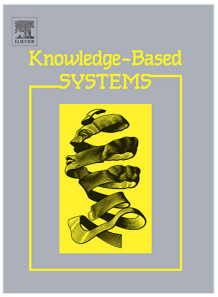
Journal Pre-proof



Machine learning-based wear fault diagnosis for marine diesel engine by fusing multiple data-driven models

Xiaojian Xu, Zhuangzhuang Zhao, Xiaobin Xu, Jianbo Yang,
Leilei Chang, Xinping Yan, Guodong Wang

PII: S0950-7051(19)30599-4
DOI: https://doi.org/10.1016/j.knosys.2019.105324
Reference: KNOSYS 105324

To appear in: *Knowledge-Based Systems*

Received date : 28 February 2019
Revised date : 30 November 2019
Accepted date : 30 November 2019

Please cite this article as: X. Xu, Z. Zhao, X. Xu et al., Machine learning-based wear fault diagnosis for marine diesel engine by fusing multiple data-driven models, *Knowledge-Based Systems* (2019), doi: https://doi.org/10.1016/j.knosys.2019.105324.

Machine learning-based wear fault diagnosis for marine diesel engine by fusing multiple data-driven models

Xiaojian Xu[a], Zhuangzhuang Zhao[a], Xiaobin Xu[a*], Jianbo Yang[a*], Leilei Chang[a], Xinping Yan[b], Guodong Wang[c]

[a] *School of Automation, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China*

[b] *National Engineering Research Center for Water Transport Safety, Wuhan University of Technology, Wuhan 430063, Hubei, China*

[c] *Institute of Computer Engineering, Vienna University of Technology, Vienna, DE0364, Austria*

Abstract: Wear fault is one of the dominant causes for marine diesel engine damage which significantly influences ship safety. By taking full advantage of the data generated in engine operation, machine learning-based wear fault diagnostic model can help engineers to determine fault modes correctly and take quick action to avoid severe accidents. To identify wear faults more accurately, a multi-model fusion system based on evidential reasoning (ER) rule is proposed in this paper. The outputs of three data-driven models including an artificial neural network (ANN) model, a belief rule-based inference (BRB) model, and an ER rule model are used as pieces of evidence to be fused in decision level. In this paper, the fusion system defines reliability and importance weight of every single model respectively. A novel method is presented to determine the reliability of evidence by considering the accuracy and stability of every single model. The importance weight is optimized by genetic algorithm to improve the performance of the fusion system. The proposed machine learning-based diagnostic system is validated by a set of real samples acquired from marine diesel engines in operation. The test results show that the system is more accurate and robust, and the fault tolerant ability is improved remarkably compared with every single data-driven diagnostic model.

Keywords: wear fault diagnosis, marine diesel engine, machine learning-based diagnostic model, fusion system, ER rule

1 Introduction

Currently most ships over the world are propelled by marine diesel engines, and the reliability of marine diesel engines has a significant influence on the safe and economical operation of ships. As indicated in the report published by The Swedish Club, claims caused by main engine damage account for 34.4% of total marine machinery claims in 2012-2014, causing over 21million USD in total [1]. Furthermore, marine diesel engines consist of many tribological systems, such as cylinder liner-piston ring system, main bearing system, and almost 50% of engine faults are caused by abnormal wear of friction pairs [2]. Therefore, it is necessary to study the wear fault diagnosis of marine diesel engines to improve their reliability.

The health condition of marine diesel engines is generally monitored by condition monitoring methods, including performance parameters monitoring, vibration monitoring, and oil monitoring. Compared with other condition monitoring methods, oil monitoring especially

performs better in wear fault diagnosis, because lots of tribological information can be extracted from wear particles in lubricating oil. Moreover, the wear condition of marine diesel engine can be monitored without disturbing the normal operation of engines and changing the engine structure. Generally, oil monitoring includes wear particle quantity analysis, morphology analysis, and physicochemical property analysis of lubricating oil, offering information to be used in wear fault mode and mechanism identification [3].

Considering wear fault mechanism and acquisition methods of tribological information, it can be found that several problems exist in wear fault diagnosis for marine diesel engine. Specifically, the complicated nonlinear relationship between wear fault characteristics and fault modes is difficult to be expressed by accurate physical models. Additionally, most data which can reflect the wear states is not well used, and the data is generally uncertain and incomplete. Besides the above two problems, engine operation requires high reliability so that engineers should handle the wear faults as soon as possible. Fortunately, machine learning-based methods offer a feasible way to solve the above problems.

Machine-learning based fault diagnosis technology is a sensitive, potentially cheaper, and high-efficient alternative for wear fault identification in comparison with regular manual maintenance such as corrective maintenance and preventive maintenance. Wear fault diagnosis is actually a decision-making issue. By using artificial intelligent algorithms, machine learning-based models analyze a large amount of data generated in engine operation, observations, and domain knowledge, and develop reusable decision-making models to represent the nonlinear relationship between fault features and fault modes. These decision-making models can deduce the data and knowledge of the specific application objects and help engineers to find out fault causes, identify fault modes, evaluate fault severity, and so on. The machine learning-based diagnostic models can not only solve problems in wear fault diagnosis, they also have several superiorities in decision making. Firstly, machine learning-based diagnosis realize the man-machine collaborative decision making instead of human oriented decision making, and therefore machine learning-based models have better capabilities of using multi-source information including qualitative and quantitative information. Secondly, the utilization of data is increased by the machine learning-based methods. More hidden information which is helpful for decision makers is mined, and decision makers can enrich their domain knowledge with the diagnostic model in return. Additionally, many decision-making problems become automated and intelligent so that the efficiency and accuracy of decision making are improved obviously. As a result, many accidents caused by wrong decisions or delayed decisions can be avoided. Finally, the machine learning-based diagnostic model can be fine-tuned by optimization algorithms and

historical data to give the optimal decision result.

In the machine learning-based wear fault diagnostic system, tribological information acquired by lubricating oil monitoring is used as the input while the wear fault mode or mechanism is the output. Data-driven algorithms are used to simulate the nonlinear relationship between inputs and outputs. An important issue in designing a machine learning-based wear fault diagnostic system is how to select an appropriate algorithm which can deal with the tribological information with uncertainty or imperfection, and build a many to many mapping relationship between tribological information and wear modes. The majority of the current wear fault diagnostic systems are mainly developed by using one single intelligent algorithm. However, different algorithms have their own merits and demerits. The robustness and fault tolerant capability of the fault diagnostic model are quite limited by only using one single algorithm. Thus, how to take advantages of different diagnostic systems with a fused method is deserved to be studied.

To improve the performance of wear fault diagnostic model, this paper proposes a new approach to fuse different fault diagnostic models in decision level by using ER rule algorithm. Specifically, three wear fault diagnostic models for marine diesel engines including bi-level BRB (BBRB) model, bi-level ANN (BANN) model, and ER rule model are developed respectively based on our previous researches [4,5]. Notably, reliability and importance of every single model are considered separately, in the process of fusing the three diagnostic models by using ER rules. Furthermore, a new method to determine the reliability factor of every model is proposed, in which model accuracy and model stability are taking into account simultaneously. To increase the fault diagnostic accuracy after decision level fusion, genetic algorithm is applied to modify the importance weight of every individual model. The newly developed model in this paper is applied to wear fault mode identification of marine diesel engines, and verified by the real data samples.

The main contributions of this paper are as follows. Firstly, the paper provides a decision-level fused model to overcome the demerits of the three data-driven diagnostic models, which helps us to further improve the performance of wear fault diagnostic models, especially model robustness and fault-tolerant capability. Secondly, the inherent property of every single model (i.e. reliability) and the subjective property (i.e. importance weight) are distinguished in the multi-model fusing process. Lastly, model accuracy and stability are simultaneously used to generate the reliability of every single model, which makes the model can be evaluated comprehensively and reasonably.

The remainder of the paper is organized as follows. In section 2, we discussed the related work in intelligent wear fault diagnosis, multi-model fusion, and ER rules. Section 3 analyzes

wear faults of marine diesel engines and formulates the wear fault diagnostic problem mathematically. The fused fault diagnostic model in decision level based on ER rule is developed in section 4. In section 5, a computational study is conducted to describe how to apply the fused model to wear fault diagnosis. The results of a series of experiments are discussed in section 6. Lastly, conclusions are given in section 7.

2 Literature Review

2.1 Intelligent wear fault diagnostic models

Wear fault diagnostic models are mainly based on statistical analysis, mathematical-physical models, and intelligent algorithms. Among these wear fault diagnostic models, intelligent wear fault diagnostic models are mostly studied. By using artificial intelligent algorithms, intelligent fault diagnosis can be conducted with the information acquired by condition monitoring systems.

