An Advanced Diagnostic Model for Gearbox Degradation Prediction Under Various Operating Conditions and Degradation Levels

Hanqi Su[1*], Jay Lee[2]

[1,2] *Center for Industrial Artificial Intelligence, Department of Mechanical Engineering,*
University of Maryland, College Park, MD, 20742, USA
hanqisu@umd.edu, leejay@umd.edu

ABSTRACT

This study introduces a novel three-stage diagnostic methodology aimed at enhancing the prediction and classification of gearbox degradation under various operating conditions and multiple degradation levels, addressing the complexities encountered in real-world industrial settings. Leveraging the latest advancements in data-driven approaches, from similarity-based methods to residual-based deep convolutional neural networks (CNNs) and pseudo-labeling techniques, our approach systematically classifies data into known, unknown, and undetermined categories, predicts known degradation levels, and refines classification models with augmented pseudo-label data. The efficacy of our methodology is demonstrated through its remarkable performance using the data from the PHM North America 2023 Conference Data Challenge. It achieves scores of 600 / 800 on the testing data and 574 / 813 on the validation data, significantly surpassing the first-place scores of 463.5 and 472 in the competition, respectively, setting a new benchmark in the field of gear fault diagnosis.

1. INTRODUCTION

Over the years, the gear fault diagnosis domain has witnessed the evolution of numerous data-driven approaches aimed at identifying and diagnosing gear faults and degradation to ensure the reliability and efficiency of mechanical systems. Existing research has shown variant machine learning (ML) and artificial intelligence approaches for gear fault diagnosis (Kumar, Gandhi, Zhou, Kumar, & Xiang, 2020; Zhu et al., 2023; Su & Lee, 2024), including convolutional neural networks (CNN) (Zhao, Kang, Tang, & Pecht, 2017; Kreuzer & Kellermann, 2023), recurrent neural networks (RNN) (Tao, Wang, Sánchez, Yang, & Bai, 2019; Durbhaka et al., 2021), autoencoders (AE) (Saufi, Ahmad, Leong, & Lim, 2020; Z. He et al., 2020), deep belief networks (DBN) (X. Wang, Qin, & Zhang, 2018; Li, Li, He, & Qu, 2019), etc. However, in real-world industrial settings, analyzing and diagnosing gearbox degradation may become more complex. Gearboxes may operate under a variety of working conditions, and sometimes, some states of gear health are not known in advance. This uncertainty adds a significant layer of difficulty to gear fault diagnosis. Meanwhile, the small size of the data sets also presents challenges for ML model deployment. Consequently, there is a growing need to explore and develop innovative solutions that reduce reliance on specialized expertise and enable the creation of more versatile, automated systems for gear fault diagnosis. These advancements hold the promise of making gear fault diagnosis more accessible and efficient, paving the way for broader applications and enhanced operational reliability.

To address these limitations, we propose a novel three-stage diagnostic approach for predicting gearbox degradation. The initial stage introduces a similarity-based model designed to classify data into known, unknown, and undetermined categories. Subsequently, the second stage employs a residual-based CNN regression model, focused on the prediction of known degradation labels. In the final stage, we transition the regression model to a classification model. We incorporate pseudo-labeling techniques to assign pseudo-labels to testing data. The data, now augmented with pseudo labels, is then used to refine the classification model further. This innovative approach enhances the model's robustness and its generalization capabilities, offering a comprehensive solution to the challenges of gearbox degradation prediction.

The rest of this article is organized as follows: Section 2 introduces the competition and dataset, reviews the relevant approaches, and data preprocessing module. Section 3 provides a comprehensive analysis of three-stage ML model construction. Section 4 reports and discusses the performances of ML models and summarizes the limitations of this study. Section 5 concludes this paper by highlighting its findings and contributions.

[*] Corresponding author: Hanqi Su (hanqisu@umd.edu)

2. BACKGROUND

2.1. Literature review

Existing work in the field has increasingly pivoted towards data-driven ML models that can handle data from complex systems under varying conditions (Kumar et al., 2020; Zhu et al., 2023; Su & Lee, 2024). Research has explored various methodologies, from statistical models to advanced deep learning techniques, for fault diagnosis and prognostics. For similarity-based approaches, (Bettahar, Rahmoune, Benazzouz, & Merainani, 2020) combined the wavelet transform, Hilbert transform, and cosine similarity metric to extract useful features for gear fault diagnosis. (Feng, Ni, Beer, Du, & Li, 2022) proposed a methodology based on similarity-based status characterization to thoroughly represent the degradation behaviors of gear systems.

