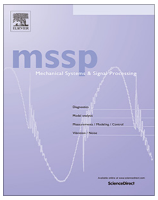
Mechanical Systems and Signal Processing 145 (2020) 106951

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp



Frequency-temporal-logic-based bearing fault diagnosis and fault interpretation using Bayesian optimization with Bayesian neural networks

Gang Chen [a,*], Mei Liu [b], Jin Chen [c]

[a] *Department of Mechanical and Aerospace Engineering, University of California, Davis, Davis, CA, USA*
[b] *Department of Automation, School of Electrical and Information Engineering, Tianjin University, Tianjin, China*
[c] *School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China*

ARTICLE INFO

Article history:
Received 5 July 2019
Received in revised form 23 March 2020
Accepted 27 April 2020
Available online 15 May 2020

Keywords:
Bayesian neural networks
Fault diagnosis and interpretation
Frequency-temporal-logic
Rolling element bearing

ABSTRACT

Rolling element bearings are widely used components in modern rotary machines, and accurate diagnosis and interpretation for faults of bearings are significant for equipment maintenance. This paper introduces a fault diagnosis method with a formal specification language, which overcomes the difficulty of understanding the decision process of fault diagnosis. The formal specification is written with a novel formal language, called frequency-temporal-logic, defining the time-frequency properties of time series signals, which not only is a classifier to diagnose the faults but also gives interpretations for the fault signals with its semantics. To find an optimal description for the fault signals, the Bayesian optimization with Bayesian neural networks has been utilized to infer the structure and parameters of the formal specification. The semantics of frequency-temporal-logic then gives the fault interpretation. Moreover, the quantitative semantics for the formal language is defined based on a novel satisfaction metric, which has a noise resistance property. Analysis of the proposed method shows that the formal description can deal with noisy signals and variable speed operations of the bearings. Finally, comparison experimental results indicate the proposed method can obtain high fault diagnosis accuracy.

Published by Elsevier Ltd.

1. Introduction

Rolling element bearings are widely used components in modern rotary machines, and faults occurring in bearings may lead to the fatal breakdown of machines [1]. Therefore, accurate detection and diagnosis of the condition of bearings are significant for the continuing operation of these machines. Moreover, a good interpretation for the fault diagnosis process is significant for timely response to the fault occurring [2,3]. This argument can be illustrated with the implement of bearing condition monitoring techniques to a wind turbine in [4]. If the monitoring system can detect deterioration of bearings, e.g., finding that the high temperature causes the faults. The experts then can find out that the high temperature is due to insufficient or inefficient lubricant properties with his domain knowledge. However, few existing methods can be found that can obtain good fault diagnosis and fault interpretation results simultaneously.

[*] Corresponding author.
E-mail address: megangchen@gmail.com (G. Chen).

https://doi.org/10.1016/j.ymssp.2020.106951
0888-3270/Published by Elsevier Ltd.

Hitherto, vibration signals, which have been demonstrated by many scholars [5,6], have carried rich information on bearing conditions and are sensitive to bearing faults. A variety of signal processing techniques have been extensively investigated to interpret the vibration signals and accurately extract fault characteristics [7–9]. Many signal processing methods are based on the interpretation that the fault signals come from the strikes of rollers on the fault surface and excite the resonant frequencies of structures between the bearing and transducers, which are called mechanism-based methods. These methods interpret the fault mechanism as a series of impacts or impulses, which excite the entire system where the bearing is mounted, thus called mechanism-based methods. Mechanism-based signal analysis methods, therefore, justify themselves by depicting the relationships between the vibration signals and the fault mechanisms, such as kurtosis [10], cyclostationary [11], envelope analysis [12], Bandwidth EMD [13]. Based on the fault mechanisms and its associated patterns, a good mechanism-based signal analysis technique should explicitly extract the periodic information of the impulsive response of a faulty bearing.

Mechanisms-based fault diagnosis methods provide good interpretations of the decision process.However, these methods usually based on knowledge of domain experts, which is hard to obtain. To overcome this issue, another class of fault diagnosis methods, which are called data-driven fault diagnosis approaches, is proposed to diagnose the faults with machine learning algorithms based on the extracted features from the vibration signals. However, data-driven approaches pay little attention to the fault mechanisms and try to determine the health state of the rolling element bearings automatically with feature diagnosis techniques. The decision process for these methods usually is not transparency to human users, which acts as a black box for end-users. For example, hidden Markov model (HMM) [14–17], k-nearest neighbors [18], support vector machine (SVM) [19,20], and Gaussian process [21], have been applied to fault diagnosis of rolling element bearings and obtained good performance, but these methods pay little attention to the interpretability of the models. Another typical example of data-driven methods is to use Deep Neural Networks (DNN) to learn features with multiple layers of abstraction and thus are capable of modeling complex patterns [22,23]. DNN based fault diagnosis methodologies for bearings have received increased attention in recent years [24,25]. However, these methods come across the same problem; they cannot provide good interpretation for the decision process.

Data-driven approaches are a powerful tool for fault diagnosis, but industry experts face some difficulties when trying to interpret the results. This raises the problem of increasing data-driven models' interpretability and interest in developing fault diagnosis able to generate 'domain-level' knowledge that is close to experts' knowledge. In other words, data-driven algorithms that are able to give explanations about why the time series signals contain fault characteristic and can build a relationship between physical phenomena and causality of faults, thus end users can take further actions to avoid or fix the faults. Therefore, one of the most important questions that remain now is, how do we provide transparency and interpretability for data-driven methods, such that human can take actions based on the interpretation of the fault diagnosis process, thus to guarantee the performance of machines.

To the best of our knowledge, the problem of finding a formal interpretable model for fault diagnosis among the time-frequency domain with data-driven algorithms largely remains untouched, which is the topic of this paper. To tackle the problem, we need to answer several essential questions. First, what kinds of formal language can be used to characterize the properties of the fault signal better? Second, how to infer the structure of the formal description with a data-driven method? Last, how can we deal with the noises among the time series data that affect the interpretation of the fault diagnosis results? In this paper, we make concrete progress in answering the above questions. Specifically, the contributions of this paper are:

1. A novel frequency temporal logic (FTL) is proposed to describe the frequency properties of fault signals, which can reveal the relationship between physical mechanisms and decision process, and whose quantitative semantic has noise resistance property.
2. Bayesian Neural Networks (BNN) has been combined with Bayesian optimization to infer the structure and parameters of the FTL description, solving the fault diagnosis problem with limited computation cost.
3. A data-driven method, i.e., the BNN based method, has been combined with a logic based method, i.e., FTL-based method, to improve the interpretability of the fault diagnosis results.
4. Two experiments are conducted to investigate the properties of the proposed method, and the performance of the method shows FTL based descriptions gives reasonable explanations while producing competitive fault diagnosis performance.

The layout of this paper is as follows: Section 2 introduces the related works of signal temporal logic and Bayesian neural network. Section 3 formulates the fault diagnosis and interpretation process as an optimization problem. Section 4 solves the optimization problem using BNN with Bayesian optimization technique. Section 5 analyzes the properties of the proposed method in condition monitoring. Section 6 demonstrates the proposed method with real experiments and Section 7 draws the conclusions of this paper.

2. Related works

In this section, we summarize related work on fault diagnosis and interpretation approach presented in this paper. We categorize related work into two groups: classification with temporal-logic-based formal language and optimization with BNN.

2.1. Time series classification with signal temporal logic

A continuous-time, continuous-value signal is a function $s \in \mathcal{F}(\mathbb{R}^+, \mathbb{R}^n)$. Denote the value of signal s at time t as $s(t)$, then signal temporal logic (STL)[26] is a temporal logic defined over signals. STL is a predicate logic (e.g. $s(t) > 4$) with interval-based temporal semantic. STL defines a formal language described with STL formulas (can be seen as sentences in natural language). The set of all STL formulas is denoted by Φ, which is recursively defined as

- $\mu \in \Phi$, where μ is a predicate in the form $\mu := (f(s(t)) \sim d), \quad f \in \mathcal{F}(\mathbb{R}^n \rightarrow \mathbb{R}), \sim \in \{<, >, \leqslant, \geqslant\}$.
- If $\varphi_1, \varphi_2 \in \Phi$, then $\varphi_1 \wedge \varphi_2, \varphi_1 \vee \varphi_2 \in \Phi$, where $\wedge$ and $\vee$ are conjunction and disjunction connectives, respectively.
- If $\varphi \in \Phi$, then $\mathbf{G}_{[a,b)}\varphi, \mathbf{F}_{[a,b)}\varphi \in \Phi$, where **G** and **F** are the temporal operators denote "globally" and "finally", and $[a,b)$ is the time bound for the formulas. Some examples to denote the properties of **G** and **F** will be given later.

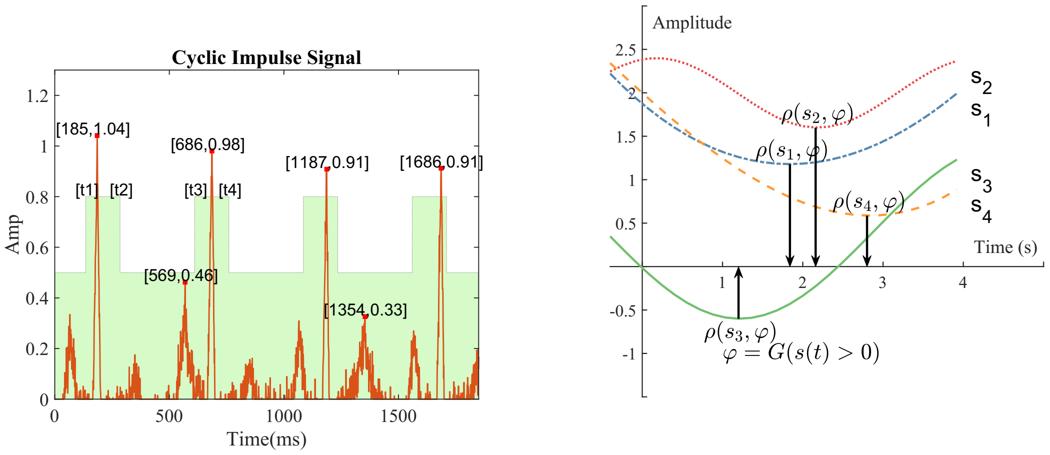
The above syntaxes define how to use basic words, e.g., $\mu, \varphi_1, \varphi_2$, to construct a sentence (or formula). With these syntaxes, STL is a powerful language to express the pattern of signals and has been widely used in safety-critical systems to specify the specifications or properties of the systems. The exist of temporal operators **G** (globally) and **F** (finally) make STL be suitable for depicting the periodical properties of rolling element bearing fault signals. For instance, the fault signal of a rolling element bearing contains a cyclic impulse energy at frequency ω, and the periodical impulse energy, has the pattern written in English whenever the frequency component ω has a energy density smaller than 0.5 for 0.325 s, the energy density at frequency ω will be bigger than 0.8 in the next 0.15 s. If we denote $f_\omega(t)$ to be the time-frequency representation of the signal at frequency ω and time t, the pattern (see Fig. 1(a)) can be easily described by an STL formula $\varphi := \mathbf{GF}_{[0,0.475]}(\mathbf{G}_{[0,0.325]}(f_\omega(t) < 0.5) \rightarrow \mathbf{F}_{[0,0.15]}(f_\omega(t) > 0.8))$, where **G** without any time bound means a time bound $[0, \infty]$. In the formula, $\mathbf{GF}_{[0,0.475]}$, read as globally finally within 0.475 s, denotes that after the pattern occurs, then within 0.475 s, the pattern will occur again. $\mathbf{G}_{[0,0.325]}(f_\omega(t) < 0.5)$ denotes the amplitude of the signal will be globally smaller than 0.5 for 0.325 s, $\rightarrow$ denotes implication relationship. $\mathbf{F}_{[0,0.15]}(f_\omega(t) > 0.8)$ denotes the energy density of frequency component ω can be over 0.8 at least once within next 0.15 s. As shown in Fig. 1(a), at time t_2, the signal is smaller than 0.5 until time t_3, and $t_3 - t_2 = 0.325$ s. After keeping the value being smaller than 0.5 for 0.325 s, the signal is bigger than 0.8 between time t_3 to t_4, and $t_4 - t_3 = 0.15$ s.