Among a variety of intelligent methods applied in wear fault diagnosis, traditional rule based expert system is one of the firstly developed intelligent models. In the early stage, researches were mainly on the development of wear fault diagnostic expert systems, Katsoulakos and Autar have constructed expert systems by extracting IF-THEN rules from expert domain knowledge to identify wear faults of engines long before 1990s [6,7]. After that, more and more researches focus on expert systems optimization, and integration with other algorithms, for example, the combination of neural network and expert systems, and the combination of fuzzy inference and expert systems [8,9]. Although the process of wear fault diagnosis based on expert system is quick and easy, the performance of diagnostic models is highly limited by expert domain knowledge. Nowadays, most wear fault diagnostic models are developed based on data-driven algorithms, because they have strong computing, nonlinear mapping and self-learning capabilities which can solve the problems in expert systems. ANN is the most representative methods among various intelligent algorithms. Till now, BP neural network are still the most widely used neural network in wear fault diagnosis. Basurko developed a BP neural network-based maintenance system to monitor the health condition of a medium-speed diesel engine, while Guo and Yuan developed a BP neural network to detect the abnormal wear of cylinder-liner piston-ring [10,11]. However, it is generally difficult to determine the structure of BP neural network, and the algorithm convergence is slow. To solve these problems, other kinds of ANN models were developed in engine fault diagnosis, such as RBF neural network, probability neural network, and fuzzy neural network [12-14]. Additionally, other data-driven methods were also used in fault diagnosis, such as support vector machine (SVM), fuzzy c-means clustering, and Bayesian network [15-17]. It should be noted that data-driven methods generally have high requirement

on dataset size and quality, and the variation of data samples could influence the performance of diagnostic model obviously. BRB inference methodology is a new method gradually applied in mechanical fault diagnosis. Compared with other methods, it can use quantitative and qualitative information simultaneously. Moreover, the inference process is transparent and interpretable. Currently, Xu has applied BRB to identify different wear modes of marine diesel engines, and find out the parts in abnormal wear condition [4,18]. Diagnostic results show that the performance of BRB fault diagnostic model for marine diesel engines is promising.

In summary, there are a variety of intelligent methods applied in wear fault diagnosis for diesel engines. However, every intelligent algorithm has its application area. Specifically, limited domain knowledge is the bottleneck of traditional rule based expert system, and it lacks flexibility, since the expert system cannot learn from real operating data to modify the diagnostic system. Data-driven models are developed on the basis of a large amount of data, and therefore, data quality significantly affects the performance of these models. In addition, most data-driven models are black boxes, such as ANN models. It is difficult for users to understand how the final diagnostic results are generated and to find potential mistakes in models. BRB model can solve the above problems well, but the complexity of belief rule base will increase obviously in large scale problems, and it will become invalid, if the input is incomplete compared with the antecedent attributes set of a BRB model. As a result, this paper proposes a machine learning-based diagnostic model with the integration of several different data-driven diagnostic models to compensate the drawback of every model.

2.2 Multi-model fusion

The ultimate goal in fault diagnosis for marine diesel engine is to achieve a diagnostic model with the best possible performance. From the above literature review, it can be known that different single models have their own diagnostic capability, and sometimes they can compensate with each other to generate a more accurate diagnostic result. Therefore, methods on multi-model fusion (i.e. multiple classifiers fusion) attract more attention.

The concept of multiple classifier fusion was first proposed in 1992 by Xu, Krzyzak and Suen which is applied in handwriting recognition [19]. Since then, a variety of fusion methods have been put forward. From the point of model structure, the fusion models can be divided into cascade structure and parallel structure. The results generated by the previous layer is used as the input of the next layer in the cascade structure while every single model works independently and then is fused with other models in parallel structure. Obviously, parallel structure is more suitable to fully use every single model's merits. In multi-model fusion, the key issue is how to integrate these single models in an effective way. Voting

method is the simplest and most widely used method which treats the decision of every classifier as a vote, and the class with most votes is determined as the final result [20]. However, voting method only considers the output label and ignores the model accuracy, therefore, it cannot make a full use of the information contained in training samples. A class-aligned method is also commonly used which is to calculate the support for one class using linear sum, product, or order statistics including minimum, maximum and median [21]. Kittler found that sum has the best identifying performance [22], but the performance of this method is easily affected by the classifier with lower accuracy. Another fusing strategy is performed on an entire decision profile to aggregate multiple models to acquire a confidence value for a decision [23,24]. D-S theory is generally used in this fusing strategy, because it provides a very efficient theoretical framework for representing and combining uncertain information from distinct sources [25]. Many other methods have been applied in this area, such as Bayesian fusion method, behavior knowledge space fusion method, and logistic regression method [26-28]. It should be noticed that every single model's contribution to the fused model is various due to their own performances which should be taken into account in the fusion process. Instead of using the output of every single model as a piece of evidence directly, many researches take the importance of every model into consideration. Xu used recognition, substitution, and rejection rate to define the sources of evidence for the proposition of interest, and therefore the overall performance of every single model was considered in the fusing process [29]. Rogova defined reference vectors for all classes, and measured the distance between the outputs of every model and the reference vectors which was used as a piece of evidence to be fused by D-S theory [30]. In this way, the single model of which the output was more similar with the reference vector will play a more important role in the fusing process.

In most current researches, the accuracy of a single model is considered to determine its importance weight, but every single model to be fused is considered to be fully reliable, and therefore, the effect of the reliability of the model on the final output is ignored. However, reliability and importance weight are two properties of evidence having different meanings. Specificity, reliability is the inherent property of models (i.e. information source), which is independent of who may use the evidence and the performance of other classifiers. On the other hand, importance weight is subjective, depending on who makes the decision and the other models' performance [31]. Compared with D-S theory, ER rule extends the weighted evidence to evidence with weight and reliability, and combines multiple single models overall considering the evidence reliability and importance weight simultaneously.

2.3 ER rule

ER rule is a probabilistic reasoning process to combine multiple independent pieces of evidence considering the reliability and importance weight of evidence. In reference [31], Yang and Xu proved that ER algorithm and Dempster rule are two particular cases of ER rule algorithm. Nowadays, ER rule has been applied in machinery fault diagnosis, disaster prevention, medical diagnosis, risk analysis, and etc [32-35]. Most researches on ER rule mainly focus on evidence acquisition, reliability and importance determination. Xu proposed a new way to determine the evidence matrix by using referential points instead of evidence intervals, which improves the application range of ER rule models [36]. Xu developed a dual objective optimization model to train the importance weights of evidence. In this model, diagnostic accuracy and the distance between the importance weight of evidence and its referential point are the two objective functions. Till now, ER rule has not been applied in multi-model fusion which is worthy of being studied.

3 Problem Statement and Formulation

3.1 Problem statement

According to wear mechanism in machinery, wear faults can generally be classified into abrasive wear, adhesive wear, fatigue wear, and corrosive wear, producing different wear particles. The typical wear particles include normal wear particles, sever sliding wear particles, cutting wear particles, spherical wear particles, fatigue spall particles, laminar particles, red oxides, and black oxides [2]. Figure 1 describes the corresponding relations among wear fault modes, wear particles and wear mechanisms. Since wear particles contain abundant tribology information, and different wear fault modes produce distinctive wear particles, the categories of wear particles are used as outputs of wear fault diagnostic models for marine diesel engines.

Wear particles can be separated from lubricating oil circulating in marine diesel engine, and then be made into ferrography or filtergram. The pictures of wear particles can be taken by ferroscope and laser scanning confocal microscope. SPIP is an image processing software to extract the wear particle characteristics from particle pictures. The characteristics of wear particles include two-dimensional (2-D) characteristics such as aspect ratio (AR), equivalent diameter (D_e), roundness (R), and three-dimensional (3-D) characteristics such as roughness average (S_a) and texture direction index (S_{tdi}). These wear particle characteristics are used as the input of the wear fault diagnostic model. The detailed descriptions of the wear particle features can be referred in reference [14].

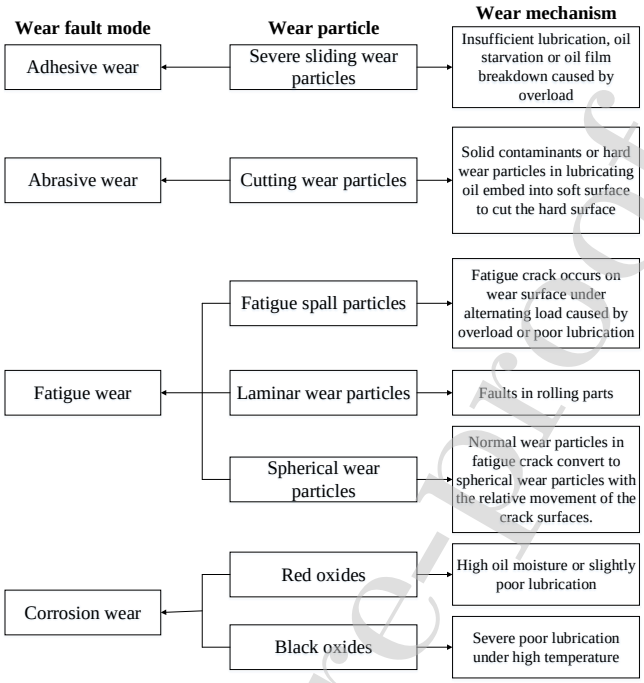


Figure 1. Corresponding relations among wear fault modes, wear particles and wear mechanisms

Machine learning-based algorithms can be used to build the relationship between wear particle features and wear fault modes. In this paper, three different fault diagnostic models based on BRB, ANN, and ER rule are fused to generate the final diagnostic result. The advantages and disadvantages of the three algorithms are compared as listed in Table 1 [4,5,18].

