Deep Unsupervised Transfer Learning for Health Status Prediction of a Fleet of Wind Turbines with Unbalanced Data

Dandan Peng[1, 2], Chenyu Liu[1, 2], Wim Desmet[1, 2], Konstantinos Gryllias[1, 2]

[1]*Department of Mechanical Engineering, Faculty of Engineering Science, KU Leuven*

[2]*Dynamics of Mechanical and Mechatronic Systems, Flanders Make, Celestijnenlaan 300, BOX 2420, 3001 Leuven, Belgium*
konstantinos.gryllias@kuleuven.be

ABSTRACT

The condition monitoring and health status prediction of a fleet of wind turbines are essential for the safety of wind turbines. At present, the Supervisory Control And Data Acquisition (SCADA) system has been widely used in wind turbines, which can monitor and collect various physical information and sensor information of wind turbines in real-time. Due to the fact that the amount of data obtained by SCADA systems is extremely large, developing an intelligent decision-making system based on deep learning is a very valuable research. Therefore, this paper is committed to exploring a health status prediction algorithm of wind turbines based on deep learning and SCADA systems. However, yet in actual industrial applications, it is very time-consuming and expensive to obtain a large amount of labeled data. In addition, as failures rarely occur, there is a serious sample imbalance problem in the datasets. More importantly, due to the difference in working environment and physical parameter setting, there are significant differences in the feature distribution of different wind turbines data, which leads to a significant drop in the performance of the deep learning model on unknown wind turbines.

Therefore, an unsupervised transfer learning algorithm based on Generative Adversarial Networks for wind turbine health status prediction (WT-GAN) is proposed. WT-GAN can not only remove the domain shift between wind turbines, but also it is an unsupervised learning method. This means that only the unlabeled data for the target domain is required, which solves the problem of labeling data. In order to evaluate the effectiveness of WT-GAN on the condition monitoring of a fleet of wind turbines, this method is applied to one dataset about blade icing detection of wind turbines. The experimental results prove that the proposed method can predict the health status of the wind turbine well. In addition, it can significantly reduce the domain shift among different wind turbines, thereby achieving excellent performance on unknown wind turbines.

1. INTRODUCTION

As a kind of renewable energy, wind energy has been valued by countries all over the world. In the past few decades, the installation capacity of wind turbines has experienced significant growth. However, wind turbines are usually built in remote and harsh environments. In addition, the wind turbines bear alternating loads for a long time, which makes them prone to failure. These problems significantly increase the maintenance and operating costs of wind turbines. Therefore, in order to ensure their normal operation and reduce maintenance costs, it is very important to monitor their health status (Márquez et al., 2016).

The gearbox, the generator, the main bearing, the blades and the tower of wind turbines are all key components that are prone to failure. At present, there are mainly three fault diagnosis methods for these key components. One is the sensor-based fault diagnosis method. Specifically, the information of the corresponding components is collected through sensors, and then the information is analyzed and processed to obtain the health status of the key components. These sensors mainly include vibration sensors, ultrasonic sensors, strain sensors, and acoustic emission sensors (Davis et al., 2016 and Muñoz et al., 2016). The second one is the image-based fault diagnosis method (Yang et al., 2021). This method uses drones to obtain images of the corresponding components of the wind turbine, and then the obtained images are analyzed to obtain the health status of the wind turbine. However, this method is not sensitive to weak faults. Only when the fault develops to be very serious, the obtained image can accurately reflect the fault situation. The third one is the fault diagnosis method based on the SCADA system (Chen et al., 2019). This method has been widely used in the condition monitoring of wind turbines. The SCADA system can collect various physical parameters of the wind turbine, such as temperature, rotational speed, wind speed, etc., which can provide data support for the subsequent health status

prediction of the wind turbine. Compared with other methods, the data obtained by the SCADA system is more comprehensive. Therefore, this paper is committed to the development of wind turbine condition monitoring method based on SCADA systems.

The SCADA system can easily obtain massive amounts of monitoring data, and deep learning technology has obvious advantages in big data processing and feature extraction (Peng et al., 2020 & Liu et al., 2020). Therefore, deep learning has been widely used in data analysis and feature extraction of SCADA systems (Peng et al., 2021). Deep learning technology does not require expert domain knowledge. It can directly obtain valuable features from a large amount of raw data through a multi-layer non-linear architecture, and automatically and accurately obtain diagnostic results. This characteristic leads to the achievement of excellent performance in the task of wind turbine health status prediction. For instance, Chen et al. (2018) proposed a deep neural network method based on SCADA data to detect whether wind turbine blades are icing. Pang et al. (2020) proposed a novel spatio-temporal fusion neural network for wind turbine health status prediction and obtained a high fault diagnosis result. Although these methods have achieved excellent diagnostic performance, they are all based on the assumption that the training and test data have the same feature distribution. In addition, they require a large amount of labeled data for training. Therefore, in actual situations, these methods face the following challenges:

1) In actual industrial applications, labeled data are very expensive and scarce. The SCADA system of each wind turbine will generate a large amount of monitoring data. It is time-consuming and laborious to analyze these data prepare a dataset.

2) Because of the different mechanical parameters, electrical parameters and working environment of wind turbines, each wind turbine has its unique characteristics, which results in the different feature distributions among wind turbines. Due to the domain shift problem, it is difficult to apply the existing methods to the health status prediction of unknown wind turbines.

The abovementioned challenges lead to the following problems that need to be resolved:

1) How to fully train a network model and make it have excellent health status prediction performance with a few labeled samples.

2) How to solve the domain shift problem among different wind turbines, so that the diagnosis model has good generalization performance and can be used for the health status prediction of unknown wind turbines.