Given a signal $s(t)$ and an STL formula φ, in order to check whether the formula is satisfied by the signal, denoted as $s(t) \models \phi$, STL is equipped with quantitative semantic, which is called robustness degree [27,28] (also called "degree of satisfaction") that quantifies how well a given signal s satisfies a given formula φ. The robustness degree is calculated recursively as

$$\begin{aligned}
\rho(s,(f(s) < d),t) &= d - f(s(t)) \\
\rho(s,(f(s) \geqslant d),t) &= f(s(t)) - d \\
\rho(s,\varphi_1 \wedge \varphi_2,t) &= \min(\rho(s,\varphi_1,t),\rho(s,\varphi_2,t)) \\
\rho(s,\varphi_1 \vee \varphi_2,t) &= \max(\rho(s,\varphi_1,t),\rho(s,\varphi_2,t)) \\
\rho(s,\mathbf{G}_{[a,b)}\varphi,t) &= \min_{t' \in [t+a,t+b)} \rho(s,\varphi,t') \\
\rho(s,\mathbf{F}_{[a,b)}\varphi,t) &= \max_{t' \in [t+a,t+b)} \rho(s,\varphi,t').
\end{aligned}$$

We use $\rho(s,\varphi)$ to denote $\rho(s,\varphi,0)$, which indicates the signals started with $s(0)$. If $\rho(s,\varphi)$ is large and positive, then s would have to deviate substantially in order to violate φ. Fig. 1 (b) shows the robustness degrees for 4 signals, where the formula is $\mathbf{G}(s(t) > 0)$, read as "the signal $s(t)$ should be globally greater than 0" (a temporal operator without time bound means all the time). The directed lines show the robustness degrees, which depicted with the signed distance to the time axis, the downward arrows indicate positive robustness degrees and the upward arrow indicates negative robustness degree.

With the definition of STL and its semantics, time series classification can be conducted by learning an STL formula φ, such that all the time series belonging to positive examples, denoted as $x \in X^+$, will satisfy the formula φ and has non-negative robustness ($\rho(x,\varphi,t) \geqslant 0$, here we define zero robustness as satisfaction). While all the time series belonging to negative examples, denoted as $x \in X^-$, will violate the formula φ and has negative robustness ($\rho(x,\varphi,t) < 0$). Existing techniques for learning STL formulas try to describe the behaviors of trajectories of differential equations or hybrid models under the name "requirement mining" [29,30]. Requirement mining has emerged as an effective approach to generate abstractions of the time series to better understand complex systems, e.g., autonomous systems. These approaches fall into two categories. The approaches in the first category assume that the output of the requirement mining problem is a formula φ_θ with a fixed structure but unknown parameter θ. Under this assumption, the learning of STL formula can be transformed into an



(a)

(b)

Fig. 1. (a) A cyclic impulse signal with pattern described with STL formula $\varphi := \mathbf{GF}_{[0,0.475]}(\mathbf{G}_{[0,0.325]}(f_\omega(t) < 0.5) \rightarrow \mathbf{F}_{[0,0.15]}(f_\omega(t) > 0.8))$; (b) Robustness degrees for different signals with formula $\phi = \mathbf{G}(s(t) > 0)$ indicated by directed line.

optimization problem with the goal of finding a parameter θ^*, such that φ_θ will distinguish the time series, optimizing certain cost function, which is usually defined with the concept of robustness [31].

Obviously, the fixed structure assumption is not realistic, since we need a domain expert, who knows formal methods and the structure of the requirements, to prescribe the requirements. To release this assumption, the approaches in the second category infer the STL formula with proper structure and the associated parameters simultaneously. Algorithms in [32,33] solved the problem using a mixture of discrete and continuous optimization using decisions trees and simulated annealing in a supervise learning way. Typically, the problem is addressed in two steps, learning the structure followed by the synthesis of parameters. The structure of the formula in [33] comes from exploring a directed acyclic graph and in [32], a decision tree approach is used to learn both the structure and the parameters. Similar to the decision tree approach, [34] proposed grammar-based decision trees (GBDTs) to derive for the logical expression with Monte Carlo, grammatical evolution and genetic programming separately. These methods have solved the structure inference problem in some way, but they do not adequately address the issue for two reasons: 1) They inferred the formulas in two steps iteratively, which was a time-consuming process; 2) The formulas obtained by these methods were combination of a set of predefined atom formulas with logic operators $\vee$ and $\wedge$, thus the number of formulas can be obtained by these methods are limited. Moreover, the robustness degree of STL is originally defined to check the satisfaction of specifications for safety-critical systems, which is very sensitive to noise. Thus the present definition for robustness degree is not suitable for vibration signal analysis, which is contaminated by a variety of noises.

2.2. Bayesian neural network optimization

Bayesian optimization is an effective methodology for the global optimization of expensive-to-evaluate black-box functions. It relies on querying a distribution over functions defined by a relatively cheap surrogate model. To overcome the drawback of Gaussian processes (GPs) based Bayesian optimization, whose inference time grows cubically in the number of observations, the neural networks have been combined with Bayesian optimization to learn an adaptive set of basis functions for Bayesian linear regression [35]. The goal of BNN is to uncover the full posterior distribution over the network weights in order to capture uncertainty, to act as a regularizer, and to provide a framework for model comparison [35]. BNN has been widely used in many fields, such as language process [36], hyperparameter optimization [37] and variational inference [38]. In this paper, the BNN is used to find the optimal formula description for the fault signals, whose objective function is high-dimensional, nonlinear and not differentiate.

3. Problem formulation

In this section, we first motivate this paper with a numerical example, then formally define the problem of FTL-based fault diagnosis and interpretation.

3.1. Motivation example

Rolling element bearing fault diagnosis based on time series signals is a hard task due to the noise and the complex operation conditions of bearings. As a result, many scholars have utilized time-frequency information to diagnose the conditions

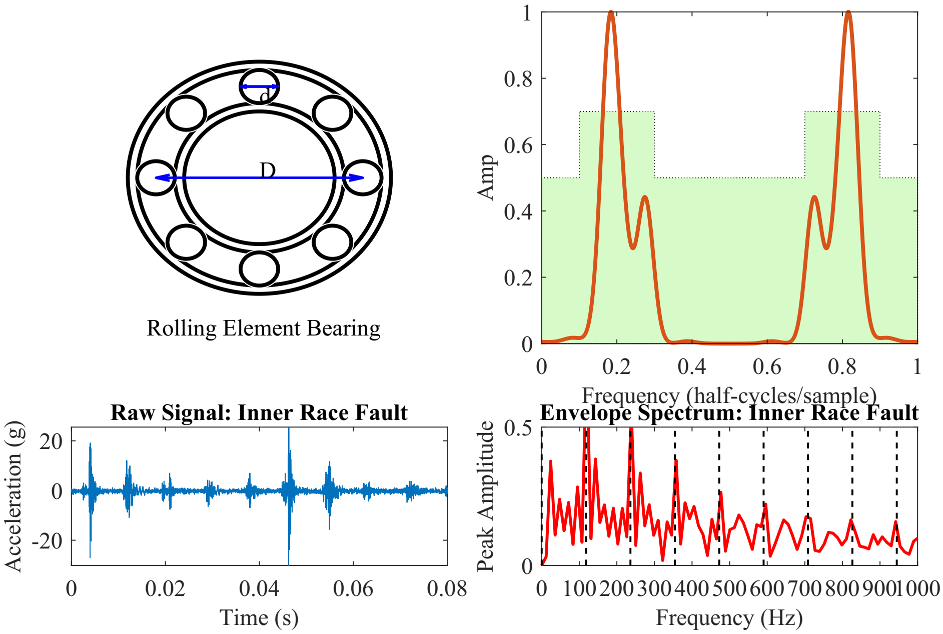


Fig. 2. (top-left) A rolling element bearing; (bottom-left) Simulated vibration signal from inner race fault; (top-right) Second temporal moment of the simulated signal for inner race fault; (bottom-right) Envelope spectrum for the inner race fault signal.

of bearings, and have achieved great success in rolling element bearing fault diagnosis. Vibration signals of fault bearings have specific patterns in the time-frequency domain and envelope spectrum (as shown in Fig. 2). Assume there exists a formal description φ, for the fault signal, written with STL as

$$\varphi = \mathbf{GF}(\mathbf{F}_{[0,0.2]}(M_2 > 0.7) \rightarrow \mathbf{G}_{[0,0.4]}(M_2 < 0.5)), \tag{1}$$

where M_2 indicates the second temporal moment of the signal (shown in top-right of Fig. 2), which calculated with the time-frequency pattern [39]. Then the formal description can be interpreted with plain English as “whenever the value of second temporal moment is larger than 0.7, then it will be smaller than 0.5 within 0.2 half-cycles/sample (normalized) for 0.4 half-cycles/sample, and this pattern will occur periodically”. When the domain experts are given with this formal description, they will know how the monitoring system makes decisions, and this description is consistent with existing knowledge, which comes from the analysis of fault mechanism. The faulting mechanism shows that for constant speed operation, the inner race defect has characteristic frequencies, which are ball pass frequencies derived from the bearing geometry and kinematics under the no-slip assumption, described by equations as [40]

$$BPFI = \frac{n}{2}(1 + \frac{d}{D}\cos\alpha)R_s, \tag{2}$$

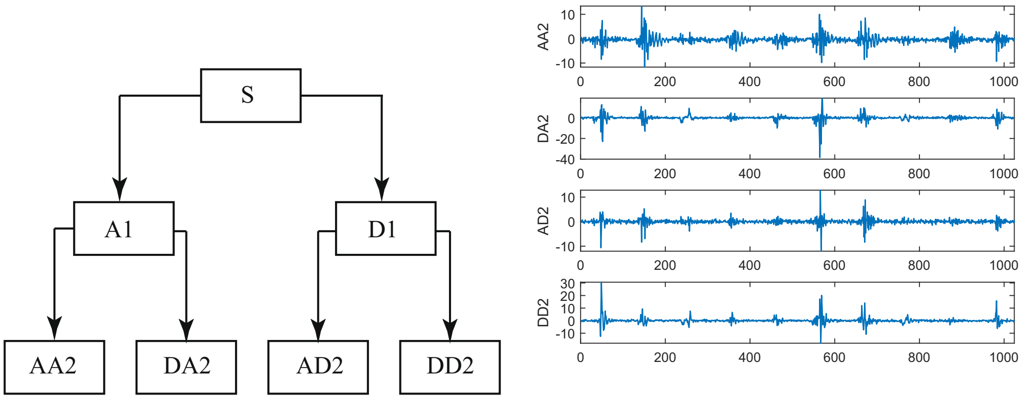
where the geometric parameters $d, D, \alpha, R_s, n$for the rolling element bearing are the diameter of ball/rolling element, pitch diameter, contact angle, the inner race rotation speed, and the number of balls. Therefore, the interpretation of the formal description will help the domain experts trust the decision made by the monitoring system.

3.2. Frequency-temporal Logic

Many fault pattern can only be characterized in the frequency domain. This is especially true for non-stationary signals whose frequency components vary over time. This class of signals motivate the need for FTL, which is an extension of STL that can be used to specify both time and frequency properties of a signal. In FTL, a signal predicate is defined over the signal presenting the evolution of the Wavelet package transform (WPT) coefficient at a particular level over frequency domain. WPT is an expansion of discrete wavelet transform whereby both the approximation and detail coefficients are decomposed. The coefficients resulting $\lambda_n^l(k)$ from the decomposition of a signal $s(t)$ is as [39]

$$\lambda_n^l(k) = \langle s(t), 2^{-l/2} W_n(2^{-l}t - k) \rangle, \tag{3}$$

where n is the node number, l is the decomposition level, k is the position parameter, and W_n is the orthogonal wavelet decomposition coefficients. Here we omit the notation for signal $s(t)$ in $\lambda_n^l(k)$, since in this paper we only deal with one



(a)

(b)

Fig. 3. (a) Level 2 decomposition using wavelet packet transform; (b) Wavelet packet transform coefficients at level 2.

dimensional signals. As shown in Fig. 3, which shows the WPT for signal in Fig. 2. Then the coefficients for $AA2, DA2, AD2$ and $DD2$ can be denoted as $\lambda_3^2(k), \lambda_4^2(k), \lambda_5^2(k)$, and $\lambda_6^2(k)$, respectively.