Table 1 Comparisons among BRB, ER rules and ANN algorithms

Classification Method	Advantages	Disadvantages
BRB	1. Inference process is transparent, and classification model has better interpretability. 2. Quantitative and qualitative information can be synchronously utilized. 3. BRB can deal with the uncertainty in information. 4. Models can be developed and optimized with less data samples. 5. BRB model is more stable, and model structure and performance are	1. BRB is not good at dealing with incomplete information. 2. BRB is not suitable for large scale problem, especially when too many input features are involved, there is high risk of combination explosion in belief rule base.

	less affected by data variation. 6. The outputs are expressed in belief distribution.	
ER rule	1. Inference process is transparent, and classification model has better interpretability. 2. ER rule can deal with the uncertainty in information. 3. ER rule has good capability of dealing with incomplete information. 4. Large scale problem can be well solved. 5. Reliability and importance weight of evidence are distinguished clearly. 6. The outputs are expressed in belief distribution.	1. ER rule is a data-driven method, requiring a large amount of data samples, and the more sufficient data samples, the more accurate the evidence matrix will be. 2. Unified methods to determine the belief distribution and the reliability of evidence are still lacking.
ANN	1. ANN has a outstanding capability of non-linear fitting. 2. Large scale problem can be well solved. 3. ANN outputs certain results.	1. ANN is a black box simulator, and the inference process is difficult to be explained. 2. ANN is a data-driven method, requiring a large amount of data samples, and the more sufficient data samples, the more accurate the evidence matrix will be. 3. ANN is not good at dealing with incomplete information. 4. The variation of data samples has a significant influence on the model structure and performance, reducing the model stability.

3.2 Problem formulation

For every single diagnostic model, suppose $\boldsymbol{x} = [x_1, x_2, ..., x_M]$ is the wear fault diagnostic features which are extracted from wear particle images, where $x_i (i = 1, 2, ..., M)$ denotes the ith input attribute and M is the number of attributes. E is assumed as the output of a single diagnostic model and P is the corresponding parameters vector. Every single diagnostic model based on BRB, ANN, and ER rule is aimed to establish causal relationship between X and E, which is generally represented by $E = f(x, P)$, where f is a function of E. The output can be represented as: $E = \{(E_1, \beta_1), (E_2, \beta_2), ..., (E_N, \beta_N)\}$, where $E_i (i = 1, 2, ..., N)$ is the category of

wear particles, $\beta_i (i=1,2,...,N)$ is the corresponding belief degree and the following constrains are satisfied: $\sum_{i=1}^{N} \beta_i = 1$ and $0 \le \beta_i \le 1$.

In the fused model, $E_i = \{(E_{i1}, \beta_{i1}), (E_{i2}, \beta_{i2}), ..., (E_{iN}, \beta_{iN})\} (i=1,...,T)$ is the evidence to be fused in decision level which is generated by every single model. It is assumed that *D* is the output of the fused diagnostic model based on ER rule, *r* is the reliability vector of evidence, and *w* is the importance weight vector of evidence. The fused diagnostic model can be described as $D = g(E, r, w)$. *g* is the function of *D*, and $D = \{(D_1, \beta_1^f), (D_2, \beta_2^f), ..., (D_N, \beta_N^f)\}$ satisfies the constrains $\sum_{i=1}^{N} \beta_i^f = 1$ and $0 \le \beta_i^f \le 1$. $r = [r_1, r_2, ..., r_T]$ is determined by the inherent property of evidence where $0 \le r_i \le 1 (i=1,...,T)$, and $w = [w_1, w_2, ..., w_T]$ is determined by decision maker's knowledge and preference where $0 \le w_i \le 1 (i=1,...,T)$. *w* maybe inaccurate and it can be optimized by genetic algorithm which will be described in section 4.

4 The Fused Fault Diagnostic Model

4.1 Evidential reasoning (ER) rules

When using ER rule to combine multiple pieces of evidence, three elements are essential which are belief distribution of evidence, evidence reliability and evidence importance weight. Suppose $\Theta = \{h_1, h_2, ..., h_N\}$ is the frame of discernment, consisting *N* mutually exclusive and collectively exclusive hypotheses. All subsets of Θ constitute its power set, represented by $P(\Theta)$ or 2^{Θ}. In the frame of discernment Θ , the belief distribution of a piece of evidence is:

$$e_j = \{(\theta, p_{\theta,j}) \mid \forall \theta \subseteq \Theta, \sum_{\theta \subseteq \Theta} p_{\theta,j} = 1\} \tag{1}$$

where $(\theta, p_{\theta,j})$ is the element of e_j , $p_{\theta,j}$ represents the belief degree of evidence e_j supporting to the proposition θ . θ can be any element of $P(\Theta)$ except the empty set.

In ER rules, reliability factor $r_j (0 \leqslant r_j \leqslant 1)$, represents how the evidence e_i provides correct assessment or solution for a given problem. Importance weight $w_j (0 \leqslant w_j \leqslant 1)$ reflects the relative importance compared with other evidence to be combined. Taking reliability factor r_j

and importance weight w_j into consideration, the belief distribution of evidence e_j can be modified to be:

$$m_j = \{(\theta, \tilde{m}_{\theta,j}), \forall \theta \subseteq \Theta; (P(\Theta), \tilde{m}_{P(\Theta),j})\} \tag{2}$$

Where $\tilde{m}_{\theta,j}$ is the belief degree of evidence e_j supporting to the proposition θ when considering the evidence reliability and importance weight. $\tilde{m}_{\theta,j}$ is defined to be:

$$\tilde{m}_{\theta,j} = \begin{cases} 0 & \theta = \varnothing \\ c_{rw,j} m_{\theta,j} & \theta \subseteq \Theta, \theta \neq \varnothing \\ c_{rw,j}(1 - r_j) & \theta = P(\Theta) \end{cases} \tag{3}$$

In (3), $m_{\theta,j}$ is the basic probability mass, and $m_{\theta,j} = w_j p_{\theta,j}$; $c_{rw,j}$ is the normalization factor, and $c_{rw,j} = 1/(1 + w_j - r_j)$, ensuring $\sum_{\theta \subseteq \Theta} \tilde{m}_{\theta,j} + \tilde{m}_{P(\Theta),j} = 1$ when $\sum_{\theta \subseteq \Theta} p_{\theta,j} = 1$. $(1 - r_j)$ describes the unreliability of evidence e_j, restricting the effect of other evidence on the final result when they are fused with evidence e_j.

The combined belief degree of two independent pieces of evidence e_1 and e_2 is $p_{\theta,e(2)}$, representing the joint support of e_1 and e_2 to proposition θ, which can be acquired by (4) - (6).

$$p_{\theta,e(2)} = \begin{cases} 0 & \theta = \varnothing \\ \dfrac{\hat{m}_{\theta,e(2)}}{\sum_{D \subseteq \Theta} \hat{m}_{D,e(2)}} & \theta \subseteq \Theta, \theta \neq \varnothing \end{cases} \tag{4}$$

$$\hat{m}_{\theta,e(2)} = [(1 - r_2) m_{\theta,2} + (1 - r_1) m_{\theta,1}] + \sum_{B \cap C = \theta} m_{B,1} m_{C,2} \quad \forall \theta \subseteq \Theta \tag{5}$$

$$\hat{m}_{P(\Theta),e(2)} = (1 - r_2)(1 - r_1) \tag{6}$$

In (5), $\hat{m}_{\theta,e(2)}$ is the orthogonal sum of the weighted belief distributions of two pieces of evidence with reliability. If evidence e_j has a higher reliability, it will reduce the effect of other evidence on the combined result more significantly. (6) describes the residual belief degree after the two pieces of evidence combine.

Similarly, multiple independent pieces of evidence $e_i (i = 1, 2, ..., L)$ can be combined

according to (7) - (9), using the recursive ER rule algorithm to determine the combined support of L pieces of evidence to proposition θ .

$$p_{\theta,e(L)}=\begin{cases}0 & \theta=\varnothing\\ \dfrac{\hat{m}_{\theta,e(L)}}{\sum_{D\subseteq\Theta}\hat{m}_{D,e(L)}} & \theta\subseteq\Theta,\theta\neq\varnothing\end{cases} \tag{7}$$

$$\hat{m}_{\theta,e(L)}=[(1-r_i)m_{\theta,(i-1)}+m_{P(\Theta),e(i-1)}m_{\theta,i}]+\sum_{B\cap C=\theta}m_{B,e(i-1)}m_{C,i}\quad\forall\theta\subseteq\Theta \tag{8}$$

$$\hat{m}_{P(\Theta),e(i)}=(1-r_i)m_{P(\Theta),e(i-1)} \tag{9}$$

In fault diagnosis, the output of every single diagnostic model is represented in belief distribution which is used as a piece of evidence to be fused by ER rule to generate the final diagnostic result.