Except for similarity-based approaches, many deep learning approaches are deployed to address gear fault diagnosis. (X. Wang et al., 2018) trained a DBN for the planetary gearbox fault diagnosis utilizing time domain features extracted from the vibration signal processed by the optimized Morlet wavelet and frequency domain features extracted from the pulse signal. To address the challenge of limited samples in gearbox diagnosis, (Saufi et al., 2020) proposed a powerful deep learning model based on a stacked sparse autoencoder (SSAE) using time-frequency images. To find the optimal hyperparameters, the Particle Swarm Optimization (PSO) algorithm was applied. (Mohamad, Abbasi, Kim, & Nataraj, 2021) developed an innovative deep learning framework that integrates a CNN module with a long short-term memory (LSTM) network module. In this framework, the CNN module extracts informative features, while the LSTM module excels in time series modeling. The framework is applied to fault classification of gearbox data, thereby enhancing condition-based maintenance (CBM). Moreover, (Chu, 2023) introduced a novel three-stage diagnostic approach, leveraging Domain2Vec structure (utilizing EfficientNet-B0 for feature extraction) for in-set health state classification, then transitioning the model to regression for out-of-set prediction, and employing KNN for error correction to boost performance. (Liu, 2023) used tree-based gradient boosting and neural networks to build the classification models for known labels and used the nearest neighbor clustering method to interpolate and extrapolate to unseen labels.(Vaerenberg et al., 2023) introduced a diagnostic approach employing a CNN optimized with an ordinal loss criterion. This method leverages the power spectral density of three-channel vibration signals to assess the severity of pitting faults in gearboxes accurately.

Moreover, pseudo-label-based methods have demonstrated impressive results across various industrial applications. The pseudo-label technique, as proposed by (D.-H. Lee et al., 2013), is designed by assigning pseudo-labels to unlabeled data and incorporating pseudo-labeled data into the training sets, which can improve ML model accuracy and generalization capabilities. Given the prevalence of unlabeled data in industrial settings compared to labeled data, the integration of information from unlabeled sources becomes crucial. (Song, Li, Jia, & Qiu, 2019) proposed a weighted pseudo-label-based retraining domain adaptation network to boost the model generalization performance. (Oh et al., 2023) employed a pseudo-label-based deep learning approach to improve fault diagnosis in hydrostatic rock drills.

2.2. Competition and dataset description

The data utilized in our study is from the PHM (Prognostics and Health Management) North America 2023 Conference Data Challenge[1]. This Challenge represents a cutting-edge intersection of ML and mechanical system health monitoring, focusing on gearbox degradation under diverse operational conditions and unseen degradation levels. The dataset comprises training, testing, and validation subsets. It includes 2,016 samples in the training set, 800 in the testing set, and 813 in the validation set. Each sample data is the time-series data recorded by a tri-axial accelerometer, capturing vibration signals at a sampling rate of 20,480 Hz. Degradation severity increases in levels from 0 to 10. The training dataset for the competition comprises 7 classes, including a healthy state (degradation level 0) and six distinct degradation levels (levels 1, 2, 3, 4, 6, 8). Notably, degradation levels 5, 7, 9, and 10 are excluded from the training dataset. There are 78 operating conditions spread across these 7 health levels, with an average of 3.69 repetitions for each operating condition of fault level included in the training dataset. In contrast, the testing and validation datasets cover whole degradation levels. Additionally, some operating conditions that are absent in the training dataset appear in both the testing and validation datasets. This exclusion is designed to replicate real-world scenarios where complete data across all operating conditions and degradation levels might not be available, thus challenging participants to develop models that can effectively handle partial or incomplete data.