Definition 1. The set of FTL formulas is denoted as Υ and the segment of FTL syntax is defined recursively as

$$\varphi = \mu_f|\varphi_1 \wedge \varphi_2|\varphi_s \vee \varphi_e|\mathbf{F}_{[\tau_s,\tau_e]}\varphi|\mathbf{G}_{[\tau_s,\tau_e]}\varphi, \tag{4}$$

where τ_s and τ_e are non-negative finite real numbers, indicating the frequency range, μ_f is a predicate over the WPT coefficients, which can be defined as,

$$\mu_f := \Omega(\lambda_n^l, k) \sim \alpha, \sim \in \{\geqslant, <\}, \tag{5}$$

where $\alpha \in \mathbb{R}$ is a constant. The Boolean operators $\wedge, \vee$ and temporal operators $\mathbf{F}, \mathbf{G}$ are the same with STL. $\Omega(\lambda_n^l, k)$ is a signal process function applied to the WPT coefficient to suppress the noise. Here we set $\Omega(\cdot, \cdot)$ as a function, mapping the coefficient to its second temporal moment and can be defined as [41]

$$\Omega(\lambda, \omega) = \frac{1}{P_\lambda(\omega)} \int t^2 P_\lambda(t, \omega) dt, \tag{6}$$

where $P_\lambda(t, \omega)$ is the spectrogram power spectrum of the coefficient λ and uses it as a time-frequency distribution and $P_\lambda(\omega)$ is the marginal distribution. In Eq. (6), the second temporal moment includes the temporal information of the time-frequency distribution, which is why we call the above logic as frequency temporal logic. In the rest of this paper, we use λ_n^l to denote $\Omega(\lambda_n^l, k)$ for short.

Definition 2. The robustness metric ρ maps an FTL formula $\varphi \in \Upsilon$, a WPT coefficient trace $\lambda_n^l \in \Lambda$ and position $k \in P$ to a real value, that is, $\rho : \Upsilon \times \Lambda \times P \to \mathbb{R} \cup \{\infty, -\infty\}$, such that

$$\begin{aligned}
\rho(\mu_f \geqslant 0, \lambda_n^l, k) &= \lambda_n^l - \alpha \\
\rho(\mu_f < 0, \lambda_n^l, k) &= \alpha - \lambda_n^l \\
\rho(\varphi_1 \wedge \varphi_2, \lambda_n^l, k) &= \min\left(\rho(\varphi_1, \lambda_n^l, k), \rho(\varphi_2, \lambda_n^l, k)\right) \\
\rho(\varphi_1 \vee \varphi_2, \lambda_n^l, k) &= \max\left(\rho(\varphi_1, \lambda_n^l, k), \rho(\varphi_2, \lambda_n^l, k)\right) \\
\rho(\mathbf{G}_{[a,b)}\varphi, \lambda_n^l, k) &= \inf_{k' \in [k+a, k+b)} \rho(\varphi, \lambda_n^l, k') \\
\rho(\mathbf{F}_{[a,b)}\varphi, \lambda_n^l, k) &= \sup_{k' \in [t+a, t+b)} \rho(\varphi, \lambda_n^l, k').
\end{aligned}$$

We use $\rho(\varphi, \lambda_n^l)$ to denote $\rho(\varphi, \lambda_n^l, 0)$, which indicates the signals started with $\lambda_n^l(0)$. If $\rho(\varphi, \lambda_n^l)$ is large and positive, then λ_n^l would have to deviate substantially in order to violate φ.

Before we formulate the fault diagnosis task, we present some definitions essential to formulate the problem. To infer the formal descriptions for the time series data, we first introduce the concept of attribute grammar for the formal language, which is defined by a 4-tuple [42] defined as

$$G =< V_N, V_T, P, g >, \tag{7}$$

where V_N is the set of non-terminal nodes, V_T is the set of terminal nodes, P denotes the set of production rules, and g is a relation mapping each node to its attributes. We use capital letters to denote non-terminal nodes, e.g., $A \in V_N$, and use lowercase letters, including Greek ones, to denote terminal nodes, e.g. $a \in V_T$. The attributes of a node are specified by the relation g. For instance, the ordered list of attributes of a non-terminal node A is $g(A)$. In this paper, we will use those attributes that are synthesized [42], meaning that if $A_0 \to A_1A_2$ is a production rule in P (in the parsing tree, A_1 and A_2 are the children of A_0), then $g(A_0)$ is the collection of $g(A_1), g(A_2)$, namely $g(A_0) = g(A_1) \cup g(A_2)$. Then the grammar for the formal language used in this paper is defined as

Definition 3. The *FTL attribute grammar* $\mathcal{G}_{\mathcal{FTL}}$ is an attribute grammar $< V_N, V_T, P, g >$ with the following specific components:

- $V_N = \{A, B\}$, where each element of V_N corresponds to an STL fragment (partial formula);
- $V_T = \{\mu_f, \mathbf{F}, \mathbf{G}, \vee, \wedge\}$, where the meanings of the symbols are the same as those in Eq. (4);
- $P = \{P_1, \cdots, P_7\}$, where the specific production rules are shown in Table 1 (there are five categories of rules, namely *Instance*, *Finally*, *Globally*, *Or*, and *And*);
- g maps each node to two types of attributes: (a) *frequency* attributes that specify the frequency bounds of the spectrum operators used in the node and (b) *predicate* attributes that specify the predicates used in the node. Specifically, the *predicate* attributes includes *WPT level l, and node number n*, *comparison operator*, and *constant*. To give an example, for a terminal node $\mu_f : \lambda_4^2 > 1$, its ordered list of *frequency* attributes is $\mu_f.freq = \{\}$, which is empty, and its set of *predicate* attributes is $\mu_f.pre = \{2, 4, >, 1\}$ (we will use the notations *.pre* and *.freq* throughout the paper). Both types of attributes are synthesized. For instance, production rule $P_5 : A \to A \vee B$ indicates that:

  $$A.freq = A.freq \cup B.freq; \quad A.pre = A.pre \cup B.pre.$$

With the definition of attributed grammar $\mathcal{G}_{\mathcal{FTL}}$, we can conclude following proposition without proof (the proposition is obvious).

Proposition 1. Any FTL formula φ can be derived with the FTL attribute grammar $\mathcal{G}_{\mathcal{FTL}}$.

Example 1. This proposition can be best illustrated with an example as shown in Fig. 4. It can be easily seen that an FTL formula $\varphi = \mathbf{G}_{[0,3]}(\mathbf{F}_{[0,2]}(\lambda_3^2 > 1) \wedge \mathbf{G}_{[0,1]}(\lambda_4^2 < 2))$ can be derived by following a sequence of production rules $P_3P_7P_2P_5P_1P_1$ applied to a set of properly attributed terminal nodes $V_T = \{\mathbf{G}_{[0,3]}, \mathbf{F}_{[0,2]}, \mu_{f1} := (\lambda_3^2 > 1), \mathbf{G}_{[0,1]}, \mu_{f2} := (\lambda_4^2 < 2)\}$.

Definition 4. A *formula* φ is an FTL formula, which is derived by a sequence of production rules $\varepsilon = r_1 \cdots, r_n$ as illustrated by Proposition 1 and Example 1.

3.3. Problem statement

We now formally define the problem of fault diagnosis and interpretation with an FTL formula, which infers an FTL formula with BNN and applied the learned formula to fault diagnosis and interpretation.

Problem 1. (Fault Diagnosis) Given a set of labeled WPT coefficients, $X = X^+ \cup X^-$, which includes a set of coefficients with positive (X^+) and negative (X^-) examples, and a attributed grammar $\mathcal{G}_{\mathcal{FTL}}$ as described in Definition 3, find an FTL formula φ, derived with a sequence of production rules $\varepsilon = r_1 \cdots, r_n$, such that the robustness degree

$$\rho(X, \varphi) = \min(\min_{x \in X^+}(\rho(x, \varphi)), \min_{x \in X^-}(\rho(x, \neq g\varphi))), \tag{8}$$

Table 1
Production rules of $\mathcal{G}_{\mathcal{STL}}$.

Rule	Category	Notation	Attributes
P_1	Instance	$A\|B \to \mu_f$	$g(A\|B) = g(\mu)$
P_2	Finally	$A \to \mathbf{F}A$	$g(A) = g(\mathbf{F}) \cup g(A)$
P_3	Globally	$A \to \mathbf{G}A$	$g(A) = g(\mathbf{G}) \cup g(A)$
P_4	Finally	$B \to \mathbf{F}B$	$g(B) = g(\mathbf{F}) \cup g(B)$
P_5	Globally	$B \to \mathbf{G}B$	$g(B) = g(\mathbf{G}) \cup g(B)$
P_6	Or	$A\|B \to A \vee B$	$g(A\|B) = g(A) \cup g(B)$
P_7	And	$A\|B \to A \wedge B$	$g(A\|B) = g(A) \cup g(B)$

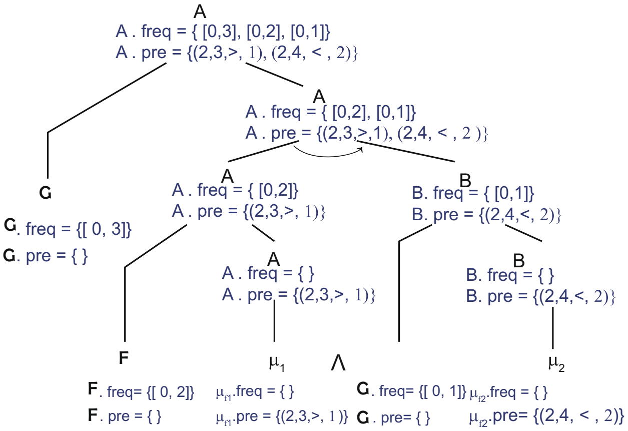


Fig. 4. The construction tree of $\varphi = \mathbf{G}_{[0,3]}(\mathbf{F}_{[0,2]}(\lambda_3^2 > 1) \wedge \mathbf{G}_{[0,1]}(\lambda_4^2 < 2))$. The arc with an arrow indicates the $\wedge$ operator. The *time* and *freq* attributes of a node are shown immediately underneath the corresponding node.

for coefficient set X against φ is maximized, where $\rho(x, \varphi)$ denotes the robustness degree for coefficient sequence x against φ. Here the time series signals are first transformed with WPT, then formula φ is applied to the WPT coefficients for robustness degrees.

Remark 1. The above formula, φ, can be seen as an interpretable classifier, therefore, it has two roles: one is a diagnoser and the other is a decision explanator. It diagnoses the conditions of the bearing and gives explanations for the decision process. Since we use labeled data to train the model, the above problem is a supervised learning problem. All the signals that come from the positively labeled bearings will satisfy the formula if the robustness calculated in Definition 1 is positive and vice versa for the negatively labeled bearings. As FTL can be understood by a human, the fault patterns can be discovered with a human-friendly approach. It is helpful to understand Eq. (8) by visualizing it in terms of Support Vector Machine (SVM): the FTL formula φ defines the boundary between desirable and undesirable behaviors; $\min_{x \in X^+}(\rho(x, \varphi))$ and $\min_{x \in X^-}(\rho(x, \neq g\varphi))$ are the distances between the boundary and the desirable and undesirable behaviors, respectively; and finally maximizing $\rho(X, \varphi)$ results in a boundary that maximally separates the desirable and undesirable behaviors. But compared with the boundaries in SVM, which are hyper-planes in some high dimensional feature spaces, which might be hard to interpret, our boundary is defined by an FTL formula φ, which is easily understandable and obtained without any human intervention.

4. Specification inference with Bayesian neural networks

In Section 3.3, fault diagnosis and interpretation requires learning an FTL formula from scratch. In this section, we will learn the structure of the formula and its parameters (attributes) using Bayesian optimization with BNN.