As an excellent deep learning paradigm, transfer learning can transfer the knowledge of the source domain to different but related target domains (Pan et al., 2009). Therefore, it provides an effective solution for the cross-domain fault diagnosis of different wind turbines. When the source domain is labeled data and the target domain is not labeled, the concept of unsupervised transfer learning is proposed. Wen et al. (2017) proposed a bearing fault diagnosis method based on unsupervised transfer learning. This method introduces the Maximum Mean Difference (MMD) to the sparse autoencoder to reduce the difference between the source domain and the target domain. Ganin et al. (2015) proposed a deep unsupervised transfer learning method with a domain adversarial training strategy, namely: Domain Adversarial Neural Network (DANN). The basic architecture of DANN is the Generative Adversarial Network (GAN), which adversarial trains a feature extractor and a domain discriminator to extract domain invariant features between different domains (Liu et al., 2020). It has better adaptability than the MMD method.

Therefore, this paper proposes an unsupervised transfer learning algorithm based on Generative Adversarial Networks for wind turbine health status prediction (WT-GAN). WT-GAN can significantly reduce the domain shift of among different wind turbines, and realize the condition monitoring of unknown wind turbines. This deep unsupervised transfer learning paradigm not only solves the problem of the lack of labeled data, but also solves the domain shift problem among different wind turbines. However, due to the scarcity of failure situations, the obtained SCADA data only contains a small number of fault samples, most of which are healthy samples. Therefore, the data set has a serious sample imbalance problem. Therefore, this paper introduces the focal loss function (Lin et al., 2017) to balance healthy samples and faulty samples instead of the cross entropy loss function, thereby improving the performance of WT-GAN on unbalanced datasets. The effectiveness of the method is verified on a case of blade icing detection of wind turbines. The main contributions of this paper are summarized as follows:

1) A wind turbine fault diagnosis model based on deep unsupervised transfer learning is proposed. This method uses deep learning to automatically extract highly abstract features from SCADA data, thereby avoiding manual selection of features.

2) This method can significantly reduce the domain shift between the source domain and the target domain, thereby improving the performance of the model on unknown wind turbines.

3) This method solves the data imbalance problem faced by wind turbine fault diagnosis.

4) In a real wind turbine blade icing case, the proposed method has obtained excellent diagnostic performance, and it also has excellent performance on unknown wind turbines.

The rest of this paper is organized as follows. Section 2 introduces the transfer learning problem of a fleet of wind turbines. In section 3, the proposed method is described. Section 4 validates the effectiveness of the proposed method on the blade icing detection of a wind farm. Finally, conclusions are drawn in Section 5.

2. PRELIMINARIES

2.1. Problem Definition

In order to clearly describe the problem studied in this article, the following introduces the basic symbols of transfer learning. First, given the source domain D_S and the target domain D_T, their feature distributions are different. Suppose $X_S = \{x_S^1, x_S^2, \cdots x_S^{n_S}\}$ is the sample of the source domain, where n_S denotes the total sample number in the source domain; $Y_S = \{1, 2, \cdots, K\}$ is the label corresponding to the source domain samples, where K is the total number of fault types of the wind turbine. Therefore, the source domain can be defined as $D_S = \{X_S, Y_S\}$. The distribution corresponding to the source domain sample is $P_S(X_S)$. For deep unsupervised transfer learning, the target domain is unlabeled data samples. Therefore, the target domain is defined as $D_T = \{X_T\}$, where $X_T = \{x_T^1, x_T^2, \cdots x_T^{n_T}\}$ and n_T represents the total sample number in the target domain. The distribution corresponding to the target domain sample is $P_T(X_T)$. Since there are domain shifts between different domains, $P_S(X_S) \neq P_T(X_T)$. This results in the model trained in the source domain not being applied to the target domain. It is worth noting that the source domain and the target domain share the same label space. The goal of this research is to reduce the difference between the source domain and the target domain as much as possible, so that the diagnostic model can be applied to unknown wind turbines.

2.2. Convolutional Neural Network

Convolutional Neural Networks (CNNs) benefit from their excellent automatic feature learning capabilities and have achieved remarkable success in the fields of image analysis, speech recognition and fault diagnosis (Chen et al., 2019 & Wang et al., 2019). Convolutional layers, pooling layers and activation functions are important components of CNNs. The convolutional layer is composed of multiple convolution kernels, and the size of each convolution kernel is fixed. The number of convolution kernels determines the number of feature maps output by the convolutional layer. Supposing that the input of the i-th convolutional layer is M_{i-1}, the feature $M_i^{conv} = W_i * M_{i-1} + B_i$ after the convolution operation, where W_i is the weight of the convolution kernel, B_i is the bias, and $*$ is the convolution operation. Then the ReLU function (Nair et al., 2010) is used to improve the nonlinear learning ability of the network. The convolutional layer is usually followed by a pooling layer to reduce the feature dimension, thereby increasing the training speed of the network and preventing the network from overfitting. After multiple convolutional layers, the obtained features will be fed to the fully connected layers and the Softmax layer to complete the classification task. The Softmax layer maps the features to the range of (0, 1) and outputs the predicted probability distribution. The Softmax function is expressed as:

$$z_j = \frac{e^{y_j}}{\sum_{k=1}^{K} e^{y_j}}, \ for\ j = 1,\ 2,\ \ldots, K \tag{1}$$

where y_j is the j-th input feature of the Softmax function, and z_j is the estimated probability distribution of an observation belonging to the j-th class.