2.3. Data preprocessing

The preprocessing phase involves a systematic procedure that includes signal segmentation, Fast Fourier Transform (FFT), and data visualization, shown in Figure 1. Initially, each time series data is divided into 22 segments, each lasting 1 second. As the data is measured on the output shaft, which operates at 5/9 the speed of the input shaft, for the lowest rotational speed condition (100 rpm), the input shaft speed equates to 4/3 Hz, while the output shaft speed is approximately 0.92 Hz. This setup ensures each segment captures at least one full period. Moreover, given the high sampling rate, segmenting the

[1] https://data.phmsociety.org/phm2023-conference-data-challenge/

original sample into smaller parts can effectively shorten the data length, increase sample size, capture local patterns, and make computational processes more efficient, finally benefiting deep learning model training. Following segmentation, the time domain data is converted into the frequency domain by using FFT, facilitating a comprehensive analysis. Through data visualization, both the time series' configuration and its frequency domain characteristics can be promptly assessed, laying a solid foundation for subsequent analytical steps.

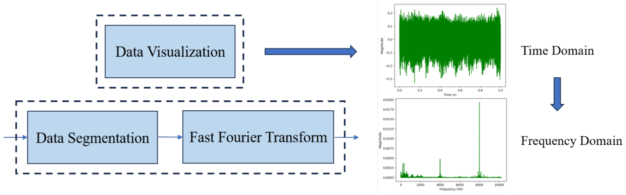


Figure 1. Process of data preprocessing.

3. METHODOLOGY

In this section, a novel three-stage diagnostic method is introduced. It involves a similarity-based model, a residual-based CNN regression model, and a refined CNN classification model using augmented pseudo-label data.

3.1. Stage 1: a similarity-based model for known, unknown, and undetermined class classification

The dataset utilized for testing encompasses data across eleven degradation levels, comprising a healthy status and ten levels of degradation. Notably, data from certain degradation levels and operational conditions are excluded from the training datasets. Initially, we devise a similarity-based model aimed at precisely classifying the test data into three categories: known, unknown, and undetermined. The categories are defined as follows: the "known" category includes degradation levels 0, 1, 2, 3, 4, 6, and 8, which are represented in the training dataset. The "unknown" category encompasses levels 5, 7, 9, and 10, which are not present in the training dataset. The "undetermined" category is applied when the similarity-based model cannot conclusively determine whether the data belongs to the known or unknown categories. It consists of 4 steps.

Step 1: Categorize the training and testing set. For a time series of length T with C distinct variables, each data in the frequency domain is represented as $X \in \mathbb{R}^{[\frac{T}{2}]\times C}$. Consequently, the dimension of the training set (X_{Train}) is $\mathbb{R}^{N_{train}\times[\frac{T}{2}]\times C}$, while the dimension of the testing set (X_{Test}) is $\mathbb{R}^{N_{test}\times[\frac{T}{2}]\times C}$.

For the training set and testing set, we segregate the frequency domain data based on varied operating conditions (19 different rotational speeds and 5 different torque) as follows:

$$\{X^{s,t}_{Train}\} = Categorize(X_{Train}) \tag{1}$$

$$\{X^{s,t}_{Test}\} = Categorize(X_{Test}) \tag{2}$$

where s denotes a specific rotational speed $\in \{100, 200, ..., 3600\}$ and t denotes a specific torque $\in \{100, 200, ..., 500\}$.

For each category defined by a specific combination of rotational speed s and torque t in the training dataset, the subset $X^{s,t}_{Train}$ is represented as $\mathbb{R}^{N^{s,t}_{train}\times[\frac{T}{2}]\times C}$, where $N^{s,t}_{train}$ is the number of samples in that category. The arrangement of rows in this matrix is based on the degradation levels, ascending from least to most severe. Similarly, the corresponding subset in the testing dataset, $X^{s,t}_{Test}$ is represented as $\mathbb{R}^{N^{s,t}_{test}\times[\frac{T}{2}]\times C}$

Step 2: Compute the training set similarity matrix. After categorizing the training set, a similarity matrix defined as $Sim\_X^{s,t}_{Train}$ is calculated for each category. We calculate the Euclidean distance between two samples to quantify the similarity. The element $Sim\_X^{s,t}_{Train}(i,j)$ is calculated as follows:

$$Sim\_X^{s,t}_{Train}(i,j) = \sqrt{(X^{s,t}_{Train}[i,:] - X^{s,t}_{Train}[j,:])^2} \tag{3}$$