4.1. Frequency-temporal logic encoder

BNN can only take numerical values, therefore we need an encoder, which maps an FTL formula φ to a vector $v \in \mathbb{R}^m$. In this paper, the encoder function $\pi : \Upsilon \rightarrow \mathbb{R}^m$, where m is the dimension of the vector. Based on some random tests among the bearing vibration data, when m is set to 40, the encoded formula can obtain good performance in fault diagnosis. Moreover, when m is larger than 40, the computational time for the learning algorithm will be increased. As shown in Fig. 5, the vector v can be divided into three parts. The first part denotes the production rules used to generate the formula sequentially, the second part denotes the frequency intervals for the spectrum operators sequentially, and the last prat denotes the predicates used in the formula sequentially. Here we only take the coefficients at Level 3 into consideration, and the encoder table is shown in Table 2. When the length of the formula is smaller than 40, 0 is auto-filled to vector v. Based on the encoder map, FTL formula $\varphi = \mathbf{G}_{[0,3]}(\mathbf{F}_{[0,2]}(\lambda_3^2 > 1) \wedge \mathbf{G}_{[0,1]}(\lambda_4^2 < 2))$ can be encoder to a numerical vector. Since the formula is derived with production rules $P_3P_7P_2P_5P_1P_1$, the first part of vector v is $[3, 5, 2, 3, 1, 1, 0, \cdots]$. Note that the production rules P_2 and P_4, and P_3 and P_5 use the same syntax rule, thus they are encoded as the same number to reduce the search space. When we map a vector to a formula, we can choose one of the production rules accordingly and the structure of the formula is uniquely defined. Then, three frequency bound $[0, 3], [0, 2]$, and $[0, 1]$ are used sequentially, then the second part of vector v is $[0, 3, 0, 2, 0, 1, 0, \cdots]$. Finally, two predicates are used with node number 3 and 4, constant values are 2 and 1, then the third part of v is $[3, 6, 1, 4, 7, 2, 0, \cdots]$. Then the vector is

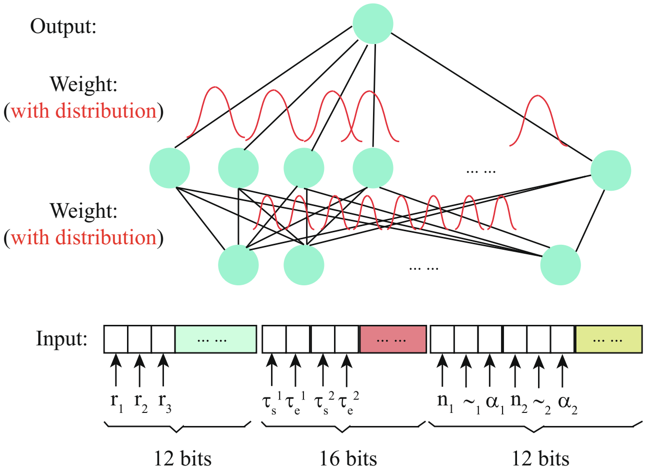


Fig. 5. FTL formula encoder and Bayesian neural network with one hidden layer.

Table 2
Encoder table for FTL formulas.

Category	Encoder	Category	Encoder
P_1	1	τ_s	τ_s
P_2/P_4	2	τ_e	τ_e
P_3/P_5	3	α	α
P_6	4	$\geqslant$	6
P_7	5	$<$	7
Node number	n		

$$v = \begin{bmatrix} 3 & 5 & 2 & 3 & 1 & 1 & \cdots & 0 & \cdots \\ \cdots & 0 & 3 & 0 & 2 & 0 & 1 & \cdots & 0 \\ \cdots & 3 & 6 & 1 & 4 & 7 & 2 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{1\times 40}. \tag{9}$$

4.2. Bayesian optimization with Bayesian neural networks

The goal of Bayesian optimization in this paper is to find an FTL formula φ, such that the robustness degree in Problem 1 is optimal, namely $v^* = \mathbf{argmax}_{v_t \in \mathcal{V}} \rho(\pi^{-1}(v), X)$, where $\pi^{-1}(v)$ is the decoder function that maps a vector to an FTL formula. With the help of the encoder proposed in Section 4.1, this FTL-based fault diagnosis problem can be transformed into an optimization problem. Before we solve this optimization problem, three important properties of this problem should be addressed as follows: 1) The domain of formula $\mathcal{V}$ is a high-dimensional mixed integer space; 2) The objective function $\rho(\pi^{-1}(v), X)$ is highly non-linear and not differentiable; 3) Query the objective function is a time-consuming process if the data set X is large.

To address the first and second properties, here we take advantage of nice properties of neural network and formulate the problem using Bayesian neural networks (see Fig. 6 for architecture) and the problem we solved is as follows: We aim to model the robustness $\rho(v, t) = \rho(\pi^{-1}(v), X, t)$ of a vector $v \in \mathcal{V} \subset \mathbb{R}^{40}$ at time step t based on noisy observation $y(v, t) \sim \mathcal{N}(\rho(\pi^{-1}(v), X, t), \sigma^2)$, where σ is the variance for the noise. Assume we obtain T_n data points for the model; denoting the combined data by $\mathcal{D} = \{(v_1, t_1, y_1), (v_2, t_2, y_2), (v_{T_n}, T_n, y_{T_n})\}$, we can then have the joint probability of the data $\mathcal{D}$ and the network weights W as

$$\mathbb{P}(\mathcal{D}, W) = \mathbb{P}(W)\mathbb{P}(\sigma^2)\Pi_{i=1}^{|\mathcal{D}|}\mathcal{N}(y_i, \hat{\rho}(\pi^{-1}(v_i), X, t_i|W), \sigma^2), \tag{10}$$

where $\hat{\rho}(\pi^{-1}(v_i), X, t_i|W)$ is the prediction of a neural network. Even though computing the posterior weight distribution $\mathbb{P}(\mathcal{D}, W)$ is an intractable task, we can use Markov chain Monte Carlo (MCMC) to sample it, in particular stochastic gradient MCMC methods, such as Stochastic Gradient Langevin Dynamics (SGLD) [43] or Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [44]. Given M samples $W^1, W^2, \cdots, W^M$, we can then obtain the mean and variance of the predictive distribution $p(\rho|v, t, \mathcal{D})$ as

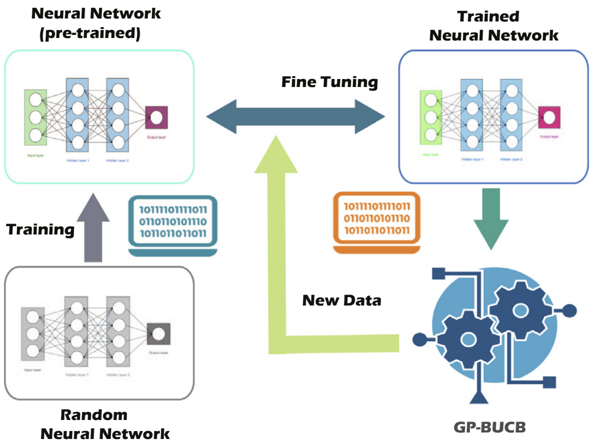


Fig. 6. Gaussian process batch upper confidence bound based transfer learning framework for FTL formula inference.

$$\begin{aligned}
\hat{\kappa}(v,t|\mathcal{D}) &= \frac{1}{M}\sum_{i=1}^{M}\hat{\rho}(\pi^{-1}(v_i),X,t_i|W^i),\ \text{and} \\
\hat{\sigma}^2(v,t|D) &= \frac{1}{M}\sum_{i=1}^{M}(\hat{\rho}(\pi^{-1}(v_i),X,t_i|W^i)-\hat{\kappa}(v,t|\mathcal{D})^2,
\end{aligned} \tag{11}$$

respectively. We will write their shorthand as $\hat{\kappa}(v,t)$ and $\hat{\sigma}^2(v,t)$ for the rest of this paper.

To address the third property of the problem, we need to find the optimal formula with limited samples. Therefore, we combine the Gaussian process batch upper confidence bound (GP-BUCB) [45] and transfer learning method to find the optimal formula. With the estimation for the mean and variance, we can find the optimal formula by sampling the formula space $\mathcal{V}$ sequentially. To simplify the sampling process, assuming the batch size is M, we just run the Gaussian process upper confidence bound (GP-UCB) one time and choose the M largest values for each sampling round. The GP-UCB algorithm samples the next sample v_{t+1} with the following rule,

$$v_{t+1} = \mathbf{argmax}_{v_t\in\mathcal{V}}\hat{\kappa}(v,t) + \beta^{1/2}\hat{\sigma}^2(v,t), \tag{12}$$

then the optimal FTL formula $\varphi = \pi^{-1}(v_{t+1})$ is used to calculate the robustness, y_{t+1}, for training the BNN. Moreover, in order to speed up the training process, we use the transfer learning framework to make use of the historical neural network parameters, which is shown in Fig. 6. In this framework, the BNN is first randomly initialized and trained with an initial set of data, then the pre-trained BNN is saved for fine-tuning. During each batch sampling and training process, the pre-trained BNN will be loaded and trained with the new data set (a combination of the newly sampled data set and old data set), then the new trained BNN will be saved for next run of batch sampling. The learning algorithm can now be summarized in Algorithm 1.

Algorithm 1 Bayesian Optimization with Bayesian Neural Networks.

```
Input: Formula width limit $W$; Limit for number of samples in sampling space $\mathcal{V}$, $L$; Initial sample number $T_0$;
  Maximal sampling step $T$; Batch size $M$.
Output: A set of sample-observation pairs $\{(v_i,\mathbf{y}_i)\}_{i=1}^{TM+T_0}$ and the optimal sample $v^*$.
Initialize the formula space $\mathcal{V}$;
Initialize the training set $\mathcal{D}_0 = \{v_i,y_i\}_{i=1}^{T_0}$;
Initialize the BNN with random parameters;
Train the BNN with the initial training set $\mathcal{D}_0$ and save the training result.
repeat
    Load the pre-trained BNN;
    Update $\hat{\kappa}(v,t)$ and $\hat{\sigma}^2(v,t)$ with Eq. (11);
    Sample the formula space $\mathcal{V}$ to get M sample pairs $\mathcal{B}_t = \{v_i,y_i\}_{i=1}^{M}$ using GP-UCB strategy shown in Eq. (12);
    Update training set $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1}\cup\mathcal{B}_t$;
    Fine tune the BNN with the new training set $\mathcal{D}_t$ and save the training result;
    $t \leftarrow t+1$;
until $t > T$.
```

During the learning process, to generate a valid set $\mathcal{V}$ in Line 1 in Algorithm 1, a random tree generate algorithm is used to generate random parsing trees for FTL formulas. Then terminal nodes' attributes of the trees are assigned with random number accordingly. Last, the set $\mathcal{V}$ can be obtained by mapping the generated trees to vectors. In Line 2, we cut each state's signal into 600 pieces and applied the WPT with a depth of level 3 to the signals to obtain their wavelet packet transform. Then we calculate the second temporal moment for each coefficient. With these signals, we construct the training data sets for each bearing condition (data set construction detail will be discussed in Section 6). Finally, for each training set, given a formula vector v_i, we calculate a robustness degree y_i. Line 3 initializes BNN's parameters with random matrices (random weight matrices). In Line 4, the training step number is much larger than the step number in Line 10, as the parameters in Line 10 are pre-trained. The performance of the algorithm can be described with the following theorem.

Theorem 1. *If the optimal robustness degree from solving problem in Definition 1 is $\hat{\kappa}$, and the robustness degree without noise is κ, pick $\delta \in (0,1)$ and set $\beta_t = 2\log(|\mathcal{V}|\pi_t/\delta)$, where $\sum_{t\geqslant 1}\pi_t^{-1} = 1$. If we sample the new training points with the strategy proposed in Eq. (12) [46], then after T steps of training, we have*

$$\mathbb{P}(|\hat{\kappa} - \kappa| < \beta_T^{1/2}\sigma(v, T-1)) \tag{13}$$

holds with probability $\geqslant 1 - \delta$.

Proof. For $v \in \mathcal{V}$ and $t \geqslant 1$. It is known that conditioned on $\mathcal{D} = \{(v_1, t_1, y_1), (v_2, t_2, y_2), (v_n, T_n, y_{T_n})\}$ is deterministic. Further, we assume that $\hat{\kappa} \sim \mathcal{N}(\hat{\kappa}(v,t), \hat{\sigma}(v,t))$ as defined in Eq. (11). Now if $r \sim \mathcal{N}(0,1)$, then

$$\begin{aligned}
&\mathbb{P}(r > c) = e^{-c^2/2}(2\pi)^{-1/2}\int e^{-(r-c)^2/2 - c(r-c)}dr \\
&\leqslant e^{-c^2/2}\mathbb{P}(r > 0) = (1/2)e^{-c^2/2},
\end{aligned}$$

for $c > 0$, as $e^{-c(r-c)} \leq 1$ for $r \geqslant c$. Set $r = (\kappa - \hat{\kappa})/\hat{\sigma}(v,t))$ and $c = \beta_t^{1/2}$. Then

$$\mathbb{P}(|\kappa - \hat{\kappa}|/\hat{\sigma}(v, t-1) > \beta_t^{1/2}) \leqslant e^{-\beta_t/2}.$$

Then

$$|\kappa - \hat{\kappa}| \leqslant \beta_t^{1/2}\hat{\sigma}(v,t) \quad \forall v \in \mathcal{V}$$

holds with probability $\geqslant 1 - e^{-\beta_t/2}$. Since $\beta_t = 2\log(|\mathcal{V}|\pi_t/\delta)$ and $|\mathcal{V}|$, the infinity norm, is less than or equal to 1 with $\pi_t = \pi^2 t^2/6$. Thus, the statement holds.

