|\right),\ \gamma>0$
Sigmoid	$\tanh\left(\gamma x_i^Tx+r\right),\ \gamma>0$

2.3 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are supervised learning algorithms that try to mimic the cognitive functions of the human brain. A neural network structure is composed of neurons (*i.e.* the unit nodes). The fundamental neural network architecture is organized into layers. The number of layers varies depending on how complex the system is. Typically, a neural network architecture is comprised of three layers: an input layer that intakes patterns from the environment, an output layer that represents the network's response to the input patterns, and a hidden layer that interconnects the input and output layers. Each layer is made of a set of neurons with logistic activations that mimic the parallel computing functionality of the brain. The following introduction to ANN is given for completeness. Interested readers are encouraged to learn more about ANNs in Haykin [12] and Koper [15]. To further describe the structure of ANN, let's start with the most basic unit: the neuron. Each neuron contains inputs, synaptic weights, a bias, a summing junction, a local induced field, an activation function, and a single output. Figure 1 illustrates a nonlinear model of a neuron.

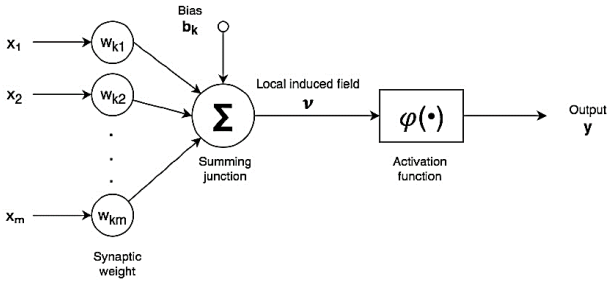


Figure 1: Components of a neuron.

It can be seen that each neuron has a number of input values that enumerated as $x_1,x_2,\ldots,x_m$. The inputs are scaled by their associated synaptic weight values, w_{kj} (input j^{th} of neuron k). The products between the inputs and their synaptic weights are then added up in the summing junction together with the bias term, b_k. The output produced by the summing junction is the local induced field, υ_k defined in Eq. (29). The local induced field serves as

an input to the activation function, φ. The activation function's product is the neuron output, y_k with a permissible limit range of or.

$$\upsilon_k = b_k + \sum_{i=1}^{m_k} w_{kj} x_{kj} \tag{29}$$

The output of neuron k is defined as:

$$y_k = \varphi(\upsilon_k) \tag{30}$$

There are various activation functions; the five most commonly used are shown in Table 2 the activation function is based on the application.

In SHM, Multilayer Perceptron (MLP) neural network is the most widely used algorithm. An illustration of the MLP neural network paradigm can be found in Figure 2.

Table 2: ANN's activation functions. Note: a is the slope parameter of the function.

Function	Definition	Range
Pure Linear	$\varphi(\upsilon) = \upsilon$	$(-\infty, \infty)$
Threshold	$\varphi(\upsilon) = \begin{cases} 0, \upsilon < 0. \\ 1, \upsilon \geq 0. \end{cases}$	$\{0, 1\}$
Piecewise-linear	$\varphi(\upsilon) = \begin{cases} 0, \upsilon \leq -\frac{1}{2} \\ \upsilon, \frac{1}{2} > \upsilon > -\frac{1}{2} \\ 1, \upsilon \geq \frac{1}{2} \end{cases}$	$[-1, 1]$
Sigmoid	$\varphi(\upsilon) = \frac{1}{1+e^{-a\upsilon}}$	$(0, 1)$
Tanh	$W2LOK$	$(-1, 1)$

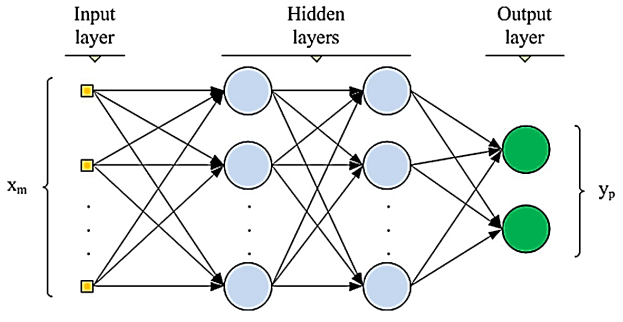


Figure 2: Architecture of a MLP neural network. Note: a perceptron is a composition of several neurons.

A MLP is a feedforward ANN model where the response output is mapped from the input data via a number of hidden layers ≥ 1. Feedforward describes the direction of how the computed activations propagate throughout the training process; they are continuously moved from one layer to another in order without any loop. The hidden neurons function as an intervener between the external input in the network output, and higher-order statistics from the input can be achieved by adding more hidden layers [12]. However, too many hidden layers may cause overfitting in training of a neural network. The learning technique utilized in training MLP network is the back-propagation algorithm. During the training process, many training examples are loaded into the network to compute the synaptic weights of the MLP with the purpose of creating an input-output mapping for the training data. This process is also known as curve fitting or generalization. The network should properly generalize the test data to the input-output mapping generated by the neural networks. Inputs to the neural networks are the features while outputs are the class labels.

Generally, the training process stops once any of the defined requirements are satisfied: maximum iteration (epoch), minimum gradient value, or the mean-square error (MSE) between the network output (predicted) and the target output (known) is minimized [12, 16]

3 Experimental implementation

3.1 Experimental setup

The vibrational data (*i.e.* the acceleration data) of the gearbox was obtained from Spectra Quest's Gearbox Dynamics Simulator (GDS), which is a two-stage parallel shaft gearbox with rolling bearings and a magnetic break. The gearbox is driven by a three-phase 3 hp motor with a maximum speed of 3600 rpm. All gears within the systems are spur gears with the same modules of 1.5. The input speed was set at 20 Hz, or 1220 rpm, under a small load of 5.7V. Table 3 summarizes the gear specifications and vibrational characteristics of the experiment.