A lower Euclidean distance value suggests a higher degree of resemblance between the paired samples. Typically, samples having the same degradation levels exhibit smaller Euclidean distance values compared to those with differing degradation levels. For samples classified under the same degradation levels, we compute the mean ($Mean^{s,t,k}_{Train}$) and standard deviation ($Std^{s,t,k}_{Train}$) of their similarity values, where k represents specific degradation levels, such as 0, 1, 2, 3, 4, 6, 8. For example, if the rows from a to row b in $X^{s,t}_{Train}$ correspond to the degradation level k, then the mean and standard deviation for this specific degradation level within the training set similarity matrix are computed as follows:

$$Mean^{s,t,k}_{Train} = \frac{1}{(a-1)(b-1)} \sum_{i=a}^{b} \sum_{j=a,j\neq i}^{b} (Sim\_X^{s,t}_{Train}(i,j)) \tag{4}$$

$$Std^{s,t,k}_{Train} = \sqrt{\frac{\sum_{i=a}^{b} \sum_{j=a,j\neq i}^{b} (Sim\_X^{s,t}_{Train}(i,j) - Mean^{s,t,k}_{Train})^2}{(a-1)(b-1)}} \tag{5}$$

These statistical measures are later utilized to assess whether the testing data correspond to the same labels as the training data.

Step 3: Compute the testing set similarity matrix. Similar to the approach in step 2, the similarity matrix for the testing set, $Sim\_X^{s,t}_{Test}$, is derived by computing the Euclidean distance between the rows of the training set ($X^{s,t}_{Train}$) and rows of the testing set ($X^{s,t}_{Test}$). The element $Sim\_X^{s,t}_{Test}(i,j)$ is calculated as follows:

$$Sim\_X^{s,t}_{Test}(i,j) = \sqrt{(X^{s,t}_{Train}[i,:] - X^{s,t}_{Test}[j,:])^2} \tag{6}$$

Calculating the similarity matrix for the testing set allows for a direct comparison with the training set, enabling the identification of similarity patterns and deviations between the two datasets.

For each column j in Sim_X_{Test}, the similarity value is calculated using all the rows from the training set and the specific row j from the testing set. Subsequently, the mean ($Mean^{s,t,k}_{Test}[j]$) for each degradation level k $\in \{0, 1, 2, 3, 4, 6, 8\}$ is computed. The degradation level $k_{min}[j]$, which has the lowest mean value, is then identified. The rows in the $Sim\_X_{Test}[:, j]$ that correspond to this degradation level k_{min} are used as a reference to classify the test data in column j. Within these reference rows, $max\_sim[j]$ and $min\_sim[j]$ denote the highest and lowest similarity values, respectively, which further assist in the classification process.

Step 4: Classify testing data into known, unknown, and undetermined classes. The classification process utilizes the previously calculated $Sim\_X^{s,t}_{Train}$ and $Sim\_X^{s,t}_{Test}$ for each set of operational conditions. For each operational condition, according to the degradation levels observed in the training data, the mean ($Mean^{s,t,k}_{Train}$) and standard deviation ($Std^{s,t,k}_{Train}$) is calculated from the training set similarity matrix ($Sim\_X^{s,t}_{Train}$) in step 2 and degradation level k_{min}, $max\_sim[j]$ and $min\_sim[j]$ are obtained in step 3. Therefore, a detailed representation of the similarity predictions is designated as $Sim\_Pred \in \mathbb{R}^{N^{s,t}_{test} \times 4}$. This matrix meticulously captures the relationship between testing data points and their most proximate degradation labels, alongside their categorization into known, unknown, and undetermined classes. The process for classifying the testing data is shown in Algorithm 1.

The decision to use thresholds of 3 × $Std^{s,t,k}_{Train}$ and 10 × $Std^{s,t,k}_{Train}$ comes from the assumption that the dataset follows a normal distribution. Under this assumption, about 99.7% of the data falls within three standard deviations from the mean. Consequently, setting a threshold at $Mean^{s,t,k}_{Train}$ + 10 × $Std^{s,t,k}_{Train}$ helps in identifying extreme outliers, as this threshold significantly exceeds the typical range of data variation.