4.2 Process of fault diagnosis based on the fused model

Figure 2 indicates the prototype of the fused diagnostic model, which includes four parts: developing single data-driven diagnostic models, determining the reliability of every single model, fusing all single models in decision level, and optimizing the fused model.

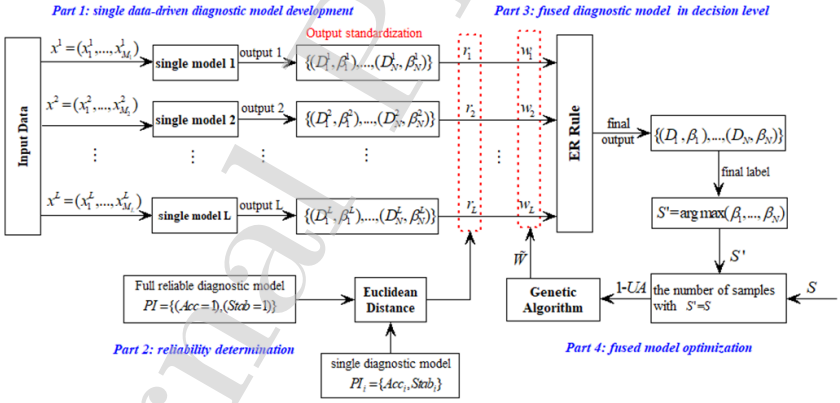


Figure 2. The prototype of fused diagnostic model

1) Developing single data-driven diagnostic models

Every single diagnostic model is developed in part 1, and these single models are built on the basis of different data-driven algorithms such as ANN, BRB, and ER rule method. Since the single diagnostic models are developed based on different classification algorithm, all single diagnostic models are independent of each other, having their own input vector

$\boldsymbol{x}^i = (x_1^i, x_2^i, ..., x_{M_i}^i)$ $(i = 1, 2, ..., L)$ and model structure according to the algorithm properties and characteristics of a diagnostic object, where L is the number of single models to be fused and M_i is the input attributes of the diagnostic model. The output generated by every single diagnostic model, which is represented in belief distribution, is used as a piece of evidence to be fused by ER rule algorithm in part 3. It is necessary to transform the output of every single diagnostic model into a uniform form before fused to ensure that all pieces of evidence consist of the same consequent attributes.

2) Determining the reliability factor

Reliability factor r_i reflects the ability of evidence to provide correct assessment for classification problem. It has a significant influence on the combined result, and the evidence with higher reliability will contributing more in the evidence fusing process. Therefore, how to determine the reliability of every single model is a key issue, and a new method is proposed in part 2 to solve the problem. In detail, there are four steps to calculate the reliability factor:

Step1: Select factors to evaluate single model's performance. Generally, accuracy is the principle indictor to evaluate a model's performance. However, stability is also an importance property of a model, because it manifests the model's robustness and sensitivity to data perturbation. Consequently, accuracy and stability are selected as the two factors to evaluate the reliability of every single model, constituting the performance index (*PI*) vector. Additionally, a fully reliable model is defined at first, of which the accuracy and stability are all set to be one, i.e. $PI = \{(Acc = 1), (Stab = 1)\}$, where Acc represents model accuracy and $Stab$ denotes model stability. The fully reliable model is the baseline to measure other model's reliability, and its reliability factor is 1.

Step 2: Calculate every single diagnostic model's accuracy Acc_i. For every single diagnostic model, *K*-cross validation experiment is conducted. In the *K*-cross validation experiment, the original dataset is randomly partitioned into *K* equal-sized sub-datasets. Of the *K* sub-datasets, a single sub-dataset is retained as testing data, and the remaining sub-datasets are used as training data. The cross-validation process is then conducted *K* times

and Acc_i is averaged over the K rounds as shown in (10), where UA_j is the accuracy of the single diagnostic model in the jth round.

$$Acc_i = \sum_{j=1}^{K} UA_j \Big/ K \tag{10}$$

Step 3: Calculate every single diagnostic model's stability $Stab_i$. Similarly, K-cross validation experiment is also carried out. The model stability is the capabilities of these K classifiers to generate the repeatable identification result for one sample [37]. The stability of the ith model is calculated according to (11) and (12).

$$Stab_i^k = \frac{1}{ne}\sum_{e=1}^{ne}\left[\frac{2}{nr(nr-1)}\sum_{p=1}^{nr}\sum_{q=1}^{p-1} agree_{e_{pq}}\right] \tag{11}$$

$$Stab_i = \sum_{k=1}^{K} Stab_i^k \Big/ K \tag{12}$$

In (11) and (12), $Stab_i^k$ is the stability of the ith classifier in the kth fold validation. ne is the number of testing samples and nr is the times of a sample to be diagnosed. When the pth and qth classifiers based on the same classification algorithm generate the same identification result for one sample, $agree_{e_{pq}} = 1$, otherwise $agree_{e_{pq}} = 0$. $Stab_i$ is the average stability of the ith diagnostic model in the K-cross validation.

Step 4: Determine the reliability of every single model. Euler distance is used to evaluate the single model's reliability compared with the fully reliable model as indicated in (13), and the smaller Euler distance is, the higher reliability of the single model will be.

$$r_i = 1 - \sqrt{(Acc_i - Acc)^2 + (Stab_i - Stab)^2} \tag{13}$$

3) Fusing multiple diagnostic model in decision level based on ER rule

In part 3, every individual output result is a piece of evidence having its own distinctive reliability factor $r_i (i = 1, 2, ..., L)$ and importance weight $w_i (i = 1, 2, ..., L)$, and is fused with other outputs by using ER rule algorithm as described in (7) - (9). The reliability factor of every single model r_i is determined in part 2 by considering the model's accuracy and stability, while the importance weight w_i is set to be 1 initially, and optimized with historical data samples as illustrated in part 4. The diagnostic result after evidence fusion is in belief

distribution $O_f = \{(D_1, \beta_1), ..., (D_N, \beta_N)\}$, and the final label is determined by $L_f = \arg\max(\beta_1, \beta_2, ..., \beta_N)$, where N is the number of consequent attributes.

4) Determining the importance weight factor

The importance weight of every single model is set to be 1 initially which may not be accurate. It is therefore essential to fine tune the importance weights to improve the performance of the fused model by using an optimizing algorithm. As indicated in part 4 of Figure 1, $W = \{w_i | i = 1, 2, ..., L\}$ represents the parameters to be adjusted, and the number of parameters which should be optimized equals to the number of models to be fused by ER rule. As a matter of fact, wear fault diagnosis is a classification problem, and the misclassification rate should be as small as possible, hence misclassification error $1 - UA$ is defined as the objective function. Specifically, the predicted sample label S' is compared with the observed sample label S . It is believed that the sample is identified correctly by the fused diagnostic model if $S' = S$. The optimization model is defined by (14) and the importance weight should meet the constrain that $0 \leqslant w_i \leqslant 1$.

$$\begin{aligned} &\min \ \xi(W) = 1\text{-}UA \\ &\text{s. t.} \quad 0 \leqslant w_i \leqslant 1 \ \ (i = 1, 2, ..., L) \end{aligned} \tag{14}$$

Genetic algorithm is selected to optimize the fused model. The optimized importance weight $\tilde{W} = \{\tilde{w}_i | i = 1, 2, ..., L\}$ will be the final importance weight of every single diagnostic model in the fused model to identify testing samples.

5 Wear Fault Diagnosis of Marine Diesel Engine Based on The Fused model

Datasets on wear particles of diesel engines are used in this study to build the single and fused wear fault diagnostic models. The experimental wear particles were generated from an EQD XXX diesel engine, a ZH XXX diesel engine, and an abrasion testing machine. A total of 150 samples, containing cutting wear particles, spherical wear particles, fatigue spall particles, laminar particles, and severe sliding wear particles were obtained and prepared for analysis.

5.1 Brief description of single wear fault diagnostic model

As shown in Figure 2, developing single data-driven diagnostic model is the first step in the development of the fused diagnostic model. In our previous work, we focus on using different single models to take their advantages on identifying wear faults. Here, we give a

brief description of every single wear fault diagnostic model that we built previously. The three single diagnostic models are BBRB model, BANN model, and ER rule model.