The motor is directly connected to the gearbox's input shaft by a self-aligning coupling. The motor speed is controlled by the variable speed drive box and is measured by a built-in optical tachometer. The gearbox has a three in-line parallel shafts configuration as a two-stage reduction gearbox with an output shaft directly connected to the magnetic break that is controlled separately by a break controller. Two AC108-1A accelerometers are mounted on the gearbox housing: one is adjacent to the driving shaft to acquire vibration data in stage one, the other is adjacent to the driven shaft to measure acceleration data in stage two. Acceleration data acquired from sensor one (by the driving shaft) is used in this research since the damaged pinion

Table 3: Specifications and vibration characteristics of gearbox.

Function	First stage	Second stage
Number of teeth	Gear: 100	Gear: 90
Modules	1.5	1.5
Speed of shaft	1220 rpm (input)	393 rpm (output)
Modules	1.5	1.5
GMF	580 Hz	
Sampling frequency	128 Hz	

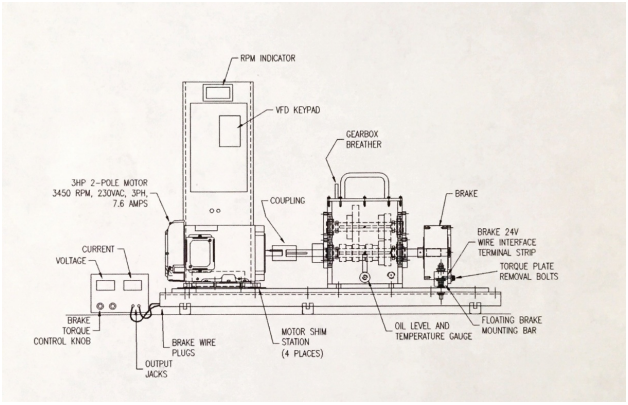


Figure 3: Experimental setup of a two-stage parallel shaft gearbox.

gear is located in stage one. Data acquisition is collected via the ADRE 408 DSPi/Sxp Dynamic Signal Processing Instrument by Bentley Nevada. Acceleration signals are sampled at a sampling frequency of 128 kHz. A full experimental setup with equipment used is demonstrated in Figures 3 and 4.

Usually, local gear failure modes fall into three categories: (a) pitting (surface wear fatigue), (b) tooth breakage (cracked, chipped, or missing tooth), and (c) scoring (surface worn on gear teeth). Type (b) gear failure, which means a sudden change in the systems stiffness, is used in this research and is shown in Figure 5.

3.2 Wavelet-based feature extraction

A flow diagram of the entire classification process including feature extraction, feature scaling, feature selection, and classification for SVM is shown in Figure 6. The main difference between SVM and ANN classification processes is that ANN does not need the penalty parameters C and the kernel widthγ.

Section 3.2.1 describes the process of extracting information from the vibration signal content using DWT and how statistical calculation is used to reduce feature dimension of the input matrix. This is referred to as Case 1. Similarly, the following Section 3.2.2 covers the process for using WPT. There are two cases in this method: Case 2 is where wavelet packet energy characteristics are used as input features to SVM and ANN for classification and Case 3 is where the statistical characteristics of the wavelet packet energy are used as input features.

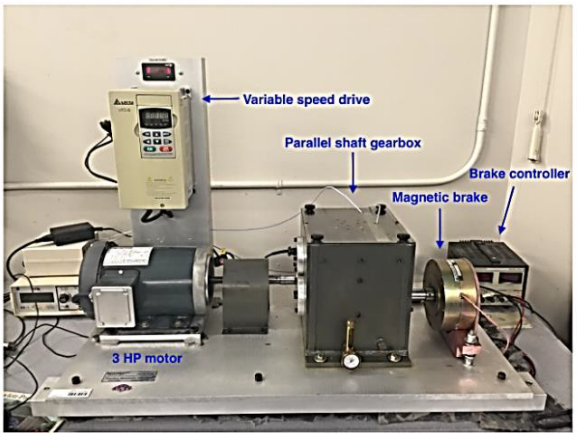
Table 4: Corresponding frequency bands at level 3 DWT decomposition.

Decomposition signal	Frequency range (Hz)
D_1	64000–128000
D_2	32000–64000
D_3	16000–32000
A_3	0–16000

3.2.1 DWT as data feature extraction method

Selections of a suitable mother wavelet and a level of decomposition are very important in analyzing signals in wavelet-based transformation. Therefore, different levels of decomposition were considered in this analysis to enhance the signal-to-noise ratio (SNR). After various trials, decomposition at level three using DWT was found to be suitable since higher detailed information consists of lower SNR. Moreover, since the frequency span of this experiment is 50000 Hz (*i.e.* the end frequency of the swept measurement), most useful frequency components will be found within this range. For these reasons, the original acceleration signals corresponding to two gearbox's conditions were decomposed into level three decomposition, consisting of four sets of wavelet coefficients. Data outputs from the high-pass filter are the detail coefficients—D while the last output from the low-pass filter stage is the approximate coefficients. A Four corresponding frequency subbands whose detailed breakdown can be found in Table 4.

(a) Data acquisition (DAQ) system



(b) Fully assembled test rig

Figure 4: Experimental setup with sensors and DAQ systems.

Figure 5: Damaged gear with missing tooth.

Table 5: Statistical feature parameters.

Feature	Function
Absolute mean	$mean = \frac{1}{N}\sum_{n=1}^{N}\lvert x(n)\rvert$
Root mean square	$rms = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n))}$
Standard deviation	
Kurtosis	$Kurtosis = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-mean)^4}{\left[\frac{1}{N}\sum_{n=1}^{N}(x(n)-mean)^2\right]^2}$
Peak-to-peak	$p2p = x(n)_{\max} - x(n)_{\min}$
Crest factor	$crest = \frac{\max(\lvert x(n)\rvert)}{rms}$
Skewness	$skew = \frac{1}{N}\sum_{n=1}^{N}\lvert x_n^3\rvert$