2.3. Domain Adversarial Neural Network

Inspired by GAN, the domain adversarial neural network minimizes the distance between the source domain and the target domain through the adversarial training of the feature generation network G_f and the domain discrimination network G_d. This technology has been successfully applied to the machinery fault diagnosis. It adds a domain discriminator G_d on the basis of ordinary CNN network. In the training process, the domain discriminator is used to distinguish the features coming from the source domain or the target domain. The feature generation network expects to fool the domain discriminator as much as possible to maximize the loss of G_d, so that the domain discriminator cannot distinguish features coming from the source domain or the target domain. Through this adversarial training idea, the model can learn the domain invariant features of the source domain and the target domain. The overall loss function of the domain adversarial neural network is described as follows:

$$\begin{aligned} L(\theta_f, \theta_y, \theta_d) = & \frac{1}{n_S} \sum_{x_i \in X_S} L_y\left(G_y\left(G_f\left(x_i; \theta_f\right); \theta_y\right), y_i\right) \\ & - \frac{\lambda}{n_S + n_T} \sum_{x_i \in X_S \cup X_T} L_d\left(G_d\left(G_f\left(x_i; \theta_f\right); \theta_d\right), d_i\right) \end{aligned} \tag{2}$$

where, $\theta_f, \theta_y, \theta_d$ are the parameters of G_y G_f and G_d respectively; d_i is the domain label of x_i (0 or 1), corresponding to the source domain or target domain; L_y is the loss function of the classifier; L_d is the loss function of the domain discriminator; λ is a hyperparameter that weighs the two loss functions L_y and L_d.

The training process of this network is a min-max adversarial optimization problem. The key of this network is to extract

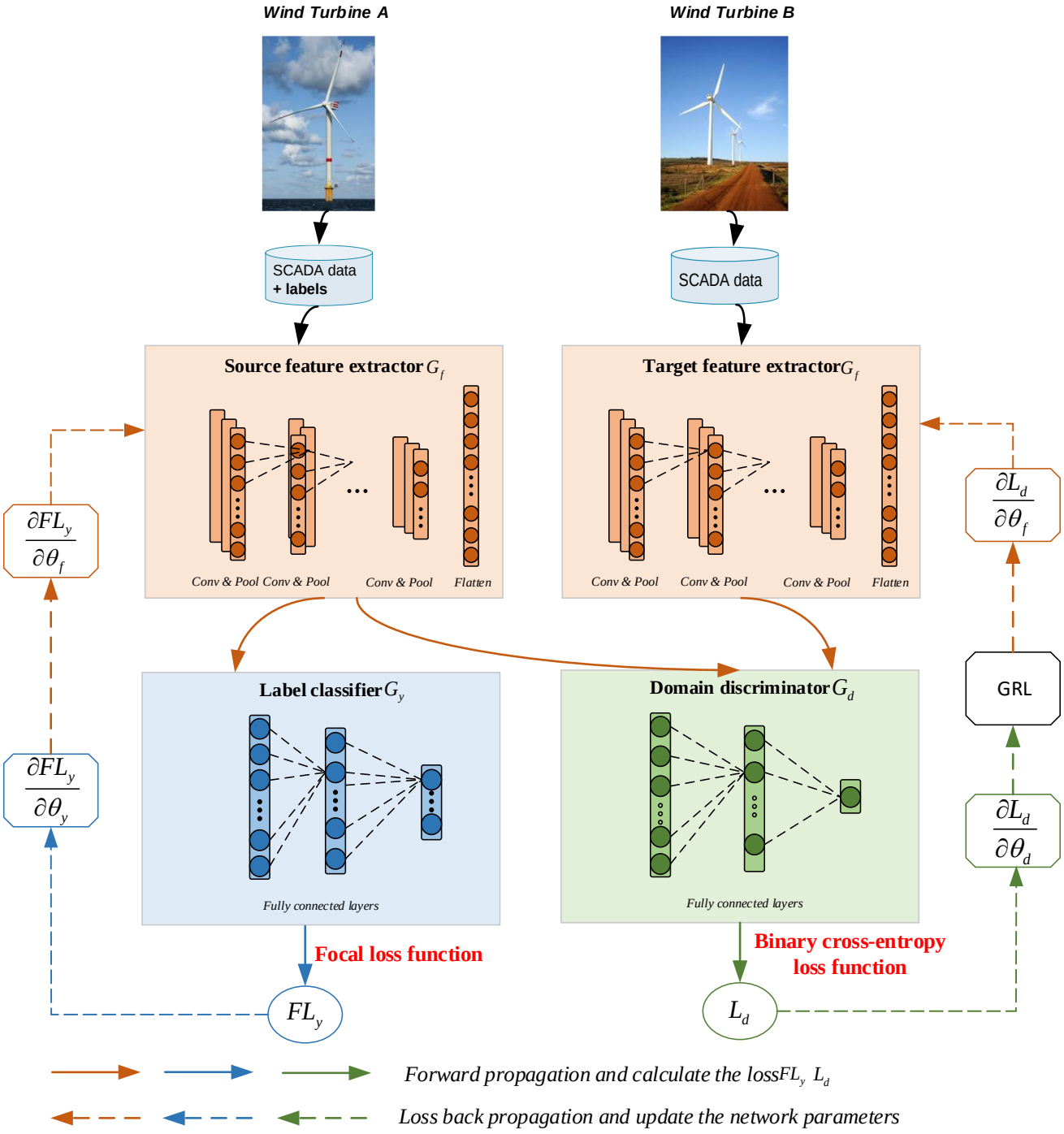


Figure 1. The architecture of the proposed WT-GAN model.

the discriminative and domain-invariant features. Specifically, the purpose of the network optimization is to minimize the loss of the classifier, so as to encourage the feature extractor to obtain discriminative deep features, while maximizing the loss of the domain discriminator to obtain domain-invariant features.