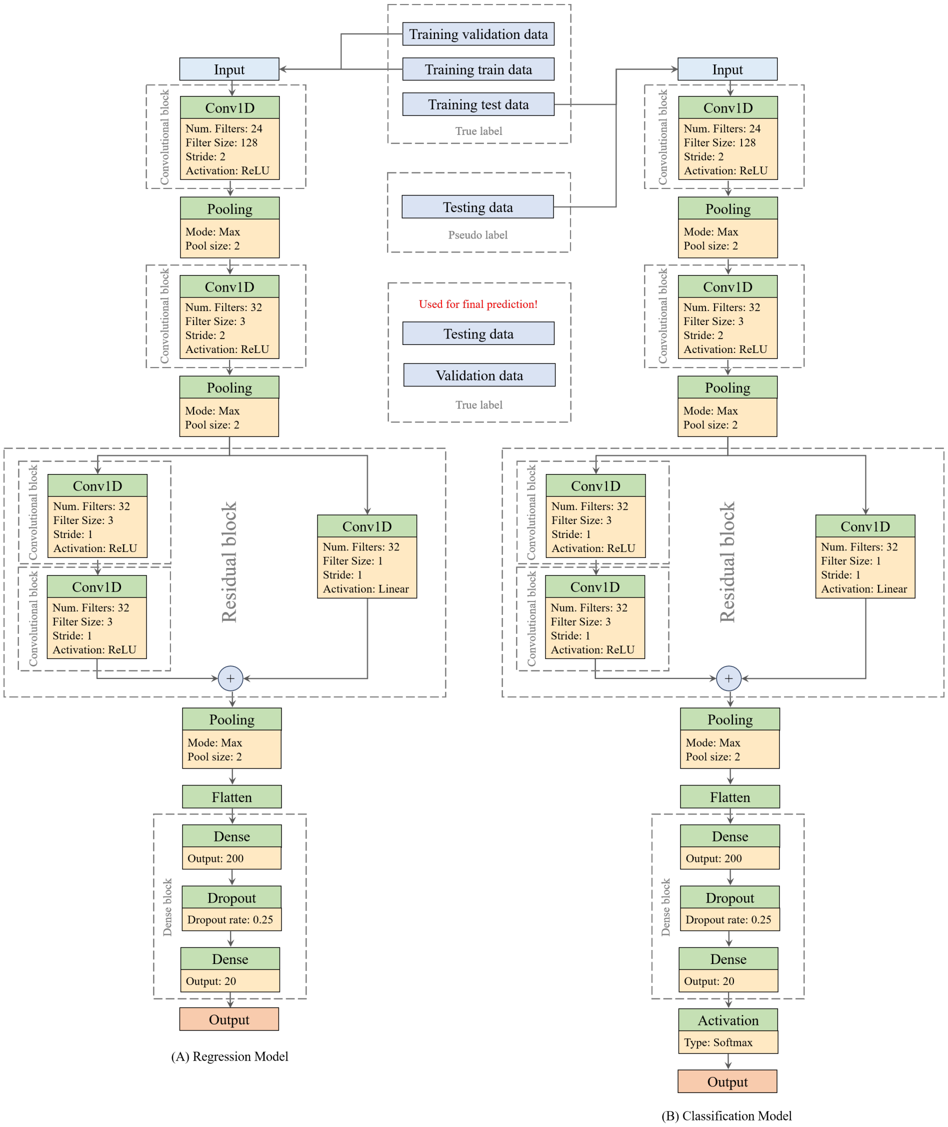
3.2. Stage 2: a residual-based CNN regression model for known degradation level prediction

In this stage, we propose a residual-based Convolutional Neural Network (CNN) model, specifically designed for the prediction of gearbox degradation levels. Our model draws inspiration from the renowned ResNet architecture—a deep learning framework known for its efficacy in tasks ranging from image recognition (K. He, Zhang, Ren, & Sun, 2016) to speech recognition (Vydana & Vuppala, 2017), as well as applications in condition monitoring and PHM (Duan, Shi, Zhou, Xuan, & Wang, 2021; Zhou et al., 2022). The original training data is divided into three subsets: ”Training Train Data,” ”Training Validation Data,” and ”Training Test Data,” following a 3:1:1 ratio. In this stage, the regression model is trained on the ”Training Train Data,” validated on the ”Training Validation Data,” and finally tested using the ”Training Test Data.”