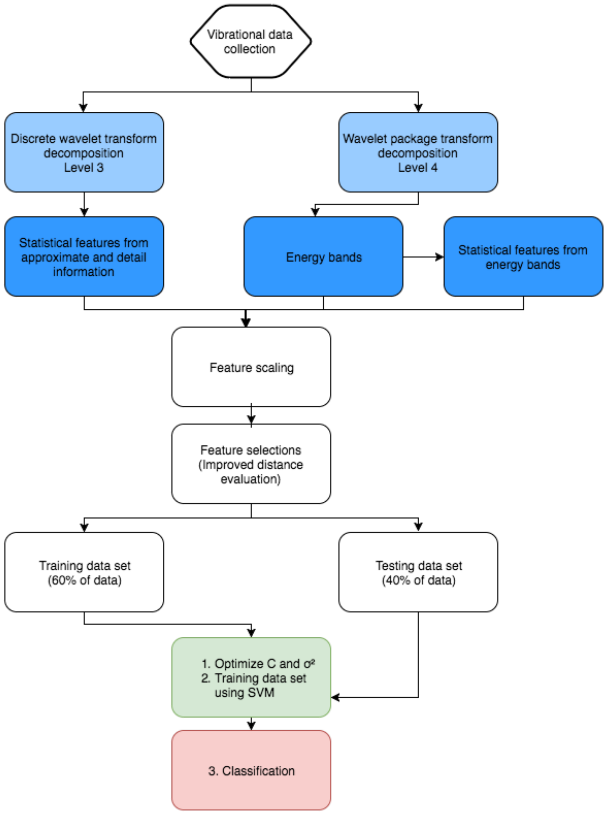


Figure 6: Framework of intelligent condition monitoring with wavelet-based data processing and statistical feature extraction for SVM. Note: $\sigma^2 = 1/\gamma$.

Case 1: Statistical characteristics of DWT wavelet coefficients

The length of the reconstructed approximate and detail coefficients is of 1048575 sample points. To reduce the volume of data, the detail and approximate coefficients obtained from DWT were divided into 255 bins with each bin consisting of 4096 sample points. Each subband was characterized by seven statistical features listed in Table 5. As a result, every decomposition signal contained seven distinct features, yielding a total of 28 attributes (4 frequency

subbands × 7 statistical characteristics = 28 attributes). Therefore, the input matrix size was 255×28 using the DWT-based features extraction method. Note that the word feature and attribute are being used interchangeably in this report.

3.2.2 WPT as data feature extraction method

While DWT decomposes the original signal into coarse frequency resolution, the WPT decomposes the signal into finer resolution in higher frequency zone that allows better signal analysis. Similar to the first method, a level of decomposition must be chosen for further analysis. WPT decomposes the signal into 2^j number of frequency bands at j level. However, there is a trade of between high frequency resolution and computation efficiency because the resolution of the frequency band increases exponentially with the decomposition level. It was suggested in existing work [14, 17, 18] that a level three decomposition is sufficient for wavelet-based features extraction. However, because the sampling frequency in this study is relatively high compared to those in the existing work (128 kHz versus ~30000 Hz), a higher decomposition level is needed to obtain higher frequency resolution. Therefore, a level of decomposition of four was selected.

Case 2: Wavelet packet energy

First, the original acceleration signal of each condition was divided into 255 bins, each consisting of 4096 sample points. Subsequently, WPT at level four decomposition was then applied to an individual bin to generate 16 sets of wavelet coefficients. Each set represented a frequency subband with a uniform width of 8000 Hz.

Statistical characteristics of the wavelet packet energy

Statistical characteristics of the energy bands were obtained in Case 3. Seven statistical calculations listed in Section 3.2.3 were exploited to extract seven features from each bin, reducing the size of the input matrix of one condition to $255W2LOK7$. Similar to the two previous cases, the process was repeated for the undamaged and damaged data sets, which led to a final input matrix size of 510×7.

3.2.3 Statistical features

The extracted information of the acceleration signal in the time-domain were characterized by seven statistical features including the absolute mean values, root mean square, standard deviation, kurtosis, peak-to-peak, crest factor, and skewness. These features represent the energy distribution, vibration amplitude, and the time series distribution of the acceleration signal [19].

All features are normalized into values between 0 and 1. The normalization equation is defined in Eq. (31).

$$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{31}$$

3.2.4 Feature selection

Generally, the number and type of input features affect the precision performance and computation efficiency of the learning algorithms and therefore, it is essential to select the best features to acquire good results. A feature selection can be seen as a search algorithm that optimizes the set of features that are sensitive to the gearbox's condition.

In the realm of machine learning, numerous feature selection techniques have been developed and implemented to eliminate redundancy of the input features (*i.e.* the curse of dimensionality). Some of the widely used techniques include, but are not limited to genetic algorithm, ant colony, principal component analysis, distance evaluation technique, etc. In the present work, the Improved Distance Evaluation technique (IDE) was adopted as the feature selection method due to its simplicity and reliability.

The IDE method was proposed by Hu *et al.* [17], Lei *et al.* [18], Ghiasi *et al.* [19] and Leia *et al.* [20] for their work in condition monitoring. The basic idea behind this technique is to find the highest ratio between the average distance of all samples within the same condition to the average distance of different conditions. The calculated ratio indicates the separability of different conditions under a specific feature: higher ratio corresponds to better separation and vice versa. Detail instruction for this algorithm can be found in Hu *et al.* [17], Lei *et al.* [18], Ghiasi *et al.* [19] and Leia *et al.* [20].

3.3 Gearbox faults diagnosis

Once the aforementioned data processing steps are completed, the next step is to perform fault detection of the gearbox with two machine learning algorithms. The fol-

lowing summarizes a step-by-step procedure for how to implement SVM using the LIBSVM library and ANN using the MATLAB neural networks toolbox.

3.3.1 Support Vector Machines

Four different built-in kernel functions Linear, Polynomial, RBF, and Sigmoid are utilized in this study. The performances of these kernel functions are evaluated at four different threshold values used in the feature selection step and with the three generated random sets of input features. This routine is applied across three aforementioned cases in Section 3.2.

Procedure for SVM implementation

Step 1: Create two sets of vectors $\{-1\}$ and $\{-1\}$ with the same length as the input matrix as the label sets (*i.e.* 510x1 per condition), where −1 represents undamaged system and represents damaged system.

Step 2: Shuffle the entire data set together with the label set. Create a training set (60% of the data) and a testing set (40% of the data).

Step 3: Perform feature scaling on the training data and testing data together. If there is a new testing data set, it must be scaled using the same minimum value and max-min range from the training data set for consistency.