Remark 2. Before conducting Algorithm, the vectors in search space $\mathcal{V}$ are first normalized by its infinity norm $|v|$ to guarantee that the scaled search space $\mathcal{V}'$ satisfies $|\mathcal{V}'| \leqslant 1$. The above theorem indicates that when the sampling step T increases, the bias of the estimation will decrease, since with more and more samples, the variance $\sigma(v, T-1)$ will decrease.

5. Properties of interpretation with frequency temporal logic

Real-time monitoring with FTL formulas not only focuses on the Boolean satisfaction of time-frequency properties of the vibration signals, but also gives quantitative semantics to the conditions of the bearings, i.e., with semantic of FTL, we can investigate the distance from a faulty system to a health system with robustness degree. Learning the parameters of FTL formulas with BNN assumes there exists Gaussian noise in the observations, which induces a noise resistance property of the proposed method. In this section, we will investigate how the variable speed operations of bearings and noise affect the performance of the proposed fault diagnosis method.

5.1. Interpretation with variable speed operations

For constant speed operation, the inner race defect, outer race defect and ball (roller) spin defect have characteristic frequencies, which can be found in [40], showing that the periodic impulses induced by the defect have fixed but different periods for the three type of defects. Therefore, the time parameters in Eq. (6) can distinguish three types of faults uniquely. In other words, the fixed periodic properties can be easily described with FTL formulas. Unfortunately, many industrial applications are characterized by harsh and variable operating conditions, with multiple heterogeneous vibration sources [47]. Therefore, the signals often present different components, among which the damage related ones are not always dominant and the time intervals of the impulses are not constant, thus whether the signals from three type of fault are distinguishable is undetermined with temporal information. For the fault diagnosis of bearings with variable speed operation, the mapping function $\Omega(\cdot)$ in the definition of robustness should be modified to deal with the variable speed operations. Many signal process functions can be used as the mapping function. For instance, the cepstrum pre-whitening operator for the signal in [47], in which a signal x, is reconstructed by the following equation,

$$x' = IFT\left\{\frac{FT(x)}{|FT(x)|}\right\}, \tag{14}$$

where x' is the reconstructed signal, FT denotes the Fourier transform and IFT denotes its inverse operation. The transformation in Eq. (14) is equivalent to pre-whitening the signal in cepstral domain. In the cepstral domain [48], the periodicity of the spectrum results in a peak at a quefrency equal to the period of the base frequency of the multi-harmonic vibration. The pre-whitening operation sets a zero value for the whole real cepstrum (except possible at zero quefrency), and transforms back to frequency domain. This is equivalent to a series of liftering operations around the quefrencies of the deterministic excitation, resulting in the almost complete deletion of their effect on signals and a removal of resonance effects. Therefore, the variation of the base frequency caused by the time-varying speed conditions will be suppressed.The quality of the signals has been demonstrated to have good performance with examples in [47]. After this operation, with x', the second temporal moment can be computed for FTL. Other signal processing techniques, such as tacholess order tracking [49] and the concentration of frequency and time [50], can also be used to define the mapping function. Therefore, the proposed logic based fault diagnosis method is compatible with other signal analysis techniques.

5.2. Interpretation with noise

How to deal with noise is an essential problem to guarantee highly reliable condition monitoring in industrial environments. Every FTL formula defines a hyperplane in the high dimensional time-frequency space. The existing of noise will affect the estimation of the hyperplane and the fault diagnosis results. For instance, the energy of a frequency component in the faulty signal can be affected by the noise, thus the distance of the signal to the hyperplane defined by an FTL formula will be changed, which may lead to miss-diagnosis. In this paper, the FTL formulas are learned with BNN, in which the estimation of robustness degree is based on the estimation of the mean robustness degree. In other words, if the noise has zero mean Gaussian distribution, we can get an unbiased estimation of the optimal hyperplane. Given a robustness degree of a vibration signal associated with an FTL formula, properties about the confident bound can be described with the following theorem.

Theorem 2. *Given a WPT coefficient x, and an FTL formula φ, if the noise S_n of the vibration signal in time-frequency domain is Gaussian with distribution $S_n \sim \mathcal{N}(0, \sigma_n)$, where σ_n is the variance of the noise, then the probability of miss-diagnosis (MD) is given by*

$$\begin{aligned} \mathbb{P}(MD) &= \begin{cases} \mathbb{P}(\bar{\kappa} > 0), & \text{when } \kappa < 0 \\ \mathbb{P}(\bar{\kappa} < 0), & \text{when } \kappa > 0 \end{cases} \\ &\leqslant \begin{cases} \frac{\delta}{2\sqrt{2\pi\sigma_n}} \int_{-\infty}^{\kappa+\eta} exp(-\upsilon^2/(2\sigma_n))d\upsilon, \\ \frac{\delta}{2\sqrt{2\pi\sigma_n}} \int_{\kappa-\eta}^{+\infty} exp(-\upsilon^2/(2\sigma_n))d\upsilon, \end{cases} \end{aligned} \tag{15}$$

where κ is the measured robustness degree for the coefficient x associated with formula φ, and $\bar{\kappa}$ is the real robustness. $\eta = \beta_T^{1/2}\sigma(v, T)$ and δ are defined in Theorem 1.

Proof. When $\kappa < 0$, set $\kappa = S_n + \bar{\kappa} + \varepsilon$, where ε is the estimation error of robustness degree, then we have $\mathbb{P}(\bar{\kappa} > 0) = \mathbb{P}(\kappa - S_n - \varepsilon > 0) = \mathbb{P}(S_n < \kappa - \varepsilon) \leqslant \mathbb{P}(S_n < \kappa + \eta)\mathbb{P}(-\varepsilon < \eta)$. Assume the estimation error of robustness degree is zero mean, and according to Theorem 1, we have $\mathbb{P}(-\varepsilon < \eta) \leqslant \delta/2$, then $\mathbb{P}(S_n < \kappa + \eta)$ $\mathbb{P}(-\varepsilon < \eta) \leqslant \frac{\delta}{2\sqrt{2\pi\sigma_n}} \int_{-\infty}^{\kappa+\eta} exp(-\upsilon^2/(2\sigma_n))d\upsilon$. When $\kappa > 0$, the proof process is similar to $\kappa < 0$.

Theorem 2 indicates that the miss-diagnosis probability is bounded, and the larger the variance, the larger the miss-diagnosis probability.

6. Case study

In this section, we will use a real experiment to demonstrate the validation of the proposed method. In the experiment, we will use the proposed method to diagnose the fault for a set of rolling element bearings. To get the training and testing data, the single pitting faults were introduced to the surface of the race or the rolling body of the bearings by electrical-discharge machining method. The signals are processed with Matlab and Python environments, and the codes for the proposed method can be found at.https://github.com/Gangchen01.

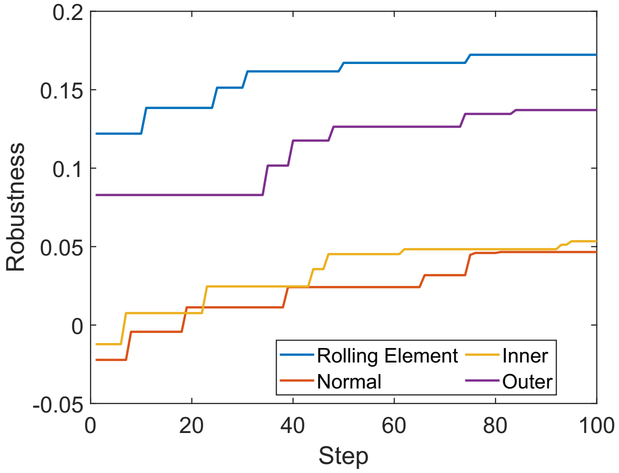


Fig. 7. Maximum robustness obtained for each sampling step during learning process.

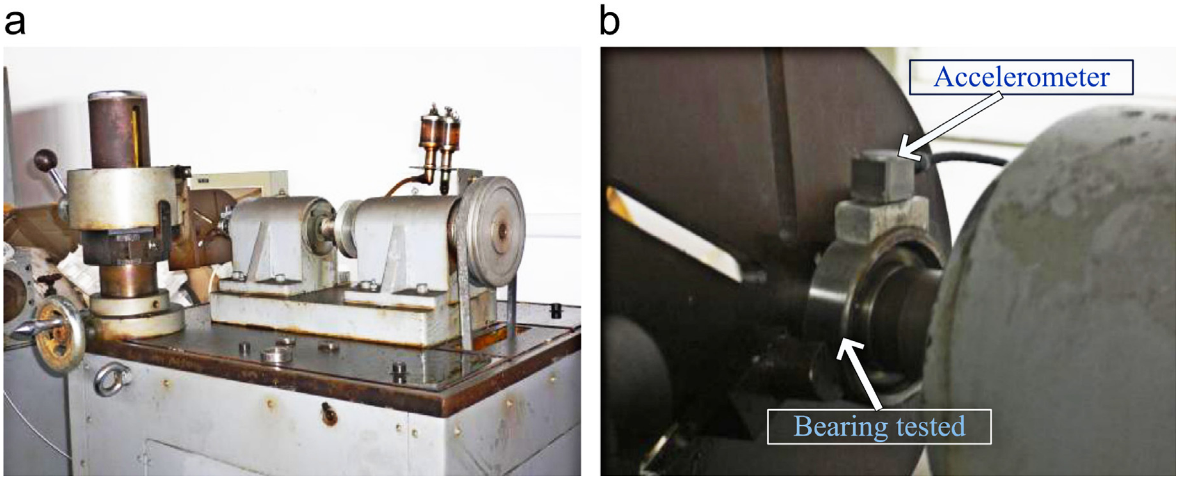


Fig. 8. (a) The rolling element bearing test rig and (b) the location of the accelerometer.

6.1. Interpretation with time-invariant rotational speed conditions

6.1.1. Experiment setup

The data generation test rig is similar to [14] as shown in Fig. 8. In the test rig, an a.c. motor drove the shaft of the rotational machine through a rub belt and a shaft coupling. The tested bearing is assembled on the shaft, whose outer race is fixed by a fixture and inner race rotated with the shaft. The data acquisition (DAQ) system for the test rig is based on NI PXI system (a NI PXI-1042 chassis with NI PXI-4472 modules). The sensor used in the experiment is an accelerometer (Kistler 8791A250), which is located on a bracket by an adhesive mounting. During the test, a series of GB203 rolling element bearings are used to collect the data and the faults are introduced to the surface of the race or the rolling element by electrical-discharge machining method. The speed of the shaft is 1800 r/min and the sampling rate is 12 kHz.