Algorithm 1 Classify testing data into known, unknown, and undetermined classes for speed rotational speed s and torque t

```
Get $k_{min}$, $max\_sim$ and $min\_sim$ from Step 3.
Get $Mean^{s,t,k}_{Train}$ and $Std^{s,t,k}_{Train}$ from Step 2.
Initialize $Sim\_pred \in \mathbb{R}^{N^{s,t}_{test} \times 4}$ with value 0.
$j \leftarrow 0$
while $j \neq N^{s,t}_{test}$ do
    if $max\_sim[j] \leq Mean^{s,t,k_{min}[j]}_{Train} + 3 \times Std^{s,t,k_{min}[j]}_{Train}$ then
        $Sim\_pred[j, 0] = 1$ // known class
    else if $min\_sim[j] \geq Mean^{s,t,k_{min}[j]}_{Train} + 10 \times Std^{s,t,k_{min}[j]}_{Train}$ then
        $Sim\_pred[j, 1] = 1$ // unknown class
    else
        $Sim\_pred[j, 2] = 1$ // undetermined class
    end if
    $Sim\_pred[j, 3] = k_{min}[j]$ // the most similar degradation level
    $j \leftarrow j + 1$
end while
```

The regression model is tailored to address the challenge of predicting the levels of gearbox degradation. The architecture of the proposed model is illustrated in Figure 2 (A). It incorporates a batch size of 32 and leverages frequency domain features. These frequency-domain features are derived from 1-second segments of time-domain vibration signals by using FFT. The core of the model comprises four one-dimensional convolutional blocks, supplemented by a residual block designed to facilitate deeper model training to prevent the vanishing gradient issue. Each convolutional block consists of a convolutional layer followed by a Rectified Linear Unit (ReLU) activation function. The detailed architecture is as follows: The initial convolutional block includes 24 filters of size 128, followed by the first Max Pooling layer (pool size = 2). The reason for using wide filters in the first convolutional block is that previous research has demonstrated promising results by using wider filters in the initial layers could enhance CNN model performance (W. Zhang, Peng, Li, Chen, & Zhang, 2017; A. Zhang et al., 2019; Yu & Zhou, 2020; Cao, He, Wang, & Yu, 2020; Shao, Ra, Kim, et al., 2024). Then it is followed by a second convolutional block (32 filters of size 3), and the second Max Pooling layer (pool size = 2). Subsequently, a residual block is introduced. The left branch contains two convolutional blocks each with 32 filters of size 3, while the right branch consists of a convolutional layer with a filter size of 1, tailored to adjust the signal dimensions for subsequent addition. Following the merging of both branches, the third Max Pooling layer is applied to reduce the feature vector size. The network concludes with a flatten layer, a dense layer with 200 neurons, a dropout



(A) Regression Model

(B) Classification Model

Figure 2. Network architecture of the proposed residual-based CNN model. (A) Regression model for stage 2. (B) Classification model for stage 3.

layer (a dropout rate of 0.25) to prevent overfitting, another dense layer with 20 neurons, and finally, an output layer that predicts the degradation level. The output of the regression model is defined as $Reg\_Pred \in \mathbb{R}^d$.

3.3. Stage 3: a fine-tuned classification model by using pseudo label techniques

Following the successful training of the residual-based CNN regression model for known degradation level prediction, we proceed to fine-tune a classification model aimed at predicting unknown degradation levels. A significant distinction between this classification model and the preceding regression model lies in the approach to leveraging model outputs. Instead of directly utilizing the regression model's final output, we employ the final embedding from the best regression model from stage 2 as the representation of the classification model's input. The classification outcomes are then derived from this representation via a Softmax activation function, with the model's architecture detailed in Figure 2(B).