Step 4: Next, implement the IDE technique for input features selection. This step is repeated with different thresholds to obtain different sets of input features. Depending on the feature extraction method used, different threshold values are applied. The threshold values and their corresponding features can be found in section 4. More focus will be devoted toward the classification accuracy since the training speed is negligible across different thresholds and kernel functions.

Step 5: Select the first 100 data points from the training set for parameters selection purpose (C and γ). The optimization process is implemented using 5-fold cross-validation and parameters that produce the best cross-validation result are selected to retain the original training set to generate the final SVM classifier.

Step 6: Use the constructed SVM classifier to classify the testing data set and collect the performance accuracy.

Step 7: Repeat step 1 through 5 for all SVM kernel functions (Linear, Polynomial, RBF, and Sigmoid).

Classification performance with noisy data

In order to study the performance of SVM, random noise was generated to add to the testing data sets. The selected thresholds for feature selection were 0.8 for Case 1 and 2, and 0.5 for Case 3. Since simulated random noise was a computer generated draw from the probability distribution, the Monte Carlo simulation methods was applied to obtain better accuracy of SVM's performance. For instance, some random values of noise added to the system affect the performance of SVM more than the other. Due to this reason, the classification accuracy of noisy data for each kernel function was the average of 500 simulations, and the noise percentage was increased from 5% to 60%. In order to avoid overfitting when working with the data contaminated with noise, the optimal value of C = 5.6569 obtained from all three cases and the kernel width = 1/3 were set to the default values in this case.

3.3.2 Artificial Neural Networks

The next machine learning method used in this study is a standard two-layered feedforward multilayer perceptron (MPL) neural networks as depicted in Figure 7.

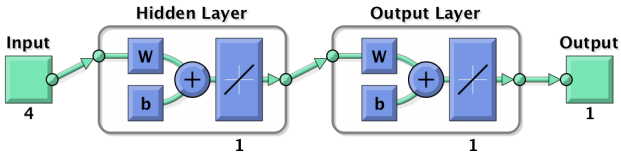


Figure 7: Schematic of a two-layer feedforward ANN. Note: The number of inputs correspond to the number of input features.

Similar to SVM, the data setup from step 1 through 3 was kept the same without step 4 because ANN does not need the kernel parameters. To make the comparison between SVM and ANN consistent throughout, ANN was also performed with and without the IDE features selection implementation throughout three cases. By contrast to SVM, the input vectors of ANN must be transposed so that the number of rows represent the normalized input features (*i.e.* one node is equivalent to one feature), while the number of columns represented the number of sample points due to the method's architecture. Multiple numbers

of hidden layers were determined through various trials; it was observed that one perceptron within the hidden and output layers was sufficient enough for the data extracted from the gearbox.

The ANN architecture in this work was generated and implemented using the MATLAB neural network toolbox. From the hidden to the output layer, linear transfer function was chosen as the mapping function from the net input to its layer's output. Within the hidden and output layer, the initial weights and biases were generated randomly. Furthermore, the desired adjustment for the weight factor was computed iteratively using the Levenberg-Marquardt training function. The network was trained and implemented by minimizing the cost function or the performance function—mean squared error (MSE)—between the predicted output and the target values. Here, the MSE goal was set to be 10^{-30}, the maximum number of iterations was 10^{30}, maximum validation failures of 6, and a maximum adaptive value—μ—of 10^{100}. The iterative process continued until any of those conditions were met.

4 Results and Discussion

4.1 Case 1: DWT coefficients' statistical characteristics

Recall that in the first case seven statistical features of each wavelet coefficients (approximations and details) were carried out, resulting in 28 input attributes. Figure 8 demonstrates the quantified sensitivities of the extracted features using IDE.

It can be observed that the statistical distributions in the lower frequency subbands are much more responsive to the gearbox's state of health because the wavelet approximations and details breakdown within these bands carry finer frequency resolution. Less sensitive features were eliminated using three different threshold values of 0.2, 0.4, and 0.8. In order to understand how much these features can affect the classification performance of SVM and ANN, three additional sets of features were selected as shown in Table 6. Consequently, Table 7 summarizes the classification performances of four different SVM kernel functions and ANN with various sets of input features.

When all features are used as inputs, Linear and Polynomial perform better than RBF, Sigmoid, and ANN. There are two explanations as to why the Linear kernel has a higher classification rate: (1) the features characterized by statistical values are linearly separable and (2) the number of instances (*i.e.* data points) is much larger than the number of features (510/28 ~ 18 times larger) which allows higher dimension space mapping to be more separable [21]. As the threshold increases, the number of input features decreases, leading to higher damage detection accuracy among all methods. In the cases where the number of input features are smaller (*e.g.* threshold = 0.8 and the last two randomly generated feature sets) RBF kernel outperforms the other kernel as well as ANN. On average comparison, Linear has the highest damage detection accuracy rate followed by ANN, Polynomial, RBF, and Sigmoid.

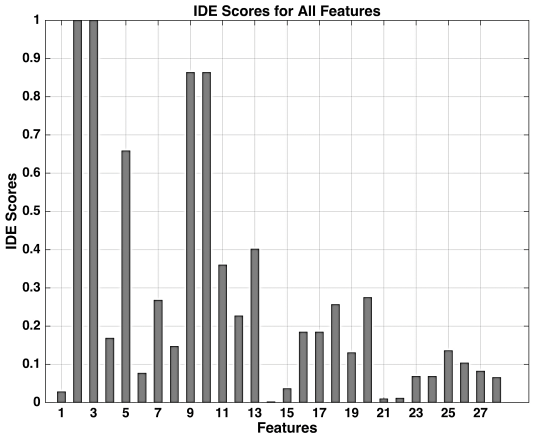


Figure 8: Improved distance evaluation (IDE) criteria of the statistical characteristics of DWT coefficients A_3 (1-7), D_1 (8-14), D_2 (15-21), and D_3 (22-27).

Table 6: Threshold values for IDE feature selections and corresponding features. Note: No specific threshold is defined in "N/A" cases; random sets of features are generated.