There are four kinds of states for the bearings, namely normal, rolling element fault, inner race fault, and outer race fault. We have collected three signals for each condition, and each of them has a length of 68 s, then cut each condition's signal into 600 pieces (0.34 s and length of 4096 for each) and applied the WPT with a depth of level 3 to the signals to obtain their wavelet packet transform coefficients. There are 8 components at level 3 and each of them has a length of 512. Finally, we calculate the second temporal moment for each coefficient. Therefore, we get 600 signal pieces for each bearing state, and each of the signal has 8 dimensions (some of the signals' trajectories can be found in Fig. 9). Then we construct the labeled set X for robustness calculation in Eq. (8). To construct the positive set for inner fault, 450 pieces of inner fault signals are used, and the negative set from the other three conditions' signal (150 signals for each). Therefore, the size of X is 900. Moreover, the labeled sets for normal, outer race fault and rolling element fault are constructed with the same method as for inner fault. To test the performance of the FTL formula in fault diagnosis, we also construct a test set for each bearing con-

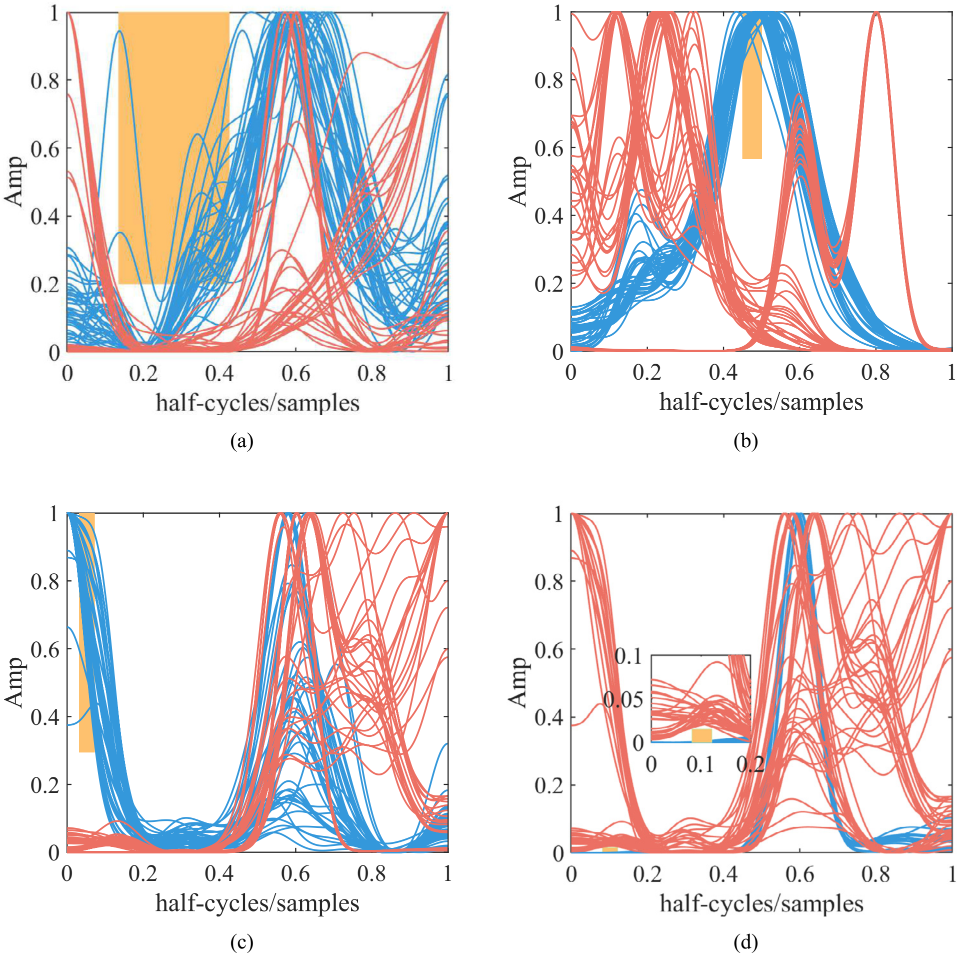


Fig. 9. (a) Visualization of FTL-based fault diagnosis for a) inner race fault,b) outer race fault, c) rolling element fault and d) normal state, where the blue trajectories are positive examples, the red trajectories are nagtive examples and the yellow regrions are defined by the FTL formulas (only the blue trajectories are allowed to pass the yellow region).

dition. For inner fault condition, the positive test examples are the rest 150 pieces, and the negative test examples come from the other three bearing conditions (50 pieces for each condition that un-used for training).

To initialize the learning algorithm in Algorithm 1, we initialize the sampling space $\mathcal{V}$ with 5000 vectors, each of which is associated with a valid FTL formula and has a width of 40. Also, we randomly sample space $\mathcal{V}$ and calculate the robustness to get the training set $\mathcal{D}_0$, whose size is set to 200. In this example, the BNN has 2 hidden layers with 100 nodes for each, and the initial BNN is trained by 1000 training step, and the fine-tuning step limit is set to 100. Moreover, the batch size is 20 and sampling step limit is set to 100.

6.1.2. Interpretation results

Fig. 7 shows the robustness obtained by Algorithm 1 for each sampling step. These results indicate that the proposed algorithm can reach positive robustness within limited sampling steps for four scenarios, indicating the obtained FTL formula can diagnosis the fault correctly among the training set. Also, the results show that the sampling strategy may not improve the performance of the obtained formula in fault diagnosis every sampling step, which due to the uncertainty of the algorithm.

Table 3 shows the learning results for faults diagnosis with FTL formulas. We can see the formulas are compact. Three reasons can explain this compact: 1) Frequency-temporal-logic is expressive in specifying the pattern of fault signals. The eventually (**F**) and always (**G**) operators offer much freedom for the signals. e.g., **F** allows the fault pattern to occur at any time within a frequency range. 2) The learning algorithm can choose the best dimension of the WPT results, which can distinguish the fault from the others. 3) These compact formulas are enough to diagnose the faults. Of course, if the fault pattern

Table 3
Interpretation formulas for the real experimental vibration signals for inner race, outer race and rolling element faults.

Fault Type	Interpretation
Inner Race	$\mathbf{F}_{[0.060,0.15]}(\mathbf{F}_{[0.13,0.43]}(\lambda_3^6 > 0.2))$
Outer Race	$\mathbf{F}_{[0.36,0.45]}(\mathbf{G}_{[0.040,0.090]}(\lambda_3^7 > 0.567))$
Rolling Element	$\mathbf{F}_{[0,0.021]}(\mathbf{G}_{[0.30,0.67]}(\lambda_3^6 \geqslant 0.945) \vee \mathbf{F}_{[0.030,0.060]}(\lambda_3^2 \geqslant 0.295))$
Normal	$\mathbf{G}_{[0.53,0.57]}(\lambda_3^6 < 0.756) \vee \mathbf{F}_{[0.080,0.12]}(\lambda_3^2 < 0.015))$

is more complex, the algorithm has the ability to increase the complexity of the formula by extending the width of the encoder in Section 4.1. These formulas can be explained in plain English as follows:

- **Inner race fault**. The sixth component of the WPT at level 3, denoted as λ_3^6 is used to distinguish the inner race fault from the other conditions. The FTL formula indicated the pattern for the inner race fault, that the second temporal moment of λ_3^6 should be finally larger than 0.2 between frequency 0.13 and 0.43 should occur at least once between frequency 0.06 and 0.15, where the frequency is normalized (as shown in Fig. 9(a)). The figure indicates that the inner race fault is different from the other conditions by having larger energy at a frequency around 0.3.
- **Outer race fault**. The seventh component of the WPT at level 3, denoted as λ_3^7 is used to distinguish the outer race fault from the other conditions. The FTL formula indicates the pattern for the outer race fault, that the second temporal moment of λ_3^7 should be globally larger than 0.567 between frequency 0.04 and 0.09 should occur at least once between frequency 0.36 and 0.45 (as shown in Fig. 9(b)). The figure indicates that the outer race fault is different from the other conditions by having larger energy at a frequency around 0.5.
- **Rolling element fault**. The second and sixth components of the WPT at level 3, denoted as λ_3^2 and λ_3^6 are used to distinguish the Rolling element fault from the other conditions. The FTL formula indicates the pattern for the rolling element fault, that the second temporal moment of λ_3^6 should be globally larger than 0.945 between frequency 0.3 and 0.67 or the second moment of λ_3^2 should be finally larger than 0.295 between frequency 0.03 and 0.06, should occur at least once between frequency 0 and 0.021 (as shown in Fig. 9(c)). The figure indicates that the inner race fault is different from the other conditions by having larger energy in the low-frequency region.
- **Normal**. The second and sixth component of the WPT at level 3, denoted as λ_3^2 and λ_3^6 are used to distinguish the normal state from the other conditions. The FTL formula indicates the pattern for the normal state is that the second temporal moment of λ_3^6 should be globally smaller than 0.756 between frequency 0.53 and 0.57 or the second moment of λ_3^2 should be finally smaller than 0.015 between frequency 0.08 and 0.12 (as shown in Fig. 9(d)). The figure indicates that the normal condtion is different from the other conditions by having a smaller energy at frequency around 0.1.

The formal interpretation results show that the fault conditions will lead to a larger energy concentration at some frequency bands, and the different fault will have different energy concentration bands. The results are in line with the fault mechanism that the fault signals come from the strikes of rollers on the fault surface and excite the resonant frequencies. Moreover, for fixed rotational speed condition, the energy concentration for the healthy bearing is bounded to a small region. Based on this interpretation, technicians can understand why fault occurs and actions that can avoid energy concentration will reduce the risk of the fault occurring, and energy concentration usually is caused by strikes. Therefore, targeted maintenance actions, such as adding lubrication, will reduce the risk of energy concentration.

The faults diagnosis results for the experimental data are shown in Table 4. The results show that the proposed method has good performance for rolling element bearing fault diagnosis, which the errors are less than 5% among the test data set. When the found formula can obtain positive robustness, the formula can diagnose the fault correctly. If the found formula can obtain negative robustness, the formula may lead to mis-diagnosis. Inner race and normal have negative robustness among testing data sets in Table 4 due to the randomness of noise or the different distribution of the fault pattern between training and testing data. There exists a high probability that the trained FTL formula can not obtain good performance for

Table 4
Performance of the interpretation formulas in fault diagnosis in terms of robustness and classification error for real bearing signals.

Fault Type	Robustness		Error Rate	
-	Training	Testing	Training	Testing
Inner Race	0.053	−0.033	0.000	0.042
Outer Race	0.137	0.223	0.000	0.000
Rolling Element	0.170	0.013	0.000	0.000
Normal	0.046	−0.013	0.000	0.032

some specifical cases. However, the overall performance is good among training and testing data sets. The intuitive representations of the formulas against the vibration signals are shown in Fig. 9. In Fig. 9, the blue trajectories indicate the positive examples, and the red trajectories indicate the negative examples. The yellow rectangles are defined by the FLT formulas. For instance, in Fig. 9(a), the formula indicates that the blue trajectories should be in the rectangles at least once, while the red trajectories should be never in the yellow region; in Fig. 9(b) indicates that the blue trajectories should be always in the yellow region, while the red trajectories should be not always in the yellow region. Moreover, the trajectories also show that the second temporal moment of the coefficients is smooth. This property is caused by the integration operator over spectrogram power spectrum. When the noise is white, the spectrogram is expected to have constant value after mapped by function defined in Eq. (6). Therefore, white noise does not affect the shape of the second temporal moment, and the smoothness of the second temporal moment can be an indicator for noise resistance property.

In order to demonstrate the effectiveness and efficiency of the method approach over state of the art methods, here we compare the performances of the proposed method with the methods developed in [51,19,20,52]. The proposed method and method in [51] are formal method based, and both methods can use formal formulas to diagnosis the faults. The comparison result is shown in Table 5. An eight-core HP desktop was used. Moreover, during the comparison, we controlled the number of sampling round in our method, and the number of learning cycles in the method developed in [51]. It can be seen that, with the former method, a satisfactory STL formula can be derived in 815 s, while such a formula cannot be obtained with the latter method within roughly the same number of seconds (the error is 11% and the robustness degree $\rho(\varphi, X)$ is still negative after 819 s).

The comparison results between the proposed method with the feature-based methods are shown in Table 6, where the formula based method is compared with Fisher, cosine similarity metric (CSM) and Genetic Algorithm (GA) based SVM methods. The results show that the proposed method have better overall performance than the other methods, and only the diagnosis for normal state has worse performance than the CSMSVM method (0.0167 vs 0.0067).

Table 5
Comparison results between our proposed method and the method in [51] for rolling element fault.

Our method			Method in [51]		
Time	Error	Robustness	Time	Error	Robustness
(s)	%	$\rho(X,d)$	(s)	%	$\rho(X,d)$
191	30	−0.357	188	50	−0.526
392	16	−0.243	396	24	−0.253
815	0	0.045	819	11	−0.168

Table 6
Fault diagnosis error rate with the methods in [19,20,52] and the method in this paper for rolling element bearing.

Fault Type	Fisher + SVM[20]	CSMSVM[19]	GA + SVM[52]	Proposed method
Inner Race	0.113	0.037	0.043	**0.007**
Outer Race	0.043	0.013	0.057	**0.000**
Rolling Element	0.013	0.000	0.007	**0.000**
Normal	0.063	0.007	0.063	**0.017**

The bold numbers come from the proposed method and have better performance.

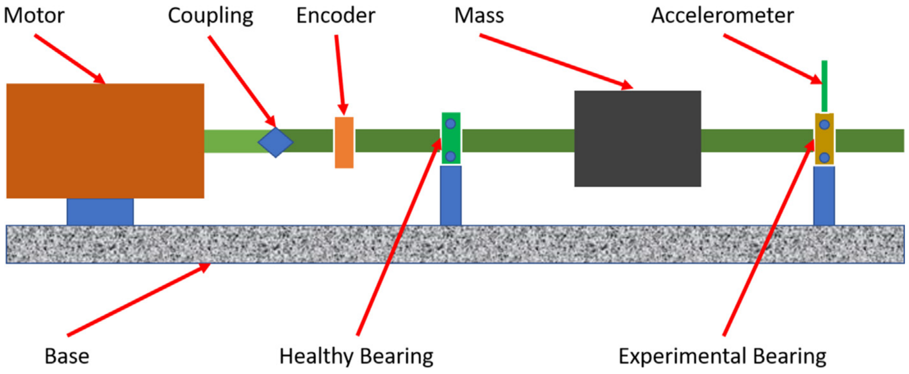


Fig. 10. Diagram for experimental setup under variable speed condition in [53].