1) BBRB wear fault diagnostic model

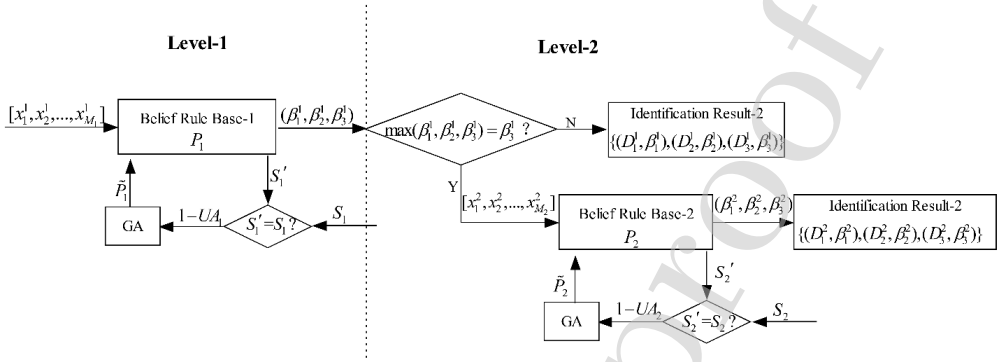


Figure 3. Structure of the BBRB wear fault diagnostic model

From literature review, fatigue spall particles, severe sliding wear particles and laminar particles are difficult to be distinguished by 2-D characteristics. A BBRB model as shown in Figure 3 is designed that each level can separately process 2-D and 3-D characteristics of wear particles. $[x_1^1, x_2^1, ..., x_{M_1}^1]$ and $[x_1^2, x_2^2, ..., x_{M_2}^2]$ denote the inputs of 2-D characteristics for the first BRB level and 3-D characteristics for the second BRB level. D_1 and D_2 are the outputs of the two levels, which can be represented by belief distributions, i.e. $D_1 = \{(D_1^1, \beta_1^1), (D_2^1, \beta_2^1), (D_3^1, \beta_3^1)\}$, $D_2 = \{(D_1^2, \beta_1^2), (D_2^2, \beta_2^2), (D_3^2, \beta_3^2)\}$. Specifically, in the output D_1, the severe sliding wear particles, fatigue spall particles and laminar particles are integrated into one category D_3^1 as they are difficult to be classified by 2-D characteristics. The output attribute with the maximum belief degree is considered as the identified wear particle type, i.e., $j = \arg\max\{\beta_1^j, \beta_2^j, \beta_3^j\}(j = 1,2)$. If β_3^1 is the maximum belief degree, the level-2 BRB will be activated. The three wear particles which are difficult to be classified in the first level BRB can be determined according to the output attribute with maximum belief degree. To improve the accuracy of the fault diagnostic model, the BBRB model is optimized by genetic algorithm as illustrated in Figure 3. Specifically, in the first level, the 2-D characteristics AR, D_e, R are represented by 2, 2, and 3 referential points respectively, and there are 12 rules in the first level belief rule base. In the second level, the 3-D characteristics S_a and S_{tdi} are represented by 2 and 4 referential points, and there are 8 rules in the second level belief rule base.

2) BANN wear fault diagnostic model

The BANN wear fault diagnostic model has the similar structure with the BBRB model as shown in Figure 4. Analogously, each level can separately process 2-D and 3-D characteristics of wear particles. 2-D particle characteristics $[x_1^1, x_2^1, ..., x_{M_1}^1]$ are used as the input of the level-1 ANN model, and 3-D particle characteristics $[x_1^2, x_2^2, ..., x_{M_2}^2]$ are used as the input of the level-2 ANN model. In the BANN model, if the output β_3^1 is largest, the second level ANN will be triggered. With the 3-D particle characteristics, the three wear particles fatigue spall particles, severe sliding wear particles and laminar particles can be further identified. As the BBRB model, the BANN model is also optimized by genetic algorithm. Considering the dataset used in wear fault diagnosis is not in larger scale, both the two levels of the BANN model only have one hidden layer respectively. The structures of the two levels ANN are 3-4-3 and 2-3-3, which means there are three, four, and three neuros in level-1 ANN and there are two, three, and three neuros in level-2 ANN. The learning rates of the two levels ANN are both 0.01.

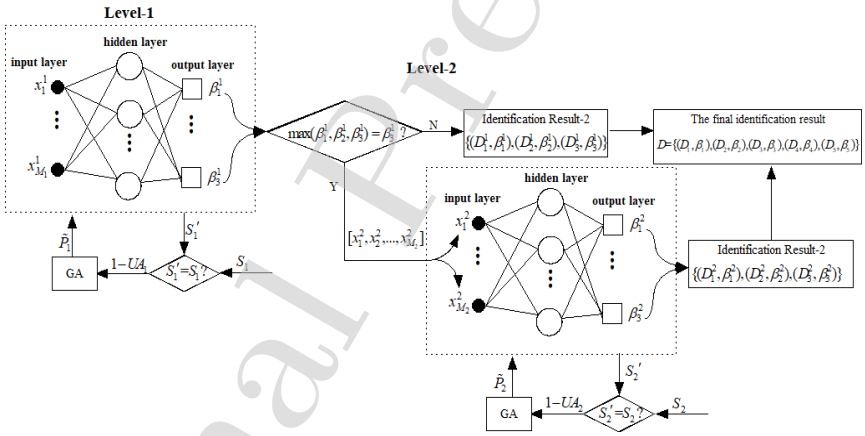


Figure 4. Structure of the BANN wear fault diagnostic model

3) ER rule wear fault diagnostic model

The 2-D and 3-D characteristics $[x_1, x_2, ..., x_M]$ are used as the input of the ER rule wear fault diagnostic model, and the categories of wear particles constitute the frame of discernment $\Theta = \{h_1, h_2, ..., h_N\}$. Every characteristic corresponds to one piece of evidence, and the belief distribution of evidence can be acquired by analyzing the historical wear fault samples statistically. Every piece of evidence is corrected by considering the reliability factor r_i and importance weight factor w_i, and fused by ER rule algorithm to generate the final

diagnostic result $\{(h_1,\hat{y}_1),...,(h_N,\hat{y}_N)\}$ which is in belief distribution. The final determined type is the one with the largest belief degree, i.e. $S_i'=\arg\max\{\hat{y}_1,\hat{y}_2,...,\hat{y}_N\}$. Being similar to BBRB and BANN model, the ER rule diagnostic model is also optimized by genetic algorithm, and the importance weight w_i of every piece of evidence will be fine-tuned. Particularly, the five characteristics of wear particles AR, D_e, R, S_a, S_{tdi} and are represented by 4,5,5,6,5 referential points. The reliability factors of the five characteristics are 0.9, 0.9, 0.9, 0.86, and 0.86 respectively.

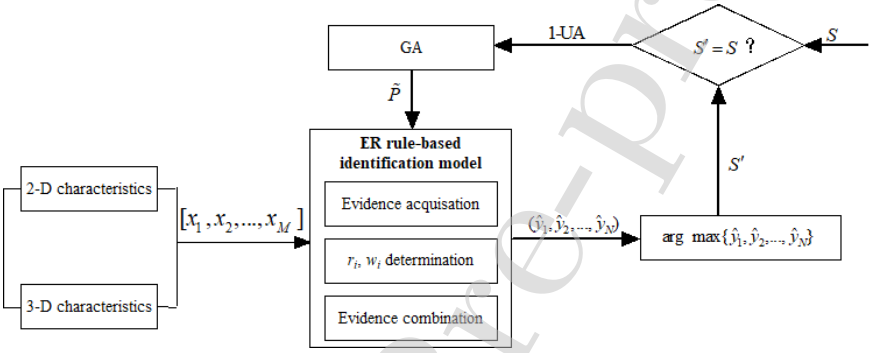


Figure 5. Structure of the ER rule wear fault diagnostic model

5.2 Wear fault diagnosis by fusing the multiple diagnostic models

As described in section 3.1, every single diagnostic model has its own output form which should be unified. Specifically, the output of BBRB model is composed of two parts: $D_1=\{(D_1^1,\beta_1^1),(D_2^1,\beta_2^1),(D_3^1,\beta_3^1)\}$ and $D_2=\{(D_1^2,\beta_1^2),(D_2^2,\beta_2^2),(D_3^2,\beta_3^2)\}$, where $\sum_{i=1}^{3}\beta_i^1=1$ and $\sum_{i=1}^{3}\beta_i^2=1$. D_1^1, D_2^1, D_3^1 represents cutting wear particles, spherical wear particles, and SBL particles which is the collection of severe sliding wear particle, fatigue spall particle and laminar particle. D_1^2, D_2^2, D_3^2 represents severe sliding wear particle, fatigue spall particle and laminar particle respectively. The output of BANN model is $D=\{(D_1,\beta_1),(D_2,\beta_2),(D_3,\beta_3),(D_4,\beta_4),(D_5,\beta_5)\}$, where $\beta_i=1$ and $\beta_k=0\ \forall k=\{1,...,5\}\setminus i$. D_1-D_5 represent the five wear particle types. ER rule model has the same output form with BANN model, i.e. $D=\{(D_1,\beta_1),(D_2,\beta_2),(D_3,\beta_3),(D_4,\beta_4),(D_5,\beta_5)\}$, but $\sum_{i=1}^{5}\beta_i=1$. To make the three

models have the same output form, the output of BBRB model will be expressed as (15), while the outputs of BANN and ER rule models will be in the form of (16). D_1-D_6 represents the wear particles in the order sever sliding wear particle, cutting wear particle, fatigue spall particle, laminar particle, spherical particle, and SBL particle. Particularly, since SBL particle is the collection of severe sliding wear particle, fatigue spall particle and laminar particle, $\underset{i=1,3,4}{\forall} D_i \cap D_6 = D_i$, when we are calculating the orthogonal sum of evidence.