Threshold value	Input features
0	All
0.2	2, 3, 5, 7, 9, 10, 11, 12, 13, 18, 20
0.4	2, 3, 5, 9, 10, 13
0.8	2, 3, 9, 10
N/A	6, 14, 15, 21, 22, 23
N/A	7,6,16,24
N/A	12, 14, 27, 28

4.2 Case 2: WPT energy contents

As shown by the IDE sensitive scores in Figure 9, the extracted features from the WPT energy are a lot more sensitive to the condition of the gearbox compared to the statistical features obtain from DWT coefficients in Case 1. Ac-

Table 7: Case 1: Performance comparisons of SVM kernel functions and ANN. Note: DDA = damage detection accuracy.

Threshold	DDA (%) of SVM				DDA (%) of ANN
	Linear	Polynomial	RBF	Sigmoid	
0	100	100	99.51	99.51	99.51
0.2	99.51	99.51	100	99.51	99.51
0.4	100	100	100	100	100
0.8	100	100	100	100	100
N/A	58.33	50	40.69	56.37	54.41
N/A	73.53	72.55	75	70.10	73.53
N/A	71.57	69.12	72.06	56.86	69.61
Avg.	86.13	84.45	83.89	83.18	85.22

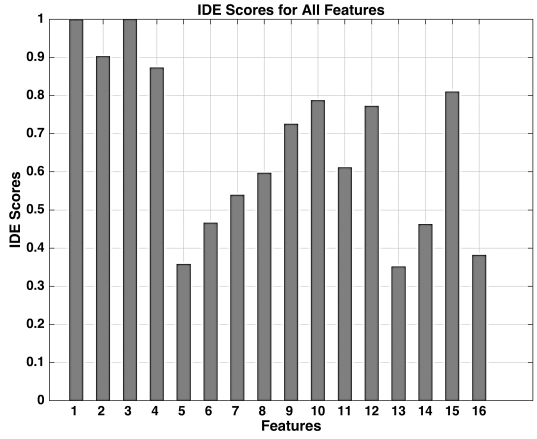


Figure 9: IDE criteria of the WPT energy bands.

Table 8: Threshold values for IDE feature selections and corresponding features. Note: No specific threshold is defined in "N/A" cases; random sets of features are generated.

Threshold value	Input features
0	All
0.2	All except 5, 13, 16
0.4	1, 2, 3, 4, 9, 10, 11, 12, 15
0.8	1, 2, 3, 4, 15
N/A	5, 13, 14, 16
N/A	5, 13
N/A	14, 16

celeration signal varies prominently in signal amplitudes for the damaged condition which impact the interval and energy distribution within frequency bands.

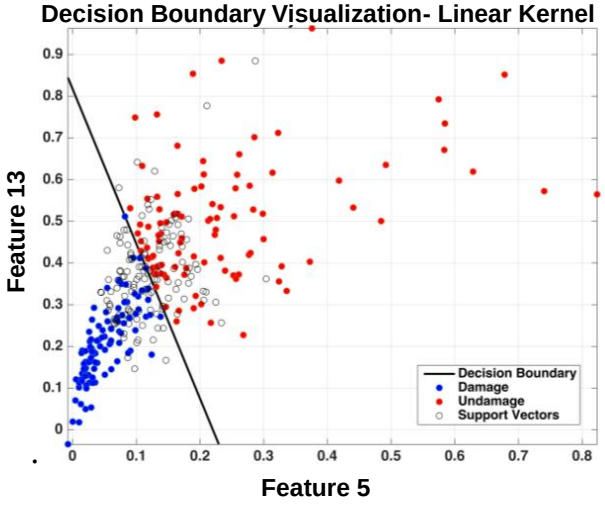
It can be observed that the lowest IDE score in Case 2 is 0.3, which is a lot higher relative to almost 0 in Case 1. Based on the general observation, the classification rates of using WPT energies as inputs are expected to be higher than using the statistic characteristics of DWT coefficients even with smaller numbers of features. The thresholds and their corresponding input features in conjunction with the three randomly selected feature sets can be found in Table 8.

Results shown in the Table 9 indicate the relationship between the threshold and classification rate is not completely linear for certain classification methods. For all SVM kernel functions and ANN, the damage detection accuracies are relatively high to perfect (100%, 99.51%, 100%, 8.04%, and 100%) when all features are used. However, when the first threshold is applied, the classification of all, except Linear and RBF, drops down. Sigmoid seems to have the larger fall (from 98.04% to 96.57%) compared to Polynomial and ANN. The classification rates begin to pick up a linear relationship with the number of sensitive features selected again when the threshold increases from 0.2 to 0.8 with two exceptions with the RBF kernel function and ANN. This phenomenon can be explained by overfitting due to a smaller number of input features. To better explain how overfitting can happen when the number of input features is small, the decision boundaries of all four SVM kernels functions are portrayed in Figure 10.

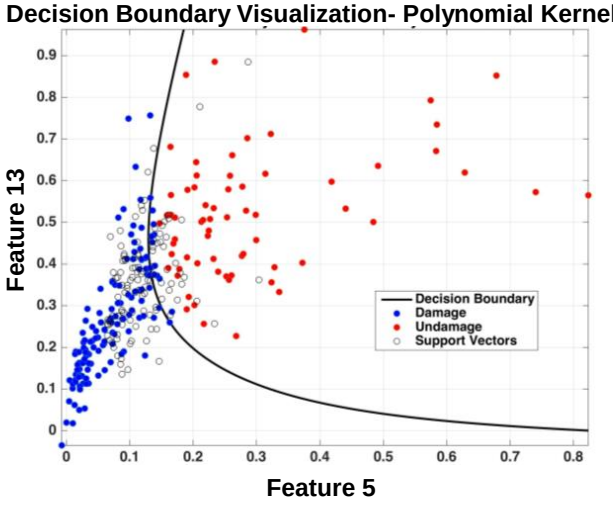
The two features used in this demonstration are Feature 5 and 13; their corresponding results are also listed in Table 10 on the third to last row. It can be seen that RBF kernel has a slightly better separation than Linear because the decision boundary is slightly curved. Decision boundaries generated by the Polynomial and Sigmoid kernels have much higher curvature and over-fit the data, yielding lower classification accuracies. Nonetheless, the average accurate detection rate among the five methods still ranges between 92.01% and 96.97%, which is significantly more precise than those in Case 1 even with less input features to begin with. The lowest damage detection rate achieved in Case 2 is 78.92% using a randomly selected feature set while the lowest rate in Case 1 is 40.69%. This indicates

Table 9: Case 2: Performance comparisons of SVM kernel functions and ANN. Note: DDA = damage detection accuracy.