6.2. Interpretation with variable speed conditions

In order to illustrate the properties of interpretation with FLT, i.e., interpretation with variable speed and noise conditions. We apply the proposed method to a data set in [53], which is collected under time-varying rotational speed conditions.

6.2.1. Experiment setup

The experiment setup is introduced in [53] and the diagram is shown in Fig. 10. The shaft is driven by a motor and the rotational speed is controlled by an AC drive. Two ER16K ball bearings are installed to support the shaft. The left one is a healthy bearing and the right one is the experimental bearing. The accelerometer is placed on the housing of the experimental bearing to collect the vibration data, where the NI data acquisition boards are used. During the experiment, the original data was sampled at 200,000 Hz and we re-sample the data to alleviate the computation load, which makes the data used in this paper is sampled at 20,000 Hz. The rotation speed for the data used in this paper is about 10 Hz to 28 Hz. One piece of the inner fault signal (left) and outer race fault signal (right) are shown in Fig. 11. Fig. 11 also shows the cepstrum pre-whitening results and their second temporal moments. The second temporal moments of the two signals are significant different, thus they are good signals for fault diagnosis.

The data sets used to train the model are from the inner race fault and outer race fault bearings [1]. We first cut each state's signal into 600 pieces (0.25 s for each), then to construct the training set, 400 pieces of inner fault signals are used as positive examples, and 400 pieces of outer race fault signals are used as negative examples. To construct the testing set, the other 200 pieces of inner race fault signals and outer race signals are used. Therefore, the size of X in this experiment is 800 for the training set. Before the tests, we apply Eq. (14) to reconstruct the signals in training and testing sets.

To investigate the effect of WPT's depth to interpretation results under time-varying rotational speed condition, we conduct three tests, which applies the WPT to the training and testing signals with a depth of level 2, 3 and 4, respectively. If considering the vibration signals to be true signals without noise, the signals for training and testing will have an average signal-to-noise (SNR) ratio 1.69, 0.422 and 0.105, respectively. Then we apply the proposed method to the decomposed signals for fault diagnosis. To investigate the effect of noise to interpretation results under time-varying rotational speed condition, we conduct another three tests, in which a white noise signal is added to each training signal and testing signal with zero mean and three different variances, namely 0.05, 0.1, and 0.2 respectively. During the noisy tests, the WPT is applied with a depth of level 3. This experiment also compares the proposed method with the algorithm in [19] under different noise conditions.

6.2.2. Interpretation results

Table 7 shows the results for variable speed conditions. The comparison of different decomposition depth for WPT indicates that the depth of level 3 can obtain the best robustness and fault diagnosis performance. A higher level (level 4) of decomposition cannot guarantee a better result, which leads to negative robustness among training and testing data. The deeper of the WPT level, since the fault information will be distributed among all the sub-component signals, the fault information among each sub-component will be less. It will be harder for the FTL formula to capture the fault properties with limited formula complexity. The results for the noise resistance tests indicate that adding of white noise do not affect of the performance for fault diagnosis greatly, which obtains high fault diagnosis rate among testing data for all noise level, revealing the noise resistance property for the proposed method. Since the noise is zero mean, and the mapping function used in FTL (denoted in Eq. (6) is the second temporal moment. Ideally, adding white noise will add a constant bias to the second temporal moment but do not affect the shape of the second temporal moment. The bias will not affect the fault diagnosis results since the learned formulas only need to add the bias to the predicates and all the fault modes share the same bias. For example, if the fault properties can be captured by formula $F_{[0,0.15]}(x > 0.5)$, and the noise introduces a bias of δ, then a new formula $F_{[0,0.15]}(x > 0.5 + \delta)$ will capture the fault properties. The comparison between the proposed method and the method in [19] demonstrates this analysis since the proposed method outperforms the other method under noisy conditions. Moreover, to deal with the noise, statistical hypothesis testing is conducted in the following, which further demonstrates the noise-resistance property of the proposed method.

Fig. 12 shows the visualization for formula obtained when $\sigma_n = 0.05$, where we select 80 of the trajectories in training set randomly for intuitive visualization, and the formula is as follow,

$$\varphi = \mathbf{F}_{[0.082,0.11]}((\mathbf{F}_{[0,0.076]}(\lambda_3^2 > 0.11) \wedge (\mathbf{F}_{[0,0.45]}(\lambda_3^5 > 0.44)) \tag{16}$$

Formula in Eq. (16) defines a cuboid as shown in Fig. 12. All the blue trajectories have points within the cuboid, while all the red trajectories should not have points within the cuboid. The formula indicates that the inner race fault is different from the outer race fault by having larger λ_3^2 and λ_3^5 in the low-frequency region.

6.2.3. Statistical hypothesis testing

Theorem 1 shows the proposed method can find a formula that can guarantee the error is bound with a given probability. The given probability can be seen as a confidential bound, e.g., the found formula is the optimal one with a probability larger than 0.95. Based on this theorem, when the signals are contaminated with white noises, the proposed method can find the optimal formula with a high probability, even though the optimal formula may lead to negative robustness. In this case, we

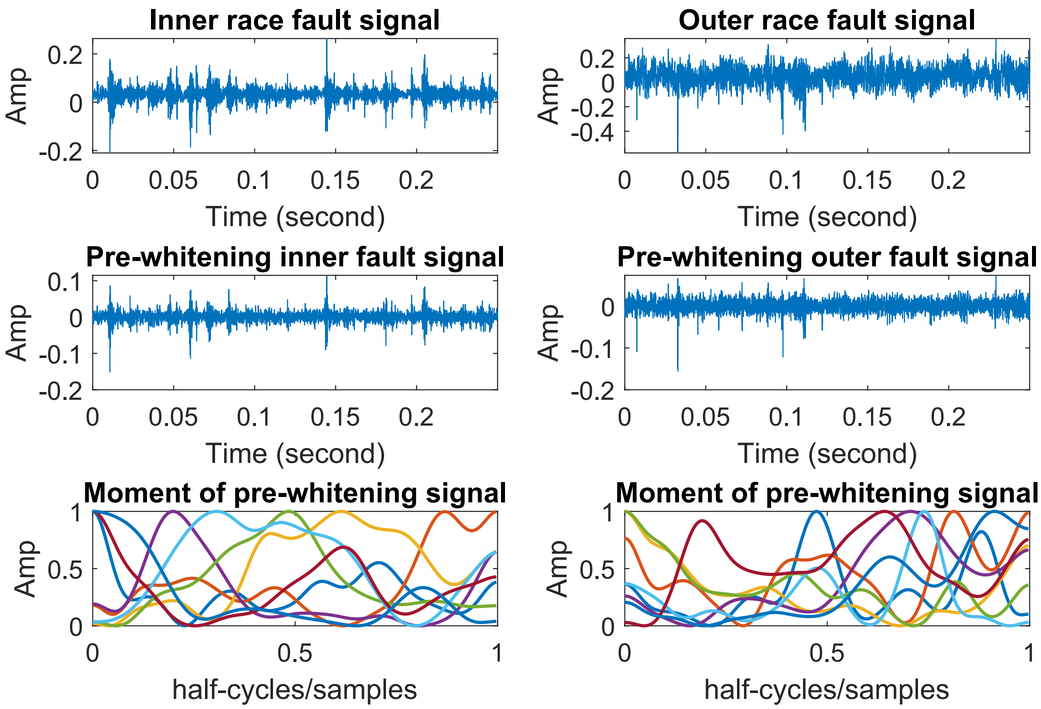


Fig. 11. Example of signal under variable speed conditions.

Table 7
Fault diagnosis error rate of the proposed method and its comparison under different experiment settings.

Experiment Setting	Statistical Hypothesis Testing		Proposed Method		CSMSVM [19]	
–	False alarm	Missing fault	Training	Testing	Training	Testing
$WPT-2$	–	–	0.024	**0.025**	–	–
$WPT-3$	–	–	0.000	**0.023**	–	–
$WPT-4$	–	–	0.041	**0.045**	–	–
$\sigma_n = 0.05 \quad (SNR = 1.69)$	0.000	0.010	0.018	**0.020**	0.031	0.045
$\sigma_n = 0.1 \quad (SNR = 0.422)$	0.005	0.013	0.028	**0.038**	0.028	0.058
$\sigma_n = 0.2 \quad (SNR = 0.105)$	0.030	0.035	0.043	**0.053**	0.055	0.070

Testing results are important for the machine learning algorithm, thus they are in bold.

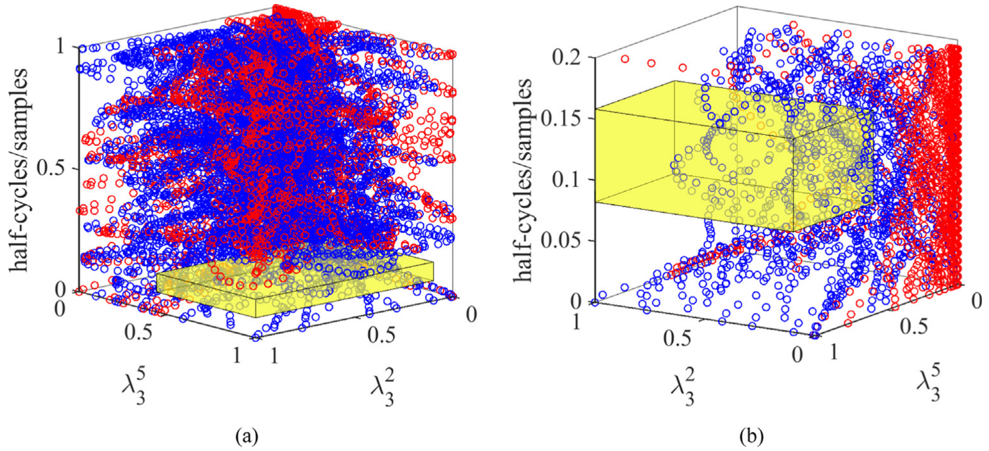


Fig. 12. Visualization of FTL formula $\varphi = \mathbf{F}_{[0.082,0.11]}((\mathbf{F}_{[0,0.076]}(\lambda_3^2 > 0.11) \wedge (\mathbf{F}_{[0,0.45]}(\lambda_3^5 > 0.44))$. a) the blue circles are positive points, and the red circles are negative points; b) a partial enlargement of (a).

Table 8
Scheme for the presentation of the results in Table 7.

	Positive Examples (H_0)	Negative Example (H_1)
Fail to reject H_0	Correct decision	Missing fault error
Reject H_0	False alarm error	Correct decision

need further analysis to reveal this property. Here we use statistical hypothesis testing to investigate the properties of the found formula.

Let us consider a t-test for the formulas obtained in above noisy conditions. We assume: (a) the robustness for a baseline fault signal is a random variable having a normal distribution with unknown mean $\bar{\rho}$ and unknown standard deviation $\bar{\rho}_\sigma$; and (b) the robustness for a signal that must be diagnosed is also normally distributed with unknown mean $\hat{\rho}$ and unknown standard deviation $\hat{\rho}_\sigma$. The problem that we will consider is to determine whether these means are equal. The t-test leads immediately to a test of hypotheses

$$H_0 : \bar{\rho} - \hat{\rho} = 0 \text{ versus } H_1 : \bar{\rho} - \hat{\rho} \neq 0, \tag{17}$$

where the null hypothesis is " the signal to be diagnosed is distributed as the baseline signal " and the alternative hypothesis is " the signal to be diagnosed is not distributed as the baseline signal". In other words, if the result of the test is that the null hypothesis is rejected, the current signal cannot be categorized as faults in the baseline samples. The scheme for the results are shown in Table 8, which leads to two kinds of error, namely false alarm error and missing fault errors.