$$D = \begin{cases} \{(D_1:0),(D_2:\beta_1^1),(D_3:0),(D_4:0),(D_5:\beta_2^1),(D_6:\beta_3^1)\} & \beta_3^1 \neq \max_{i=1,2,3}(\beta_i^1) \\ \{(D_1:\beta_1^2),(D_2:0),(D_3:\beta_2^2),(D_4:\beta_3^2),(D_5:0),(D_6:0)\} & \beta_3^1 = \max_{i=1,2,3}(\beta_i^1) \end{cases} \tag{15}$$

$$D = \begin{cases} \{(D_1,\beta_1),(D_2,\beta_2),(D_3,\beta_3),(D_4,\beta_4),(D_5,\beta_5),(D_6,\beta_6)\} & \beta_6 = 0,\ \beta_i = 1 \ and \\ & \beta_k = 0\ \forall k = \{1,...,5\} \setminus i \\ \{(D_1,\beta_1),(D_2,\beta_2),(D_3,\beta_3),(D_4,\beta_4),(D_5,\beta_5),(D_6,\beta_6)\} & \sum_{i=1}^{5}\beta_i = 1 \ and\ \beta_6 = 0 \end{cases} \tag{16}$$

Five-fold cross validation is conducted in all the three single diagnostic models. Table 2 lists the *UA* value of every model for every fold testing dataset. From Table 2, it can be known that the average *UA* values of BBRB, BANN and ER rule model are 0.893,0.920,0.933. Based on (11) - (13), the stability of the three models are 0.984,0.885,0.973. As a result, the performance index of BBRB, BANN, ER rule model is $PI_1 = \{(Acc_1:0.893),(Stab_1:0.984)\}$, $PI_2 = \{(Acc_2:0.920),(Stab_2:0.885)\}$, and $PI_3 = \{(Acc_3:0.933),(Stab_3:0.973)\}$ respectively. According to (10), the reliability factors of the three models are $r_1 = 0.892$, $r_2 = 0.860$, $r_3 = 0.928$ when their outputs are used as three pieces of evidence to be fused by ER rule algorithm in decision level.

Table 2 *UA* values of every fold testing dataset for the three models

Models	1fold	2fold	3fold	4fold	5fold	Average
BBRB	0.933	0.900	0.900	0.867	0.867	0.893
BANN	0.933	0.933	0.833	0.900	1.000	0.920
ER rule	0.933	0.967	0.900	0.867	1.000	0.933

In the five-fold cross validation, the importance weight of every piece of evidence is initialized to be 1, i.e. $w_1 = w_2 = w_3 = 1$. The importance weight is fine-tuned according to (14), and the optimization result is shown in Table 3. From Table 3, it can be seen that the importance weights of BBRB model and ER rule model are steady in the five-fold cross validation, and quite similar, indicating that the two models play the similar roles in the fused diagnostic model. The importance weights of BANN in the first four folds testing datasets are quite small, because the belief degree of every consequence attribute in the output is either 0

or 1 so that the mis-identified result will significantly decrease the accuracy of the fused model. Oppositely, BANN model has the largest importance weight due to its high accuracy in the fifth fold testing dataset.

Table 3 Importance weight of every fold testing dataset for the three models

Models	1fold	2fold	3fold	4fold	5fold
BBRB	0.839	0.979	0.849	0.902	0.630
BANN	0.074	0.104	0.050	0.157	0.898
ER rule	0.782	0.871	0.934	0.790	0.702

6 Results of the Fused Diagnostic Model and Discussions

6.1 Accuracy analysis

Figure 6 shows the accuracy of the fused diagnostic model and the three single diagnostic model in the five-fold cross-validation. It can be seen from Figure 6 that the accuracy of the diagnostic result can be improved after integrating the three single data-driven models by the ER rule method. Specifically, the accuracy of the fused model on the fold-1, fold-2 and fold-4 testing datasets is 1, 1, 0.933 respectively, which is increased by 7%, 10% and 6.6% compared with the smallest accuracy given by the single models. Considering that the BANN model and the ER rule model can make correct identification on the fold-5 testing dataset, the poor diagnostic result given by BBRB model in this dataset can be improved, and all samples can be distinguished by the fused model consequently. Similarly, the good performance of BBRB model and ER rule model on fold-3 testing dataset (i.e. UA_{BBRB}=0.9 ; $UA_{\text{ER rule}}$=0.9) can compensate that of the BANN model on this dataset after the three models are integrated.

Table 4 lists the samples of the testing datasets that are mis-classified by the three single models and the fused model in the five-fold cross-validation. As described in Table 4, the samples mis-identified by one single model can be corrected by the fused model. For example, for the fold-1 testing dataset, the BBRB model can correct the identification result of samples 14 and 17 which are wrongly classified by the BANN model and the ER rule model. On the other hand, the BANN model and the ER rule model can correct the diagnostic result given by the BBRB model on sample 16. Additionally, if a single model has an outstanding capability of identifying a specific type of wear particle, it can compensate the deficient capabilities of other single models. Taking the fold-2 testing dataset as an example, the ER rule model can correctly identify the laminar particles. While samples 23 and 24 are mis-classified by the BBRB model and samples 19 and 20 are mis-classified by BANN model, they all can be

accurately determined by the fused diagnostic model because of the satisfying performance of the ER rule model on identifying laminar particles. However, it should be noticed that the fused diagnostic model cannot give a correct diagnostic result if all single diagnostic models make wrong identification for a specific sample, which means that the performance of the fused model is determined by the single models, but it will not be inferior to that of the single model.

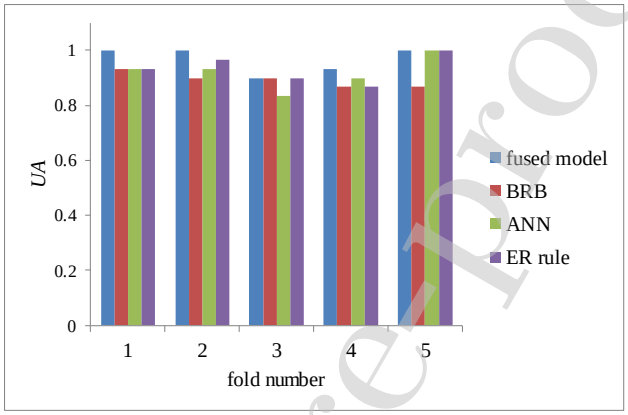


Figure 6. *UA* of the fused and the three single data-driven diagnostic models

Table 4 Mis-classified samples in five-fold cross-validation*

Fold number	Diagnostic model	Mis-classified samples
Fold 1	BBRB	(Sample 16 : FS) , (Sample 19: L)
	BANN	(Sample 14 : FS) , (Sample 17: FS)
	ER rule	(Sample 14 : FS) , (Sample 17: FS)
	Fused model	/
Fold 2	BBRB	(Sample 18 : FS) , (Sample 23 : L) (Sample 24 : L)
	BANN	(Sample 19 : L) , (Sample 20 : L)
	ER rule	(Sample 18 : FS)
	Fused model	/
Fold 3	BBRB	(Sample 11 : C) , (Sample 17 : FS) (Sample 18 : FS)
	BANN	(Sample 11 : C) , (Sample 13 : FS) (Sample 14 : FS) (Sample 15 : FS) (Sample 16 : FS)

Fold number	Diagnostic model	Mis-classified samples
	ER rule	(Sample 11：C), (Sample 17：FS) (Sample 19：L)
	Fused model	(Sample 11：C), (Sample 17：FS) (Sample 18：FS)
Fold 4	BBRB	(Sample 7：C), (Sample 18：FS) (Sample 23：L), (Sample 24：L)
	BANN	(Sample 7：C), (Sample 9：C) (Sample 17：FS)
	ER rule	(Sample 7：C), (Sample 14：C) (Sample 17：FS), (Sample 18：FS)
	Fused model	(Sample 7：C), (Sample 17：FS)
Fold 5	BBRB	(Sample 5：SSL), (Sample 17：FS) (Sample 23：L), (Sample 24：L)
	BANN	/
	ER rule	/
	Fused model	/

* C-cutting wear particles, L-laminar particles, FS-fatigue spall particles; SP-spherical wear particles, SSL-severe sliding wear particles.

The fused diagnostic model trained by the fold-1 training dataset is selected as the final diagnostic model, and the reliability and importance weight of the three single diagnostic models are $r_{\mathrm{BBRB}}=0.892$, $r_{\mathrm{BANN}}=0.860$, $r_{\mathrm{ER\ rule}}=0.928$, $w_{\mathrm{BBRB}}=0.839$, $w_{\mathrm{BANN}}=0.074$, $w_{\mathrm{ER\ rule}}=0.782$. Figure 7 shows the square error $SE_i(i=1,2,...,n)$ between the predicted belief distribution and the observed belief distribution. For the whole dataset, the SE_i of only two samples are over 0.1, as highlighted by the elliptic curve after integration of the single models, while the SE_i of the other samples are all below 0.1, and are smaller than the SE_i generated by the BBRB model and the ER rule model. It can be concluded that the fused diagnostic model developed in this paper has higher accuracy for wear particle identification, and could increase the difference between the belief degree of the real wear particle type and that of other wear particle types, making the identification results more credible.