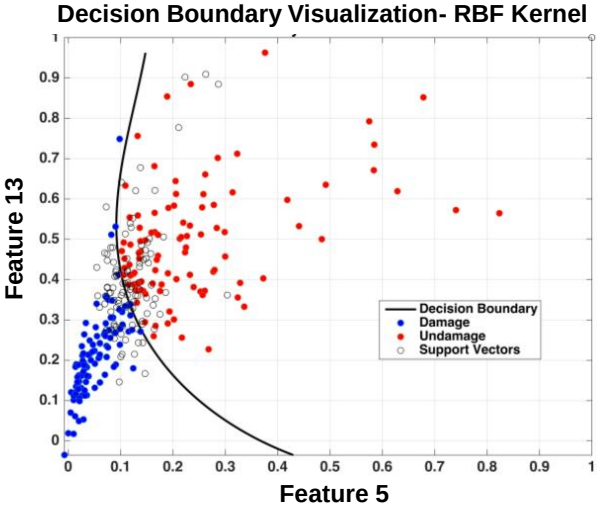
Threshold	DDA (%) of SVM				DDA (%) of ANN
	Linear	Polynomial	RBF	Sigmoid	
0	100	99.51	100	98.04	100
0.2	100	99.02	100	96.57	99.51
0.4	100	99.51	100	98.04	100
0.8	100	100	99.51	100	99.02
N/A	92.16	90.69	95.59	86.76	87.75
N/A	85.78	78.92	88.24	79.90	81.86
N/A	91.18	90.20	91.18	85.29	87.25
Avg.	95.59	93.97	96.36	92.01	93.63



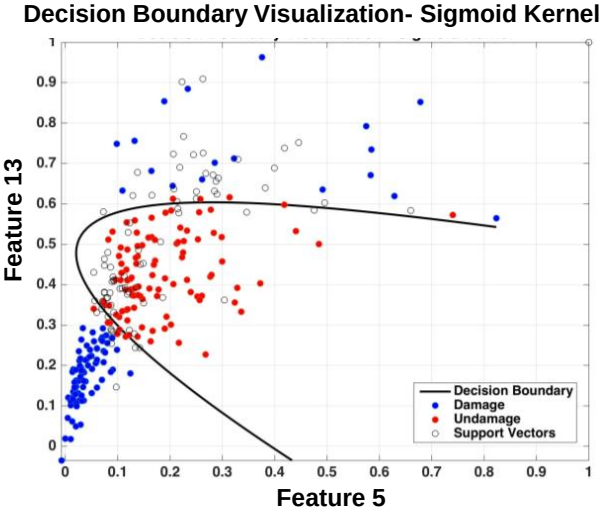
(a) Linear. DDA = 85.78%



(b) Polynomial. DDA = 78.92%



(c) RBF. DDA = 88.24%



(d) Non-Linear. DDA = 79.90%

Figure 10: Decision boundaries of all four SVM kernel functions using Feature 5 and 13.

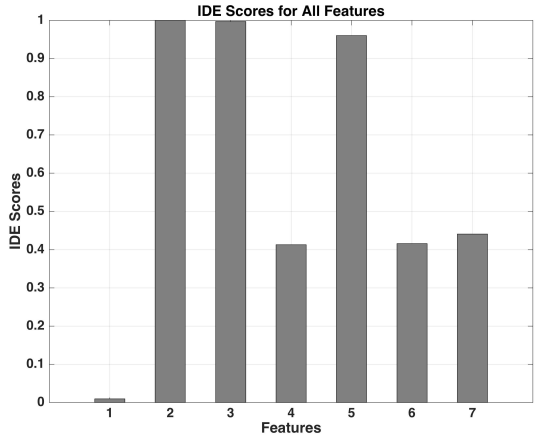


Figure 11: Improved distance evaluation criteria of the statistical characteristics of the WPT energy bands.

that detecting gear fault using wavelet energy packet is more effective than using the statistical features of wavelet coefficients. On the average performance in Case 2, RBF yields the best accuracy followed by Polynomial, Linear, ANN, and Sigmoid

4.3 Case 3: WPT energy contents' statistical characteristics

The third case study conducted in this research involved using the statistical characteristics computed from the wavelet energy packet in Case 2. Therefore, there were a total of seven input features in this case. Similar to the previous IDE features selection figures, Figure 11 exemplifies the responsibilities of the input features: absolute mean, RMS, standard deviation, kurtosis, peak-to-peak, crest factor, and skewness.

Based on the IDE scores, the absolute mean of all the energy subbands is almost not sensitive at all to the gearbox's condition. RMS, standard deviation, and peak-to-peak values are extremely sensitive to the system while the other three (kurtosis, crest factor, and skewness) are somewhat susceptible to damage within the gearbox. In light of basic statistical values within digital signal processing (DSP), the RMS and standard deviation measure how far the signal amplitude fluctuates from its mean while the peak-to-peak describe the excursion of the waveform (acceleration in this case) so they are sensitive to the signal energy. Kurtosis measures how flat or peaked the amplitude is compared to its normal distribution, crest factor indicates the extremity of peak values in a waveform, and the skewness indicates the symmetry of the waveform amplitude's probability density function. Thus, when the acceleration signal varies it impacts the amplitude distribution, impact interval and energy, which affect the aforementioned statistical values [14].

Table 10: Threshold values for IDE feature selections and corresponding features. Note: No specific threshold is defined in "N/A" cases; random sets of features are generated.