3. Proposed Method

As mentioned before, due to the difference in equipment parameters, electrical parameters and working environment, it is different for the feature distribution of SCADA data collected by different wind turbines. Therefore, the model trained on the dataset of one specific wind turbine performs

poorly on unknown wind turbines, easily resulting in wrong decisions. The domain adversarial network introduced above can effectively reduce the distribution difference of SCADA data from different wind turbines. Therefore, a GAN-based deep unsupervised transfer learning algorithm is proposed to proceed the health status prediction of wind turbines. In addition, the proportion of abnormal data in the collected SCADA data is tiny, so a serious data imbalance problem exists in the dataset. Therefore, the focal loss function is introduced to solve the problem of data imbalance, thereby increasing the network's attention on abnormal samples, and ultimately improving the knowledge transfer ability of the network model on different wind turbines. In this section, the proposed WT-GAN network and the optimization process of the model is described in detail.

3.1. WT-GAN Network

As shown in Figure 1, the WT-GAN mainly includes four modules, source feature extractor G_f , target feature extractor G_f , label classifier G_y , and domain discriminator G_d . The source feature extractor and the target feature extractor share the same network parameters.

3.1.1. Feature Extractor

As CNN has powerful feature extraction capabilities, the feature extractor network are constructed with multiple convolutional layers to extract the deep abstract features from the SCADA data in the source and target domains. The specific network parameters of G_f are shown in Figure 2. *Conv* refers to the convolutional layer. Regardless of the source domain data or the target domain data, they all need to be input into G_f to extract domain-invariant features. Specifically, G_f contains five convolution modules. In the first four convolution modules, the convolution kernel size is 3×3. The features after the convolution operation are fed to the ReLU activation function to enhance the nonlinear ability of the model. Then, the BN layer is used to improve the generalization ability of the model and speed up the convergence of the network to avoid the disappearance of the gradient. After BN layers, in order to reduce the feature dimension, the maximum pooling layer is then used in the second and fourth convolution module respectively. In the second convolution module, the pooling stride is 2×2, while it is 1×2 in the fourth convolution module. Therefore, in order to avoid losing important information when extracting deep features, the number of convolution kernels gradually increases from 16, 32 to 64, 128, 256 as the network deepens. In addition, it is worth noting that although CNN is selected as the feature extraction network in this paper, it is not limited to a specific network structure. The proposed network architecture can be easily extended to other deep models, such as recurrent networks.

3.1.2. Domain Discriminator

As shown in Figure 1, the features obtained from the feature extractor are input into the domain discriminator. The domain discriminator is mainly composed of three fully connected layers, as shown in Figure 2. The number of neurons in the three layers is set to 128, 64 and 2, respectively. The first two fully connected layers use Leaky ReLU, which is defined as:

$$lr(x) = \begin{cases} x & x \geq 0 \\ \xi x & x < 0 \end{cases} \tag{3}$$

It can be found that unlike ReLU, Leaky ReLU does not set values (less than zero) to zero, which retains the original information to a certain extent. In this work, the parameter ξ is set to 0.2. In the domain discriminator, this feature of Leaky ReLU makes it have better performance. In the binary classification task, the Sigmoid activation function makes the training more stable and easier to converge, so the Sigmoid function is used in the last layer of the domain discriminator. It is defined as:

$$S(x) = \frac{1}{1+e^{-x}} \tag{4}$$

In addition, a binary cross-entropy loss function is used. Therefore, the total loss output by the domain discriminator is calculated as follows:

$$\begin{aligned} L_d &= mean\{l_0, l_1, \cdots, l_N\} \\ l_n &= -(d_n \times log(z_n) + (1-d_n) \times log(1-z_n)) \end{aligned} \tag{5}$$

where N is the number of samples, z_n is the probability that the n-th sample belongs to a positive example, and d_n is the label of the n-th sample, which is 0 or 1, that is, it comes from the source domain or the target domain. The total loss output by the domain discriminator is the mean value of the loss of all samples.

3.1.3. Label Classifier and Focal Loss Function

As shown in Figure 1, the features obtained from the feature extractor are also input into the label classifier. The specific structure is shown in Figure 2. The label classifier uses two fully connected layers, and the number of neurons is 100 and K respectively. K is the total number of fault categories. The first fully connected layer maps the 1×256 features into 1×100 features with ReLU activation function, and the second fully connected layer maps the 1×100 features into 1×K features. For multi-classification tasks, the Softmax layer is often used as a classifier, which maps the output of the fully connected layer to the range of 0 to 1, and the sum of all feature values is 1. The commonly used loss function for classification tasks is the cross-entropy loss function, which is used to evaluate the error between the predicted probability distribution output by the Softmax layer and the true probability distribution. However, as mentioned above, the SCADA data of wind turbines has serious data imbalance problems. To solve it, the focal loss function is introduced to