Additionally, the performance of the fused diagnostic model under the effects of disturbance signals is verified. For every characteristic of wear particles, a random noise

following uniform distribution $x \sim U(a,b)$ is added to the original testing data samples, where a equals to negative ten percent of the mean input value of the whole dataset, and b equals to positive ten percent of the mean input value of the whole dataset. The accuracy of every single diagnostic model is 0.867 (BBRB), 0.933 (BANN), and 0.833 (ER rule) respectively, and the accuracy of the fused diagnostic model is 0.933. Although compared with the results described in Figure 6, disturbance signals reduce the performance of every model slightly, the result is acceptable, and the diagnostic accuracy can be still improved by fusing the three single models.

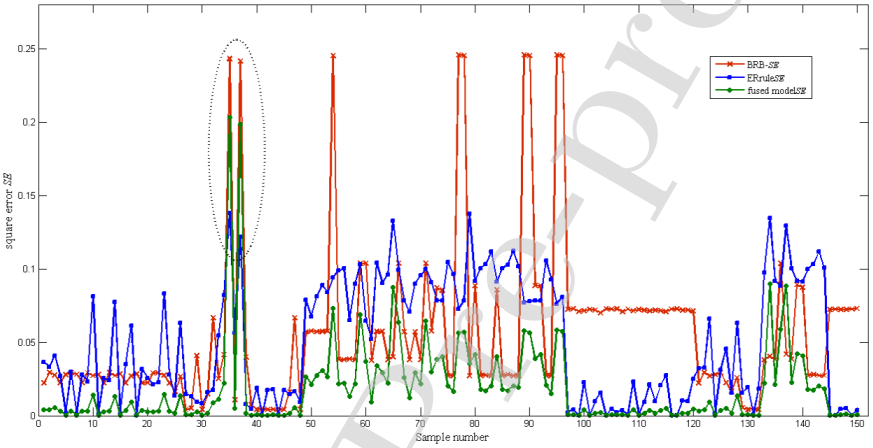


Figure 7. *SE* of every sample in the whole dataset determined by fused diagnostic model and the three single diagnostic models

The performance of the fused diagnostic model is further verified by a verification dataset containing 39 samples which is independent of the training and testing datasets. These 39 samples were collected during experiments conducted in a four-stroke diesel engine. Figure 8 shows the diagnostic result of every sample in the verification dataset given by the fused model and the three single models which are represented by different colors and sizes. From Figure 8, it can be seen that the fused model cannot identify the wear fault correctly if all or most single models give wrong diagnostic results, such as sample 1, 20, 25, 28 which are highlighted by black arrows. It means that the performance of the fused model is influenced by single model's performance. However, it can be noticed that the overall diagnostic accuracy can be improved by fusing the single models. Samples 2, 6, 13, 16, 30, and 32 mis-identified by ER rule model, samples 12, 14, 21, 26, 33, and 37 mis-identified by BBRB models, and samples 23, 24, and 36 mis-identified by BANN models are all identified correctly by the fused model. Furthermore, Table 5 shows the accuracy and mean square error

(*MSE*) of all the diagnostic models on verification dataset. The accuracy of the fused model increases by 12.8%, 7.7% and 10.2% compared with that of the BBRB model, the BANN model and the ER rule model. The *MSE* of the fused model is also decreased clearly. From Figure 9, it can be seen that the belief distribution of most samples predicted by the fused diagnostic model is closer to the observed belief distribution, and the difference between the predicted value and the observed value is also smaller than that of the BBRB model and the ER rule model except for the six samples as marked by circles which are similar to the points marked by arrows in Figure 8.

Table 5 Performance of the fused the three single diagnostic model on verification dataset

Indicator	BBRB	BANN	ER rule	Fused model
MSE	0.418	-	0.488	0.271
UA	0.718	0.769	0.744	0.846

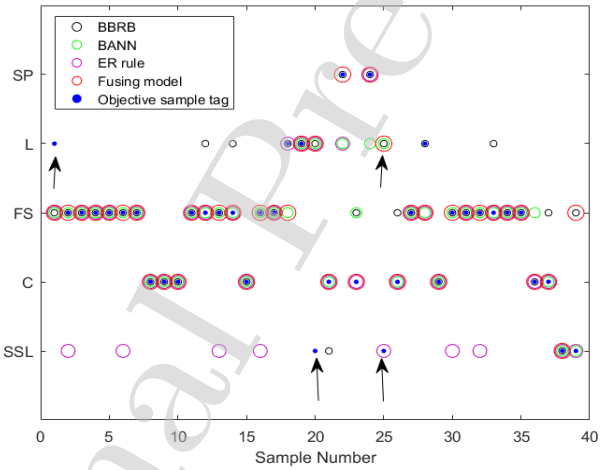


Figure 8. Diagnostic results of the verification dataset given by the fused diagnostic model and three single diagnostic models

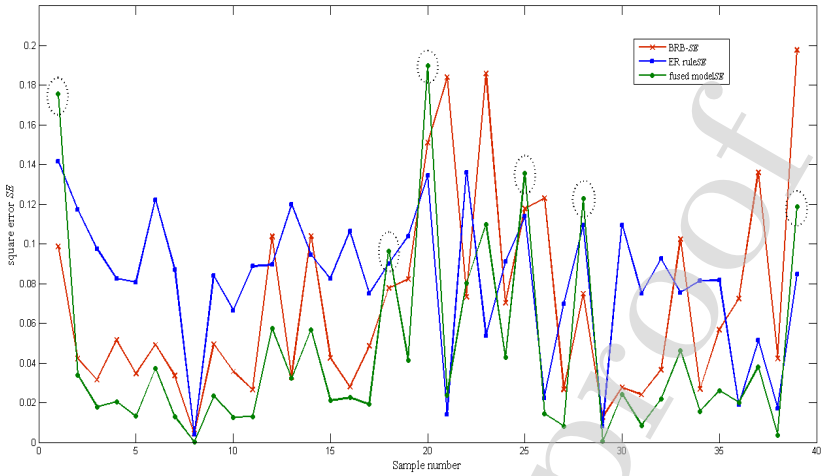


Figure 9. *SE* of every sample in the verification dataset determined by the fused diagnostic model and the three single diagnostic models

6.2 Computational complexity analysis

With data scale increasing, the time that an algorithm processes the data rises as well. The computational complexity of an algorithm directly affects its efficiency and flexibility. Here, we analyze the computational complexity of ER rule, BRB, and BP-ANN algorithms. Suppose that *m* characteristics of wear particles (input) and *n* wear fault modes (output) are considered in wear fault diagnosis. It means *n* is the number of elements in the frame of discernment (FoD), and $N=2^n-1$ is the number of elements in the power set of the FoD.

For ER rule method, *m* is the number of pieces of evidence to be fused because every characteristic of wear particle is used as a piece of evidence. Since we combine 2 pieces of evidence every time, we need to find the intersection between any pairs of elements in the two power sets (one for each piece of evidence) which gives us N^2 computations. *m*-1-time combinations have to be conducted to complete the fusion of the *m* pieces of evidence. Consequently, the computational complexity of ER rule algorithm is $O((m-1)(2^n-1)^2)$. Because the power set is not considered in wear fault diagnosis, the computational complexity can be simplified as $O((m-1)n^2)$.

For BRB method, the number of rules activated by the input data point is 2^m. Being similar to ER rule algorithm, only 2 rules are combined every time, and therefore, N^2 (i.e. $(2^n-1)^2$) computations are conducted. The computational complexity of BRB can be represented as $O(2^m(2^n-1)^2)$ which can be simplified as $O(2^m n^2)$.

For BP neural network, suppose *L* hidden layers are developed and each hidden layer has $N_l(l=1,2,...,L)$ neurons. The number of neurons in input layer and output layer are *m* and n

respectively. The number of multiplications between input layer and the first hidden layer is mN_1, and the number of multiplications between output layer and the last hidden layer is nN_L. The number of multiplications between every two hidden layers is $N_l N_{l+1}$. As a result, the computational complexity of BP-ANN is $O(mN_1 + nN_L + \sum_{l=2}^{L-2} N_l N_{l+1})$. Considering only one hidden layer is developed in wear fault diagnostic model, the model computational complexity can be simplified as $O(N(m+n))$.

From the above analysis, it can be found that ER rule and BRB have similar computational complexity, but the rising number of input features will obviously increase the model complexity and computing time of BRB. The number of hidden layers and neurons in each layer determine the computational complexity of BP-ANN method, and the more hidden layers, the more complexity.