During the t-test, the significance level is set to 0.05. To obtain the estimation of $\hat{\rho}$ and $\hat{\rho}_\sigma$, we took a 3000 width window of the signals that to be diagnosed from the rolling element bearings and calculated the robustness value. Next, we moved this window forward 100 sampling points along the time axis and repeated the calculations. The resulting trace of robustness values was used to determine $\hat{\rho}$ and $\hat{\rho}_\sigma$. The mean $\bar{\rho}$ and standard deviation $\bar{\rho}_\sigma$ are calculated based on the baseline trajectories. To test the formula found in the noise test experiments, the positive examples in the training set are used as baseline samples and the signals in the testing set are tested. The results in Table 7 indicate that fault diagnosis with statistical hypothesis testing will lead to a lower error rate than diagnosing with only FTL formula, e.g., when the SNR is 0.105, statistical hypothesis testing obtains a missing fault rate of 0.035, while FTL formula alone obtains an error rate of 0.053. This result demonstrates that the proposed algorithm can find an FTL formula that is robustness to noise.

CRediT author statement

Gang Chen: Conceptualization, Data curation, formal analysis, methodology, software, and writing. **Mei Liu:** Conceptualization, methodology, supervision, review & editing. **Jin Chen:** Review & editing.

7. Conclusions

This paper introduces a novel method for bearing fault diagnosis, which maps the vibration signals to some logic formulas. The logic formulas can be seen as interpretable classifiers, which can be used for fault diagnosis. Moreover, as the logic formulas are written in a formal language, it can be understood by a human and are easy to be used for online condition monitoring. To infer the structure of the logic description, BNN is combined with Bayesian optimization method, and the transfer learning technique is used to speed up the training procedure. The performance of the proposed method is demonstrated with two experiments.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Li, Y. Lei, J. Lin, S.X. Ding, An improved exponential model for predicting remaining useful life of rolling element bearings, IEEE Trans. Industr. Electron. 62 (12) (2015) 7762–7773.

[2] N.-K. Wesley, S. Bhandari, A. Subramaniam, M. Bagheri, and S.K. Panda, Evaluation of statistical interpretation methods for frequency response analysis based winding fault detection of transformers, in: 2016 IEEE International Conference on Sustainable Energy Technologies (ICSET), IEEE, 2016, pp. 36–41..

[3] J. Gonzales, E. Mombello, Fault interpretation algorithm using frequency-response analysis of power transformers, IEEE Trans. Power Deliv. 31 (3) (2015) 1034–1042.

[4] H.D.M. de Azevedo, A.M. Araújo, N. Bouchonneau, A review of wind turbine bearing condition monitoring: state of the art and challenges, Renew. Sustain. Energy Rev. 56 (2016) 368–379.

[5] J. Wu, C. Wu, S. Cao, S.W. Or, C. Deng, X. Shao, Degradation data-driven time-to-failure prognostics approach for rolling element bearings in electrical machines, IEEE Trans. Ind. Electron. 66 (1) (2019) 529–539.

[6] I. El-Thalji, E. Jantunen, A summary of fault modelling and predictive health monitoring of rolling element bearings, Mech. Syst. Signal Process. 60 (2015) 252–272.

[7] Y. Li, X. Liang, M.J. Zuo, Diagonal slice spectrum assisted optimal scale morphological filter for rolling element bearing fault diagnosis, Mech. Syst. Signal Process. 85 (2017) 146–161.

[8] W. Ahmad, S.A. Khan, J.-M. Kim, A hybrid prognostics technique for rolling element bearings using adaptive predictive models, IEEE Trans. Ind. Electron. 65 (2) (2018) 1577–1584.

[9] Y. Miao, M. Zhao, J. Lin, Y. Lei, Application of an improved maximum correlated kurtosis deconvolution method for fault diagnosis of rolling element bearings, Mech. Syst. Signal Process. 92 (2017) 173–195.

[10] V.C. Leite, J.G.B. da Silva, G.F.C. Veloso, L.E.B. da Silva, G. Lambert-Torres, E.L. Bonaldi, L.E. d. L. de Oliveira, Detection of localized bearing faults in induction machines by spectral kurtosis and envelope analysis of stator current, IEEE Trans. Ind. Electron. 62 (3) (2015) 1855–1865.

[11] D. Abboud, J. Antoni, M. Eltabach, S. Sieg-Zieba, Angle/ time cyclostationarity for the analysis of rolling element bearing vibrations, Measurement 75 (2015) 29–39.

[12] J. Wang, L. Qiao, Y. Ye, Y. Chen, Fractional envelope analysis for rolling element bearing weak fault feature extraction, IEEE/CAA J. Automatica Sin. 4 (2) (2017) 353–360.

[13] L. Lu, J. Yan, C.W. de Silva, Dominant feature selection for the fault diagnosis of rotary machines using modified genetic algorithm and empirical mode decomposition, J. Sound Vib. 344 (2015) 464–483.

[14] H. Jiang, J. Chen, G. Dong, T. Liu, G. Chen, Study on hankel matrix-based SVD and its application in rolling element bearing fault diagnosis, Mech. Syst. Signal Process. 52 (2015) 338–359.

[15] M. Yuwono, Y. Qin, J. Zhou, Y. Guo, B.G. Celler, S.W. Su, Automatic bearing fault diagnosis using particle swarm clustering and hidden markov model, Eng. Appl. Artif. Intell. 47 (2016) 88–100.

[16] G. Xin, N. Hamzaoui, J. Antoni, Semi-automated diagnosis of bearing faults based on a hidden markov model of the vibration signals, Measurement 127 (2018) 141–166.

[17] H.O. Omoregbee, P.S. Heyns, Fault detection in roller bearing operating at low speed and varying loads using bayesian robust new hidden markov model, J. Mech. Sci. Technol. 32 (9) (2018) 4025–4036.

[18] P. Baraldi, F. Cannarile, F. Di Maio, E. Zio, Hierarchical k-nearest neighbours classification and binary differential evolution for fault diagnostics of automotive bearings operating under variable conditions, Eng. Appl. Artif. Intell. 56 (2016) 1–13.

[19] G. Chen, J. Chen, A novel wrapper method for feature selection and its applications, Neurocomputing 159 (2015) 219–226.

[20] Z. Jian, X.-B. Li, X.-z. Shi, W. Wei, B.-b. Wu, Predicting pillar stability for underground mine using fisher discriminant analysis and SVM methods, Transactions of Nonferrous Metals Society of China, vol. 21, no. 12, 2011, pp. 2734–2743..

[21] P. Boškoski, M. Gašperin, D. Petelin, *D.* Juričić, Bearing fault prognostics using rényi entropy based features and gaussian process models, Mech. Syst. Signal Process. 52 (2015) 327–337..

[22] X. Li, W. Zhang, Q. Ding, A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning, Neurocomputing 310 (2018) 77–95.

[23] S. Wang, J. Xiang, Y. Zhong, Y. Zhou, Convolutional neural network-based hidden markov models for rolling element bearing fault identification, Knowl.-Based Syst. 144 (2018) 65–76.

[24] M. He, D. He, Deep learning based approach for bearing fault diagnosis, IEEE Trans. Ind. Appl. 53 (3) (2017) 3057–3065.

[25] H. Shao, H. Jiang, H. Zhang, T. Liang, Electric locomotive bearing fault diagnosis using a novel convolutional deep belief network, IEEE Trans. Ind. Electron. 65 (3) (2018) 2727–2736.

[26] O. Maler, D. Nickovic, Monitoring temporal properties of continuous signals, in: Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems, Springer, 2004, pp. 152–166..

[27] G. Chen, Z. Sabato, Z. Kong, Active learning based requirement mining for cyber-physical systems, in: 2016 IEEE 55th Conference on Decision and Control, IEEE, 2016, pp. 4586–4593..

[28] A. Donzé, O. Maler, Robust satisfaction of temporal logic over real-valued signals, in: International Conference on Formal Modeling and Analysis of Timed Systems, Springer, 2010, pp. 92–106..

[29] E. Bartocci, L. Bortolussi, G. Sanguinetti, Data-driven statistical learning of temporal logic properties, in: International conference on Formal Modeling and Analysis of Timed Systems, Springer, 2014, pp. 23–37..

[30] L. Nenzi, S. Silvetti, E. Bartocci, L. Bortolussi, A robust genetic algorithm for learning temporal specifications from data, in: International Conference on Quantitative Evaluation of Systems, Springer, 2018, pp. 323–338..

[31] D. Neider, I. Gavran, Learning linear temporal properties, in: Formal Methods in Computer Aided Design , IEEE, 2018, pp. 1–10.

[32] G. Bombara, C.-I. Vasile, F. Penedo, H. Yasuoka, C. Belta, A decision tree approach to data classification using signal temporal logic, in: Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control , ACM, 2016, pp. 1–10.

[33] Z. Kong, A. Jones, A. Medina Ayala, E. Aydin Gol, C. Belta, Temporal logic inference for classification and prediction from data, in: Proceedings of the 17th international conference on Hybrid Systems: Computation and Control , ACM, 2014, pp. 273–282.

[34] R. Lee, M.J. Kochenderfer, O.J. Mengshoel, J. Silbermann, "Interpretable categorization of heterogeneous time series data," in, in: Proceedings of the 2018 SIAM International Conference on Data Mining , SIAM, 2018, pp. 216–224.

[35] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, R. Adams, Scalable bayesian optimization using deep neural networks, in: International Conference on Machine Learning, 2015, pp. 2171–2180.

[36] M. Fortunato, C. Blundell, O. Vinyals, Bayesian recurrent neural networks, arXiv preprint arXiv:1704.02798, 2017..

[37] I. Loshchilov, F. Hutter, Cma-es for hyperparameter optimization of deep neural networks, arXiv preprint arXiv:1604.07269, 2016..

[38] Y. Gal, Z. Ghahramani, Bayesian convolutional neural networks with bernoulli approximate variational inference, arXiv preprint arXiv:1506.02158, 2015..

[39] M. Gokhale, D.K. Khanduja, Time domain signal analysis using wavelet packet decomposition approach, Int. J. Commun. Netw. Syst. Sci. 3 (03) (2010) 321.

[40] C. Mishra, A. Samantaray, G. Chakraborty, Rolling element bearing defect diagnosis under variable speed operation through angle synchronous averaging of wavelet de-noised estimate, Mech. Syst. Signal Process. 72 (2016) 206–222.

[41] P.J. Loughlin, What are the time-frequency moments of a signal?, in: Advanced Signal Processing Algorithms, Architectures, and Implementations XI, vol. 4474, International Society for Optics and Photonics, 2001, pp. 35–45. .

[42] S. Park, S.-C. Zhu, Attributed grammars for joint estimation of human attributes, part and pose, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2372–2380.

[43] M. Welling, Y.W. Teh, Bayesian learning via stochastic gradient langevin dynamics, in, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 681–688.

[44] J.T. Springenberg, A. Klein, S. Falkner, F. Hutter, Bayesian optimization with robust bayesian neural networks, in: Advances in Neural Information Processing Systems, 2016, pp. 4134–4142. .

[45] T. Desautels, A. Krause, J.W. Burdick, Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization, J. Mach. Learn. Res. 15 (1) (2014) 3873–3923.

[46] N. Srinivas, A. Krause, S.M. Kakade, M. Seeger, Gaussian process optimization in the bandit setting: no regret and experimental design, arXiv preprint arXiv:0912.3995, 2009..

[47] P. Borghesani, P. Pennacchi, R. Randall, N. Sawalhi, R. Ricci, Application of cepstrum pre-whitening for the diagnosis of bearing faults under variable speed conditions, Mech. Syst. Signal Process. 36 (2) (2013) 370–384.

[48] D.G. Childers, D.P. Skinner, R.C. Kemerait, The cepstrum: a guide to processing, Proc. IEEE 65 (10) (1977) 1428–1443.

[49] M. Zhao, J. Lin, X. Xu, Y. Lei, Tacholess envelope order analysis and its application to fault detection of rolling element bearings with varying speeds, Sensors 13 (8) (2013) 10 856–10 875.

[50] Z. Feng, X. Chen, T. Wang, Time-varying demodulation analysis for rolling bearing fault diagnosis under variable speed conditions, J. Sound Vib. 400 (2017) 71–85.

[51] Z. Kong, A. Jones, C. Belta, Temporal logics for learning and detection of anomalous behavior, IEEE Trans. Autom. Control 62 (3) (2016) 1210–1222.

[52] E. Alba, J. Garcia-Nieto, L. Jourdan, E.-G. Talbi, Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms, in: 2007 IEEE Congress on Evolutionary Computation, IEEE, 2007, pp. 284–290. .

[53] H. Huang, N. Baddour, Bearing vibration data collected under time-varying rotational speed conditions, Data Brief 21 (2018) 1745–1749.