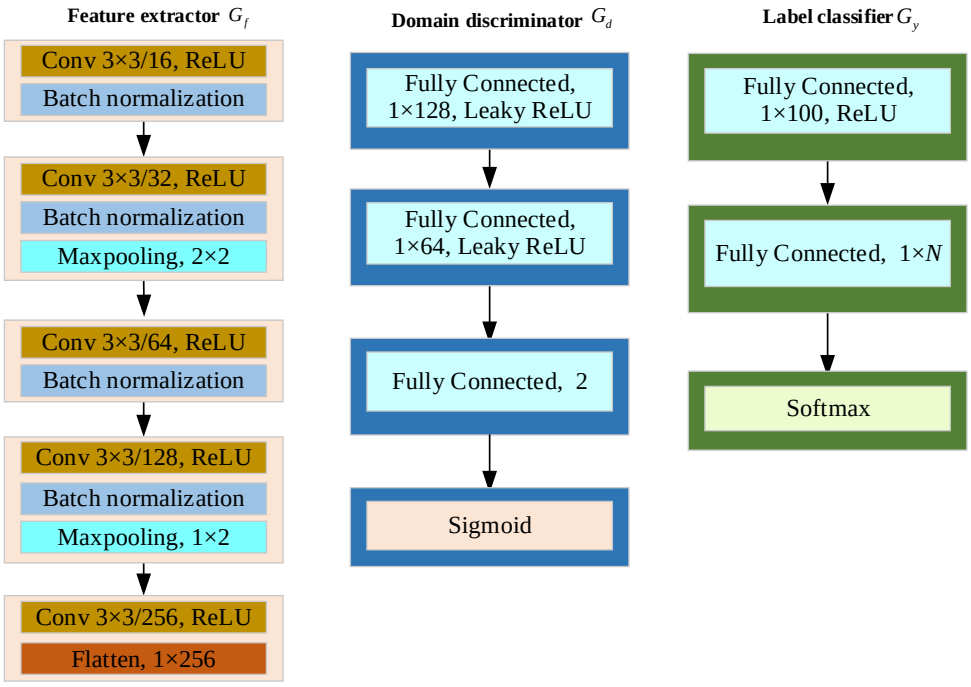


Figure 2. The detailed parameters of feature extractor, domain discriminator and label classifier.

the label classifier. Taking the binary classification task as an example, cross entropy loss and focal loss can be expressed as Eq. (6) and Eq. (7), respectively.

$$CE = \begin{cases} -\log y_i', & y=1 \\ -\log(1-y_i'), & y=0 \end{cases} \tag{6}$$

$$FL_y = \begin{cases} -\alpha(1-y')^{\gamma}\log y', & y=1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), & y=0 \end{cases} \tag{7}$$

where y' represents the value predicted by the model, and y represents the true label. It can be seen that focal loss function introduces two hyperparameters α and γ. These hyperparameters can make loss focus on the learning of fault samples, so that the model has excellent performance on unbalanced datasets.

3.2. Model Optimization

It can be seen from *Section 3.1* that, the loss function of WT-GAN consists of two parts, namely FL_y obtained by the label classifier and L_d obtained by the domain discriminator, as shown in Figure 1. In the network training process, two tasks are required to be finished. The first one is to achieve accurate classification of the source domain dataset, that is, to minimize FL_y. The second task is to try to confuse the source domain dataset and the target domain dataset. That is to achieve the maximization of L_d of the domain discriminator. The loss function of WT-GAN can be defined as:

$$L(\theta_f,\theta_y,\theta_d) = FL_y(\theta_f,\theta_y) - \lambda L_d(\theta_f,\theta_d) \tag{8}$$

Therefore, the optimal values of the parameters $\theta_f,\theta_y,\theta_d$ can be expressed as:

$$\begin{aligned} \left(\hat{\theta}_f,\hat{\theta}_y\right) &= \arg\min_{\theta_f,\theta_y} L\left(\theta_f,\theta_y,\hat{\theta}_d\right) \\ \hat{\theta}_d &= \arg\max_{\theta_d} L\left(\hat{\theta}_f,\hat{\theta}_y,\theta_d\right) \end{aligned} \tag{9}$$

When optimizing and updating the network parameters, the optimization method Adam (Adaptive Moment Estimation) is used (Kingma et al., 2014). The advantage of Adam is that each iterative learning rate has a certain range, which makes the parameters relatively stable. For example, the updating function of θ_f is defined as:

$$\theta_f = \theta_f - \mu\frac{\hat{m}_f}{\sqrt{\hat{v}_f}+\varepsilon} \tag{10}$$

where, μ is the learning rate, defined as 0.0001 in this article; $\varepsilon = 10^{-8}$ to prevent the divisor from becoming 0; $\hat{m}_f$ and $\hat{v}_f$ are defined as:

$$\begin{aligned} \hat{m}_f &= \frac{\beta_1 m + (1-\beta_1)g_f}{1-\beta_1} \\ \hat{v}_f &= \frac{\beta_2 v + (1-\beta_2)g_f^2}{1-\beta_2} \\ g_f &= \frac{\partial FL_y^i}{\partial\theta_f} - \lambda\frac{\partial L_d^i}{\partial\theta_f} \end{aligned} \tag{11}$$

Table 1. The data collection time of the two wind turbines and the percentage of icing data.

Wind Turbine	Date	Healthy	Ice	Ratio	Total
#15	Nov. 1st, 2015 - Jan. 1st, 2016	350255	23892	14.7:1	374147
#21	Nov. 1st, 2015 - Dec. 1st, 2015	168930	10683	15.8:1	179613

Similarly, the updating of θ_y, θ_d is similar to that of θ_f, replacing g_f with g_y or g_d, defined as:

$$g_y = \frac{\partial FL_y^i}{\partial \theta_y} \tag{12}$$

$$g_d = \frac{\partial L_d^i}{\partial \theta_d} \tag{13}$$

When updating the parameter θ_f, the gradient direction in the back propagation process is required to be automatically reversed. This benefits from the Gradient Reversal Layer (GRL), as shown in Figure 1. The GRL is located between the feature extractor and the domain discriminator.

In the backpropagation process, the gradient of L_d of the domain discriminator will be automatically inverted before being backpropagated to the feature extractor because of GRL, thus realizing adversarial loss. The label classifier and the domain discriminator compete with each other in the training process and finally achieve the balance between the label loss and the domain discriminator loss.

4. EXPERIMENTAL VALIDATIONS

In this section, the dataset and experimental settings are firstly elaborated. Then, the effectiveness of the proposed method is verified. Finally, the comparative experiments verify that the WT-GAN has excellent performance on unknown wind turbines.