Threshold value	Input features
0	All
0.25	2, 3, 4, 5, 6, 7
0.5	2, 3, 5
0.99	2, 3
N/A	1, 3, 4
N/A	1, 2, 7, 4, 5
N/A	1, 4, 6, 7

Based on the results summarized in Table 11, the performances of four SVM and ANN fluctuate between 94.81% to 99.51% without any 100% damage detection accuracy regardless of the number of input features. Aside from the performance of the Sigmoid kernel when using the last set of random selected features set, the differences among success classification rates of all SVM kernel functions and ANN with different input features are relatively small. RBF kernel carries out the same damage detection rates of 99.02% throughout all seven demonstrations. Compared with the results presented in Case 1 and 2, Case 3 by far has the highest successful rate in gearbox damage detection. RBF also outperforms the other SVM kernel functions and ANN on average similar to Case 2 at 99.02%. The second best tool is Polynomial kernel (98.99%) follows by Linear (98.81%), ANN (98.53%), and Sigmoid (98.04%).

It can be observed that on average, using statistical features of wavelet energy produces higher success rate in gearbox damage detection in comparison with using wavelet energy packet and statistical values of wavelet coefficients. When dealing with statistical characteristics of wavelet coefficients, Linear kernel performs with the highest damaged detection accuracy (86.13%). RBF kernel, on the other hand, has the highest success rate in detecting gearbox damage using wavelet energies and their statistical features (96.36% and 99.02% respectively).

4.4 Classification performance with noise

In addition to comparing the performances of all SVM kernels and ANN, classification performances of the four SVM kernel functions are presented. Figures 12–14 below de-

Table 11: Case 3: Performance comparisons of SVM kernel functions and ANN. Note: DDA = damage detection accuracy.

Threshold	DDA (%) of SVM				DDA (%) of ANN
	Linear	Polynomial	RBF	Sigmoid	
0	98.53	99.51	99.02	98.53	98.04
0.25	98.53	99.02	99.02	99.02	98.04
0.5	98.53	98.53	99.02	98.53	99.02
0.99	99.51	98.53	99.02	98.53	99.02
N/A	99.02	98.53	99.02	98.53	99.02
N/A	99.02	99.51	99.02	98.53	98.04
N/A	98.53	98.53	99.02	94.61	98.53
Avg.	98.81	98.88	99.02	98.04	98.53

pict their performances when random noise is added to all three cases of study; Case 1, 2, and 3. For the noisy case, the SVM parameter C and γ are kept fixed among all kernel functions to avoid overfitting. Since the optimal value of C across all three cases is 5.6569, it is also used in this study. The kernel parameter γ is set to be $\frac{1}{\#offeatures} = \frac{1}{3}$ (*i.e.* default value in LIBSVM library guide [21]. Three feature sets chosen for these three noisy cases are those corresponding to the thresholds: 0.8 for Case 1 and 2, 0.5 for Case 3. Each feature sets contain three dominant features.

The performances of all four SVM kernels in Case 1 decrease drastically as the percentage of noise increases. When less than 15% noise is added to the testing data, Polynomial has lower classification performance than those of RBF, Linear, and Sigmoid. This is because its higher degree order (d = 2) causes overfitting. As the percentage of noise increases from 15% onward, Polynomial begins to outperform the rest, however, with a small margin. It can also be observed that Linear and RBF perform with roughly the same accuracy. Even with 60% noise added, the classification accuracies of all four kernels are still above 65%.

Compared with Case 1, Case 2 and 3 carry out higher classification accuracies; their lowest rates are above 75% even when 60% noise is added to the testing data. Although the Polynomial kernel in these two cases, however, are not performing as well as the other kernels, however, their differences are pretty negligible.

The classification accuracy trends for RBF-SVM of Case 1, 2, and 3 in this work are similar to the one presented in Hu *et. al.* [17] but with a slightly higher success test rate (classification accuracy ~ 57% versus ~ 70% in Case 1, ~ 77% in Case 2, and ~ 76% in Case 3).

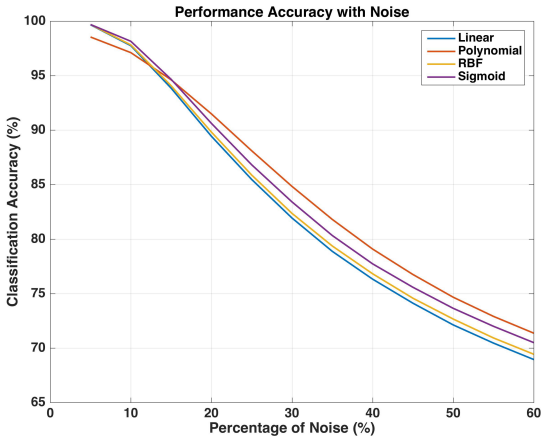


Figure 12: Classification performance with noise of the statistical characteristics of the DWT coefficients.

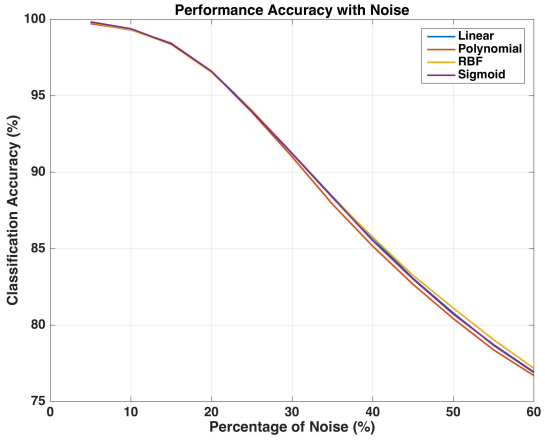


Figure 13: Classification performance with noise of the WPT energy bands.

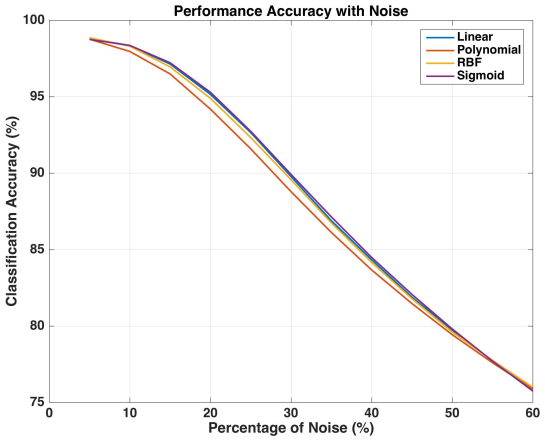


Figure 14: Classification performance with noise of the statistical characteristics of the WPT energy bands.