7 Conclusions

A machine learning-based wear fault diagnostic model is proposed by using the ER rule to integrate three single data-driven diagnostic models which are BBRB model, BANN model, and ER rule model. The reliability and importance weight of every single model is considered respectively. A fully reliable model is defined and Euler distance is used to determine the single model's reliability, while the importance weight of every single model is optimized by genetic algorithm. Five-fold cross validation is conducted to evaluate the effectiveness of the fused diagnostic model, and a verification dataset which is independent with the training and testing datasets is used to further verify the performance of the fused model. The advantages of the fused diagnostic model proposed in this paper are summarized as follows:

1) The performance of wear fault diagnostic model can be enhanced from different aspects by fusing the three data-driven wear fault diagnostic models. The demerits of the single diagnostic models can be overcome
2) The inherent property of the individual model can be better expressed by involving model stability to determine the model reliability, because the model stability is generally related to the model structure, modeling algorithm, inference process and etc.
3) The ER rule algorithm distinguishes model reliability and model importance weight, therefore, not only the performance of every single model will have an effect on the final output of the fused model, the relationship among all the individual models will also influence the fusing model's performance.

4) The fused diagnostic model is more accurate and robust, and the fault tolerance ability can be improved remarkably compared with a single data-driven diagnostic model.

Acknowledgements

We acknowledge financial support from the NSFC (No.61903108), the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (U1709215), the NSFC (No.61433001,71601180), the Science & Technology Project of Zhejiang Province (No.2019C03104,2018C04020). And Open Fund of National Engineering Research Center for Water Transport Safety (No.A2019007).

Reference:

[1] L. A.Malm, A. Hultman, and J. Enstrom, "Main Engine Damage," Swedish Club, 2015.

[2] Yan X P. Condition Monitoring and Fault Diagnosis for Mechanical System [M]. Wuhan: Wuhan University of Technology Press, 2009.

[3] Jiang R Y, Yan X P. Condition monitoring of diesel engines[J]. Complex System Maintenance Handbook, 2008：533-557.

[4] Yan X P, Xu X J, Sheng C X, et al. Intelligent wear mode identification system for marine diesel engines based on multi-level belief rule base methodology[J]. Measurement Science & Technology, 2018, 29(1)：1-13.

[5] Xu X J, Yan X P, Sheng C X, et al. Identification on wear mode for marine diesel engine based on evidential reasoning rule [J]. Tribology, 2017,11(6)：814-822.

[6] Katsoulakos P S, Newland J, Stansfield J T, et al. Monitoring, databases and expert systems in the development of engine fault diagnostics[J]. British Journal of Nondestructive Testing, 1988,30(4)：263-273.

[7] Autar R K. Computer aided maintenance of diesel engine by use of an expert system[C]. Australia：International Mechanical Engineering Congress and Exhibition, 1991.

[8] Xu K, Luxmoore A R, Jones L M, et al. Integration of neural networks and expert systems for microscopic wear particle analysis[J]. Knowledge-Based Systems, 1998, 11(3)：213-227.

[9] Tasdemir S, Saritas I, Ciniviz M, et al. Artificial neural network and fuzzy expert system comparison for prediction of performance and emission parameters on a gasoline engine[J]. Expert Systems with Applications, 2011, 38(11)：13912-13923.

[10] Basurko O C, Uriondo Z. Condition-based maintenance for medium speed diesel engines used in vessels in operation[J]. Applied Thermal Engineering, 2015, 80：404-412.

[11] Guo Z W, Yuan C Q, Li Z X, et al. Condition identification of the cylinder liner-piston ring in

a marine diesel engine using bispectrum analysis and artificial neural network[J]. Insight：Non-destructive Testing and Condition Monitoring, 2013, 55(11)：621-626.

[12] Li Y, Pont M J, Jones N B, et al. Applying MLP and RBF classifiers in embedded condition monitoring and fault diagnosis systems[J]. Transactions of the Institute of Measurement & Control, 2001, 23(5)：315-343.

[13] Wu J D, Chiang P H, Chang Y W, et al. An expert system for fault diagnosis in internal combustion engines using probability neural network[J]. Expert Systems with Applications, 2008, 34(4)：2704-2713.

[14] Han J, Li X D, Xia L, et al. Application of rough set based fuzzy neural network in fault diagnosis [J]. Journal of Hefei University of Technology, 2012, 35(5)：577-580.

[15] G. P. S. Ã, G. W. Stachowiak, and P. Podsiadlo. Automated classification of wear particles based on their surface texture and shape features[J]. Tribology International, 2008, 41：34–43.

[16] Zeng R L, Zhang L L, Xiao Y K, et al. A method combining order tracking and fuzzy c-means for diesel engine fault detection and isolation[J]. Shock and Vibration, 2015, 1-7.

[17] Zeng Y H, E J Q, Zhu H, et al. Fault diagnosis on cooling system of ships diesel engine based on bayes network classifier[J]. Journal of Central South University (Science and Technology), 2010,41(4)：1379-1384.

[18] Xu X J, Yan X P, Sheng C X, et al. A Belief Rule-Based Expert System for Fault Diagnosis of Marine Diesel Engines[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, PP(99)：1-17.

[19] Xu L, Krzyzak A, Suen C Y. Methods of combining multiple classifiers and their applications to handwriting recognition[J]. IEEE transactions on systems, man, and cybernetics, 1992, 22(3)：418-435.

[20] Lu Y, Shi P F, Zhao Y M. Voting Principle for Combinnation of Multiple Classifiers[J]. Journal of Shanghai Jiaotong University,2000,34(5)：680-683.

[21] Kuncheva L I. A Theoretical Study on Six Classifier Fusion Strategies[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(2)：281-286.

[22] Kittler J. Improving recognition rates by classifier combination: A theoretical framework[J]. Frontiers of Handwriting Recognition, 1997, 5：231-247.

[23] Pizzi N J, Pedrycz W. Aggregating multiple classification results using fuzzy integration and stochastic feature selection[J]. International Journal of Approximate Reasoning, 2010, 51(8)：883-894.

[24] Kuncheva L I, Bezdek J C, Duin R P W. Decision templates for multiple classifier fusion: an

experimental comparison[J]. Pattern Recognition, 2001, 34(2)：299-314.

[25] Liu Z, Pan Q, Dezert J, et al. Classifier fusion with contextual reliability evaluation[J]. IEEE transactions on cybernetics, 2017, 48(5)：1605-1618.

[26] Lam L, Suen C Y. Optimal combinations of pattern classifiers[J]. Pattern Recognition Letters, 1995, 16(9)：945-954.

[27] Huang Y S, Suen C Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1995 (1)：90-94.

[28] Verlinde P, Cholet G. Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application[C]. International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA). 1999：188-193.

[29] Xu L, Krzyzak A, Suen C Y. Methods of combining multiple classifiers and their applications to handwriting recognition [J]. IEEE transactions on systems, man, and cybernetics, 1992, 22(3): 418-435.

[30] Rogova G. Combining the results of several neural network classifiers[M]. Classic Works of the Dempster-Shafer Theory of Belief Functions. Springer, Berlin, Heidelberg, 2008：683-692.

[31] Yang J B, Xu D L. Evidential reasoning rule for evidence combination[J]. Artificial Intelligence, 2013, 205：1-29.

[32] Xu X B, Zheng J, Xu D L, et al. Information fusion method for fault diagnosis based on evidential reasoning rule[J]. Control theory and application, 2015, 32(9)：1170-1182.

[33] Xu D L, Zhang Y, Yang J B. Probability of natural disasters：a forecasting model based on data and the evidential reasoning rule[C]. Exeter：6th International BAASANA Conference, 2016.

[34] Zhu H Y, Yang J B, Xu D L, et al. Application of evidential reasoning rules to identification of asthma control steps in children[C]. Colchester：International Conference on Automation and Computing, 2016：444-449.

[35] Zhao Z N, Qiao P L, Wang J, et al. Security situation assessment of all-optical network based on evidential reasoning rule[J]. Mathematical Problems in Engineering, 2016(4)：1-7.

[36] Xu X B, Zheng J, Yang J B, et al. Data classification using evidence reasoning rule[J]. Knowledge-Based Systems, 2017, 116：144-151.

[37] Last M, Maimon O, Minkov E. Improving stability of decision trees[J]. International Journal of Pattern Recognition & Artificial Intelligence, 2002, 16(2)：145-159.

Highlights (for review)

We think there are three highlights in this paper:

1) The demerits of single data-driven diagnostic models in fault diagnosis can be overcome by fusing their outputs in decision level;

2) Evidential reasoning (ER) rule distinguishes reliability factor and importance weight factor of single diagnostic models when they are fused;

3) a new method considering model accuracy and stability simultaneously has been proposed to determine the reliability factor of every single diagnostic model.

***conflict of Interest Statement**

Declaration of interests

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the officer). He/she is responsible for communicating with other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from yangjianbo129@163.com.

Signed by all authors as follows:

Xiaojian Xu	xxj03@hdu.edu.cn	Xiaojian Xu	2019.11.26
Zhuangzhuang Zhao	252981680@qq.com	Zhuangzhuang Zhao	2019.11.26
Xiaobin Xu	xuxiaobin1980@163.com	Xiaobin Xu	2019.11.26
Jianbo Yang	yangjianbo129@163.com	Jianbo Yang	2019.11.26
Leilei Chang	leileichang@hotmail.com	Leilei Chang	2019.11.26
Xinping Yan	xpyan@whut.edu.cn	Xinping Yan	2019.11.26
Guodong Wang	guodong.wang@tuwien.ac.at	Guodong Wang	2019.11.26