4.1. Dataset Description

In order to verify the effectiveness of the proposed method, this paper uses a real wind turbine SCADA dataset. Table 1 shows the data collection time of the two wind turbines and the percentage of icing data. This dataset is mainly used for blade icing detection of a wind farm. Specifically, the dataset includes SCADA data for two wind turbines. In this study, the dataset of wind turbine #15 is used as the source domain, and the dataset of wind turbine #21 is used as the target domain. The dataset contains multiple physical characteristics and parameters of the operating state of the wind turbine, such as wind speed, motor temperature, power generation, etc. Each sample has a corresponding time record, and the label of the sample is given, icing or no icing. Table 2 shows the physical characteristics and the operating parameters collected by the SCADA system. If the blades of wind turbines freeze, it will affect the rotor speed, the power generation efficiency, and the blade balance of wind turbines as well. In order to make the network learn features as much as possible, all the monitoring information are fed into the network model. It can be found from Table 1 that the ratio of health data to icing data in wind turbine #15 is 14.7, and the ratio of wind turbine #21 is 15.8, so this dataset presents a serious data imbalance problem. In addition, due to the difference in mechanical parameters and working environment, the data distribution between the two wind turbines is different.

Table 2. Wind turbine parameters collected by the SCADA system.

No.	Parameter	No.	Parameter
1	Wind speed	2	Generator speed
3	Grid side active power	4	Wind direction
5	Mean of wind direction	6	Yaw position
7	Yaw speed	8	Pitch1 angle
9	Pitch2 angle	10	Pitch3 angle
11	Pitch1 speed	12	Pitch2 speed
13	Pitch3 speed	14	Pitch motor 1 temperature
15	Pitch motor 2 temperature	16	Pitch motor 3 temperature
17	X-direction acceleration	18	Y-direction acceleration
19	Environment temperature	20	Cabin temperature
21	Ng5 1 temperature	22	Ng5 2 temperature
23	Ng5 3 temperature	24	Ng5 1 charger DC current
25	Ng5 1 charger DC current	26	Ng5 1 charger DC current
27	Data group identification		

As shown in Table 2, the SCADA system will collect information of 27 parameters each time. In order to enable the network model to learn the changes of these parameters over time, a training sample is composed of 10 adjacent time monitoring data, and thus the sample dimension is 10×27. The models are all written with Python 3.6 and the deep learning framework Pytorch, and run on Ubuntu 16.04 system with GTX 2080 GPU. In this experiment, five metrics (*Accuracy*, *Precision*, *Recall*, *F*1, and *Score*) are used to comprehensively evaluate the network performance. Those can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Table 3. The Experimental results of WT-CNN and WT-GAN (Focal Loss).

	WT-CNN (test on source only)	WT-CNN (test on target only)	WT-GAN
Accuracy	0.7156 ± 0.0120	0.6763 ± 0.0377	**0.7776** ± 0.0643
Precision	0.7187 ± 0.0255	0.6869 ± 0.0393	**0.7762** ± 0.0459
Recall	0.8915 ± 0.0945	0.8914 ± 0.1359	**0.8966** ± 0.0655
*F*1	0.7918 ± 0.0251	0.7674 ± 0.0481	**0.8315** ± 0.0506
Score	0.6638 ± 0.0148	0.6129 ± 0.0448	**0.7425** ± 0.0669

Table 4. The experimental results of WT-CNN and WT-GAN (Cross-Entropy Loss).

	Baseline (test on source only)	Baseline (test on target only)	WT-GAN
Accuracy	0.6845 ± 0.0191	0.6989 ± 0.0686	**0.7222** ± 0.0460
Precision	0.7287 ± 0.0263	0.7234 ± 0.0505	**0.7426** ± 0.0290
Recall	0.7849 ± 0.1088	0.8413 ± 0.1359	**0.8435** ± 0.1222
*F*1	0.7501 ± 0.0369	0.7704 ± 0.0649	**0.7844** ± 0.0548
Score	0.6549 ± 0.0129	0.6569 ± 0.0771	**0.6865** ± 0.0342

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

$$Score = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{18}$$

where *TP*, *TN*, *FP*, *FN* refer respectively to the number of true positive samples, true negative samples, false positive samples and false positive samples. If the proportions of each category in the dataset are similar, the *Accuracy* can evaluate whether the model performs well or not. However, for the wind turbine dataset, the health data and the fault data are severely unbalanced. Therefore, the *Accuracy* metric cannot accurately evaluate the performance of the model. Therefore, the *Precision*, the *Recall*, the *F*1, and the *Score* are introduced to comprehensively evaluate the model performance as much as possible. In the *Score* evaluation index, the weights of the health category and the fault categories are the same and have nothing to do with the number of samples, this metric is more reasonable for performance evaluation of sample imbalance tasks.

4.2. Experimental Results

4.2.1. Performance of Domain Transfer

In this section, in order to verify the effectiveness of the proposed transfer learning algorithm, in addition to the method proposed in this paper, a WT-CNN model is also constructed. Compared with the WT-GAN, the only difference is that the WT-CNN has no domain transfer capability. The WT-CNN is mainly composed of the feature extractor and the label classifier presented in Figure 1. The experimental settings and the training parameters are exactly the same. 75% of the samples in the SCADA dataset of wind turbines #15 and #21 is used for training, and the remaining 25% for testing. Therefore, for wind turbines #15 and #21, their training samples are 20641, and their test samples are 6880. In this experiment, the batch size is 128. In order to avoid the randomness of the experimental results, each experiment is repeated for five times, and its average and standard deviation are calculated. The experimental results are shown in Table 3.