5 Conclusion

In this research, three methods for gearbox damage detection based on DWT coefficients and WPT energy in conjunction with SVM and ANN were conducted for comparison. The methodologies developed in this paper could be applied to other rotating elements as well. The damage level considered was level 1, *i.e.* detecting the presence of damage in a gear box. DWT and WPT were implemented to extract information from the gearbox acceleration data. This information was further processed into in- put features for SVM and ANN using statistical analysis. The improved distance evaluation technique (IDE) was used to eliminate non-sensitive features. In order to improve the performance of SVM, its penalty parameter C and the kernel parameter γ were optimized specifically for each kernel function with different input features via the grid-search method. Finally, the selected input features and optimal SVM parameters *C*and γ were used to generate classifiers to perform damage detection for the gearbox. The classification performances of the four SVM kernel functions with noisy data were also presented. The following are the conclusions accumulated from this research.

1. When applying wavelet decomposition to extract features from gearbox's acceleration data for machine learning algorithms, a decomposition level between 3 and 4 is sufficient for gearbox damage detection. In contrast, if wavelet trans- form is used as a main tool to analyze gearbox signals, then a higher level of decomposition will be needed. For this system, the smallest decomposition level of 8 was needed to identify gearbox damage based on wavelet packet energy.
2. Linear-SVM has the highest classification accuracy when using statistical characteristics of DWT coefficients as input features to SVM and ANN. RBF-SVM, on the other hand, outperform Linear, Polynomial, Sigmoid-SVM and ANN when using wavelet packet energy characteristics or their statistical values as inputs.

In general, input features extracted from WPT energy characteristics are more sensitive to the condition of a gearbox than those from the DWT coefficient. Therefore, higher damage detection accuracy can be achieved using features obtained from the wavelet energy packet.

References

[1] Z. Hou, A. Hera and M. Noori, Wavelet-Based Techniques for Structural Health Monitoring, [auth.] Atchintya H. [ed.] Atchintya H. Health Assessment of Engineered Structures: Bridges, Buildings and Other Infrastructures, World Scientific Publishing Company Pte Limited, 2013, (2013) 179-199.

[2] C. R. Farrar, and K. Worden, Structural Health Monitoring: A Machine Learning Perspective, Wiley, (2013)17-21.

[3] J. C Li., and J. Ma, Wavelet Decomposition of Vibrations for Detection of Bearing-localized Defects, J. NDT & E International, 30 (3) (1997) 143-149.

[4] B. Samanta, Gear Fault Detection using Artificial Neural Networks and Support Vector Machines with Genetic Algorithms, J. Mechanical Systems and Signal Processing, 18 (2004) 625-644.

[5] A. Saxena, and A. Saad, Evolving an Artificial Neural Network Classifier for Condition Monitoring of Rotating Mechanical Systems, J. Applied Soft Computing, 7 (1) (2005) 441-454.

[6] B. Satish, and N.D.R. Sarma, A Fuzzy BP Approach for Diagnosis and Prognosis of Bearing Faults in Induction Motors, IEEE Conference Publications, 3(2005) 2291-2294.

[7] M. Saimurugan, K.I. Ramachandran, V. Sugumaran, and N.R. Sakthivel, Multi component fault diagnosis of rotational mechanical system based on decision tree and support vector machine, J. Expert Systems with Applications, 38 (4) (2011) 3819–3826.

[8] L. Debnath and, F. A. shah, Wavelet Transforms and Their Applications, Volume 2, volume 2, Springer, New York NY, (2015), ISBN: 978-1-4612-6610-5 (Print) 978-1-4612-0097-0 (Online).

[9] R. X. Gao, and, R. Yan, Fourier Transform to Wavelet Transform: A Historical Perspective, Chapter 2, Wavelets Theory and Applications for Manufacturing, Springer, (2011) 17-22.

[10] N. Cristianini, and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, (2000), ISBN: 9780521780193.

[11] N. Vapnik, An Overview of Statistical Learning Theory, IEEE Transactions on Neural Networks, 10 (5) (1999) 988 - 999.

[12] S. S. Haykin, Neural Networks and Learning Machines, 3rd Edition. Pearson Education, Inc., Upper Saddle River, New Jersey,

(2009), ISBN-13: 978-0-13-147139-9.

[13] S. Choa, Asfour, S., A. Onar and N. Kaundinya, Tool Breakage Detection using Support Vector Machine Learning in a Milling Process, International Journal of Machine Tools and Manufacture, 45 (3) (2005) 241–249.

[14] C. Shen, S. Wang, F., and P. W. Tse, Fault diagnosis of rotating machinery based on the statistical parameters of wavelet packet paving and a generic support vector regressive classifier, J. Measurement, 46 (4) (2013) 1551-1564.

[15] A. Koper, "Neural Networks Performance and Structure Optimization using Genetic Algorithms", Master's thesis, San Luis Obispo, CA, USA : California Polytechnic University, (2012).

[16] F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Washington, Spartan Books, (1962).

[17] Q. Hu, Z. He, Z. Zhang, and Y. Zi, Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble, J. Mechanical Systems and Signal Processing, 21 (2) (2007) 688–705.

[18] Y. Lei, Z. Heb, and Y. Zib, Application of an Intelligent Classification Method to Mechanical Fault Diagnosis" J. Expert Systems with Applications, 36 (6) (2009) 9941–9948.

[19] R. Ghiasi, P. Torkzadeh, and M. Noori, Structural Damage Detection using Arti cial Neural Networks and Least Square Support Vector Machine with Particle Swarm Harmony Search Algorithm, International Journal of Sustainable Materials and Structural Systems, 1 (4) (2014) 303–320.

[20] Y. Leia, Z. Hea, and, Y. Zia, A New Approach to Intelligent Fault Dianogsis of Rotating Machinery, J. Expert Systems with Applications, 35 (4) (2008) 1593–1600.

[21] C. Hsu, C. Chang, and C. Lin, A Practical Guide to Support Vector Classification, Technical report Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, (2003).