The second column represents the results obtained by the WT-CNN trained on the source domain and then tested on the source domain. The third column represents the results obtained by WT-CNN trained on the source domain and then tested on the target domain. The fourth column represents the result of the WT-GAN trained on the source domain and then tested on the target domain. It can be found that the performance of the WT-GAN with domain transfer capability is significantly better than the WT-CNN, and WT-GAN obtains the best results in all five evaluation indicators. Specifically, compared to WT-CNN, WT-GAN improves the *Accuracy*, the *Precision*, the *Recall*, the *F*1 and the *Score* by approximately 6.20%, 5.75%, 0.51%, 3.97%, and 7.87%, respectively. Especially for the *Score* indicator, the WT-GAN increases it by 7.87%. This shows that the proposed deep unsupervised transfer learning method can learn the domain invariant features of different wind turbines, thereby reducing the domain shift between wind turbines. In this way,

the model trained on a specific wind turbines can be applied to the fault diagnosis of the unseen wind turbines, thereby greatly improving the practical application value of the deep learning model.

4.2.2. Effectiveness of Focal Loss Function

In order to prove the effectiveness of the focal loss function in this section, a set of experiments are also proceeded using WT-CNN and the WT-GAN with cross-entropy loss function. The experimental settings and the training parameters are exactly the same as in the previous section. Similarly, five repeated experiments are conducted. The experimental results are shown in Table 4.

Obviously, even if the cross-entropy loss function is used in the WT-CNN and the WT-GAN, the performance of the WT-GAN is still significantly better than the WT-CNN. This again confirms that the WT-GAN has a very good knowledge transfer ability, which is consistent with the experimental conclusions of the previous section.

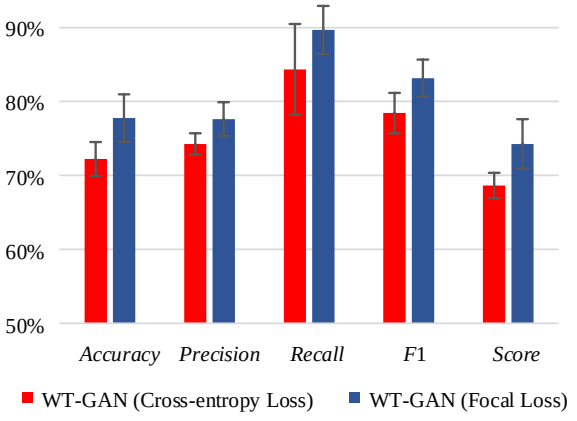


Figure 3. Experimental results of WT-GAN.

Figure 3 shows the result when the focal loss function and the cross-entropy loss function are respectively used by the WT-GAN. It can be seen that, the WT-GAN with the focal loss function performs significantly better than that of the WT-GAN with the cross-entropy loss function. All the evaluation metrics of the WT-GAN with focal loss function have higher values than that using the cross-entropy loss function. Specifically, compared with the WT-GAN (cross-entropy Loss), the five evaluation indicators of the WT-GAN (Focal Loss) are 5.54%, 3.36%, 5.31%, 4.71%, and 5.60% higher, respectively. The performance of WT-GAN (Focal Loss) is competitive because the focal loss function pays more attention to the feature learning of the fault samples. Focal loss function can effectively improve the ability of WT-GAN when dealing with data imbalance task, so that WT-GAN has excellent wind turbine fault detection performance.

5. CONCLUSION

Considering the problem of lack of labels and data imbalance in the SCADA data of wind turbines, this paper proposes a deep unsupervised transfer learning network with the focal loss function. The proposed model uses the idea of adversarial training to reduce the domain shift among different wind turbines. This enables the network model trained on one wind turbine to be well transferred to the fault diagnosis of other unknown wind turbines. In addition, the focal loss function makes the network model to pay more attention on the feature learning of fault samples, which significantly improves the model's performance with unbalanced dataset. Therefore, the proposed method makes it easy to be applied in practical industrial applications. In order to verify the effectiveness of the proposed model, the proposed method is applied to the case of blade icing detection of a wind farm. The experimental results confirms the model transfer ability and the ability to solve the data imbalance of the proposed method.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of the China Scholarship Council, the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme and the Research Foundation–Flanders (FWO) under the ROBUSTIFY research grant no. S006119N.

NOMENCLATURE

D_S	source domain
D_T	target domain
n_S	total sample number in the source domain
n_T	total sample number in the target domain
X_S	sample in the source domain
X_T	sample in the target domain
K	total fault types of the wind turbine
Y_S	sample label in the source domain
$P_S(X_S)$	distribution of the source domain sample
$P_T(X_T)$	distribution of the target domain sample
M_{i-1}	input of the *i*-th convolution layer
M_i^{conv}	features obtained by the convolution operation
W_i	weight of the convolution kernel
B_i	bias
$*$	convolution operation
y_j	the *j*-th input feature of the Softmax function
z_j	estimated probability distribution of an observation belonging to the *j*-th class
G_f	feature generation network

G_y	label classifier
G_d	domain discriminator
θ_f	parameter of G_f
θ_y	parameter of G_y
θ_d	parameter of G_d
d_i	domain label of domain discriminator
L_y	loss function of the classifier
L_d	loss function of the domain discriminator
λ	a hyperparameter that weighs the two loss functions L_y and L_d
Conv	convolutional layer
ξ	hyperparameter in Leaky ReLU
N	number of samples
z_n	probability that the *n*-th sample is a positive example
d_n	domain label of the *n*-th sample, which is 0 or 1
y'	predicted fault label of label classifier
y	real fault label
α	a hyperparameter in Focal Loss
γ	another hyperparameter in Focal Loss
FL_y	focal loss function of the classifier
μ	learning rate

REFERENCES

Márquez, F. P. G., Pérez, J. M. P., Marugán, A. P., & Papaelias, M. (2016). Identification of critical components of wind turbines using FTA over the time. Renewable Energy, vol. 87, pp. 869-883. doi: 10.1016/j.renene.2015.09.038

Davis, N.N., Byrkjedal, Ø., Hahmann, A.N., Clausen, N., & Mark, Ž. (2016). Ice detection on wind turbines using the observed power curve. Wind Energy, vol. 19, pp. 999-1010. doi: 10.1002/we.1878

Muñoz, C., Márquez, F., & Tomás, J. (2016). Ice detection using thermal infrared radiometry on wind turbine blades. Measurement, vol. 93, pp. 157-163. doi: 10.1016/j.measurement.2016.06.064

Yang, X., Zhang, Y., Lv, W., & Wang, D. (2021). Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier. Renewable Energy, vol. 163, pp. 386-397. doi: 10.1016/j.renene.2020.08.125

Chen, L., Xu, G., Zhang, Q., & Zhang, X. (2019). Learning deep representation of imbalanced SCADA data for fault detection of wind turbines. Measurement, vol. 139, pp. 370-379. doi: 10.1016/j.measurement.2019.03.029

Peng, D., Wang, H., Liu, Z., Zhang, W., Zuo, M. J., & Chen, J. (2020). Multibranch and multiscale CNN for fault diagnosis of wheelset bearings under strong noise and variable load condition. IEEE Transactions on Industrial Informatics, vol. 16(7), pp. 4949-4960. doi: 10.1109/TII.2020.2967557

Liu, C., & Gryllias, K. (2020). A semi-supervised support vector data description-based fault detection method for rolling element bearings based on cyclic spectral analysis. Mechanical Systems and Signal Processing, 140, 106682. doi: 10.1016/j.ymssp.2020.106682

Peng, D., Liu, C., Desmet, W., & Gryllias, K. (2021). An improved 2DCNN with focal loss function for blade icing detection of wind turbines under imbalanced SCADA data. *Proceedings of the ASME 2021 3rd International Offshore Wind Technical Conference*, Virtual, Online. doi: 10.1115/IOWTC2021-3527

Chen, L., Xu, G., Liang, L., Zhang, Q., & Zhang, S. (2018). Learning deep representation for blades icing fault detection of wind turbines. *In 2018 IEEE International Conference on Prognostics and Health Management (ICPHM)* (pp. 1-8), June 11-13, Seattle, WA, USA. doi: 10.1109/ICPHM.2018.8 448394

Pang, Y., He, Q., Jiang, G., & Xie, P. (2020). Spatio-temporal fusion neural network for multi-class fault diagnosis of wind turbines based on SCADA data. Renewable Energy, vol. 161, pp. 510-524. doi: 10.1016/j.renene. 2020.06.154

Wen, L., Gao, L., & Li, X. (2017). A new deep transfer learning based on sparse auto-encoder for fault diagnosis. IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49(1), pp. 136-144. doi: 10.1109/TSMC.2017.2754287

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *In International Conference on Machine Learning* (pp. 1180-1189), July 6-11, Lille, France.

Liu, Q., Ma, G., & Cheng, C. (2020). Data fusion generative adversarial network for multi-class imbalanced fault diagnosis of rotating machinery. IEEE Access, vol. 8, pp. 70111-70124. doi: 109/ACCESS.2020.2986356

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2018). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42(2), pp. 318-327. doi: 10.1109/tpami. 2018.2858826

Chen, Z., Gryllias, K., & Li, W. (2019). Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network. IEEE Transactions on Industrial Informatics, vol. 16(1), pp. 339-349. doi: 10.1109/TII.2019.2917233

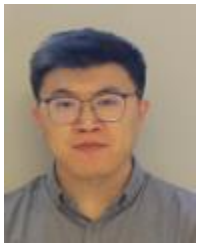
Wang, H., Liu, Z., Peng, D., & Qin, Y. (2019). Understanding and learning discriminant features based on multi attention 1DCNN for wheelset bearing fault diagnosis. IEEE Transactions on Industrial Informatics, vol. 16(9), pp. 5735-5745. doi: 10.1109/TII.2019.2955540

Nair, V., & Hinton, G. (2010). Rectified linear units improve restricted Boltzmann machines, *in Proc. International Conference on International Conference on Machine Learning* (pp. 807-814), June 21-24, Haifa, Israel.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

BIOGRAPHIES

Dandan Peng was born in Shaanxi, China, in 1992. She received the B.S. and M.S. degrees in mechanical engineering from the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2016 and 2019, respectively. She is currently working toward the Ph.D. degree in mechanical engineering from Catholic University of Leuven, Leuven, Belgium. Her research interests include Hilbert–Huang transform, convolutional neural network, machinery condition monitoring, and fault diagnosis.



Chenyu Liu received the B.S. and M.Sc degree in aircraft manufacturing engineering and solid mechanics both from Northwestern Polytechnical University, China. He joined the noise and Vibration Research Group in the Department of Mechanical Engineering, KU Leuven, Belgium as a PhD researcher in 2017. His research interests include data driven based condition monitoring, rotating machinery prognostics, and machine learning and deep learning application.

Wim Desmet holds a full professor position on vibro-acoustics of machines and transportation systems at the Department of Mechanical Engineering of KU Leuven, Belgium. He is also the head of noise & vibration research group. His research interests lie in the fields of numerical and experimental vibro-acoustics, uncertainty modelling of dynamic systems, aeroacoustics, active noise and vibration control, noise control engineering, multibody dynamics, dynamics of lightweight materials and systems, vehicle mechatronics and virtual prototyping.

Konstantinos Gryllias holds a 5 years engineering diploma degree and a PhD degree in Mechanical Engineering from National Technical University of Athens, Greece. He holds an associate professor position on vibro−acoustics of machines and transportation systems at the Department of Mechanical Engineering of KU Leuven, Belgium. He is also the manager of the University Core Lab Dynamics in Mechanical & Mechatronic Systems DMMS-M of Flanders Make, Belgium. His research interests lie in the fields of condition monitoring, signal processing, prognostics and health management of mech. & mechatronic systems.