A further distinction involves the dataset employed for training. Whereas the regression model in Stage 2 is trained solely on training data, Stage 3 also incorporates testing data with pseudo labels into the training process of the classification model, without prior knowledge of the actual degradation level labels. This is achieved by leveraging the predictions made by the residual-based CNN regression model (Reg_Pred) on the testing dataset from Stage 2, alongside the similarity predictions (Sim_Pred) from Stage 1. A compre-

Algorithm 2 The overall process for pseudo label assignment

```
Get $Sim\_Pred \in \mathbb{R}^{d\times 4}$ from Stage 1.
Get $Reg\_Pred \in \mathbb{R}^{d}$ from Stage 2.
Initialize $Pseudo\_label \in \mathbb{R}^{d\times 11}$ with value 0.
$k \leftarrow 0$
while $k \neq d$ do
    // known class
    if $Sim\_Pred_{(k,0)} == 1$ then
        $Pseudo\_label_{(k,Reg\_Pred_{(k)})} = 1$
    // undetermined class
    else if $Sim\_Pred_{(k,2)} == 1$ then
        $Pseudo\_label_{(k,Reg\_Pred_{(k)})} = 1$
    // unknown class
    else if $Sim\_Pred_{(k,1)} == 1$ then
        Pseudo label assignment based on Table 1
    end if
    $k \leftarrow k + 1$
end while
```

hensive evaluation of these predictions is conducted to generate pseudo labels for the testing data, as outlined in Algorithm 2. Specifically, for testing data classified as either known or undetermined, the pseudo label is directly assigned based on the regression model's prediction, while the detailed strategy for the unknown class is shown in Table 1.

Table 1. Pseudo-label assignment strategy for unknown class.

Pseudo Label Assignment		Stage 1						
		0	1	2	3	4	6	8
Stage 2	0	5 (p=1)						9 (p=0.5) 10 (p=0.5)
	1							
	2							
	3							
	4							
	5	5 (p=1)						
	6	5 (p=0.5) 7 (p=0.5)				5 (p=1)	7 (p=1)	
	7	7 (p=0.5) 9 (p=0.3) 10 (p=0.2)					7 (p=1)	7 (p=0.3) 9 (p=0.4) 10 (p=0.3)
	8	9 (p=0.5) 10 (p=0.5)						9 (p=0.5) 10 (p=0.5)
	9	9 (p=1)						
	10	10 (p=1)						

In the table, the format "Num (p=x)" is used, where "Num" denotes a specific label and "x" indicates the probability associated with that label. Since the degradation label directly reflects the severity of degradation, the relationship between the predictions of Sim_Pred and Reg_Pred is critical. When the predictions are close (i.e., the difference is less than or equal to 2), it suggests that the highest probability for the true label lies between these two values. However, when the prediction difference is significant (i.e., greater than 2), additional considerations must be made. If Sim_Pred exceeds Reg_Pred, it indicates that the regression model's prediction in stage 2 may be less reliable, necessitating greater reliance on Sim_Pred. Conversely, if Sim_Pred is lower than Reg_Pred, it suggests that the similar model's prediction in stage 1 may be less accurate, and Reg_Pred should be given more consideration. Additionally, data from unknown classes will be assigned pseudo labels of 5, 7, 9, or 10 with certain probabilities. Overall, both the training test data with actual labels and the testing data with pseudo labels are utilized to fine-tune the classification model, enhancing its accuracy and robustness in identifying unknown degradation levels.

4. RESULT AND DISCUSSION

In this section, we discuss the performance of ML models across various stages using different evaluation metrics. For the stage 2 regression model's training and evaluation, the 2016 data points from the training dataset were segregated into training, validation, and testing subsets following a 3:1:1 ratio. In the meantime, the division ensures that the stratified distribution of operating conditions within each subset mirrors the overall dataset composition. To fine-tune the stage 3 classification model, we utilize 398 training test data points with true labels alongside 800 testing data points with pseudo labels. And another 813 validation data points are used to validate, ensuring our model's robustness and generalization. Both the regression and classification models are trained with a batch size of 32, initial learning rates set between 0.0001 and 0.00005, and a decay rate of $e^{-0.015}$ determined through preliminary experiments. The training session concludes once the validation loss ceases to decline throughout 15 epochs for stage 2 and 10 epochs for stage 3. Each experiment is repeated 10 times to demonstrate the stability of the model.

4.1. Performance of stage 1 similarity-based model

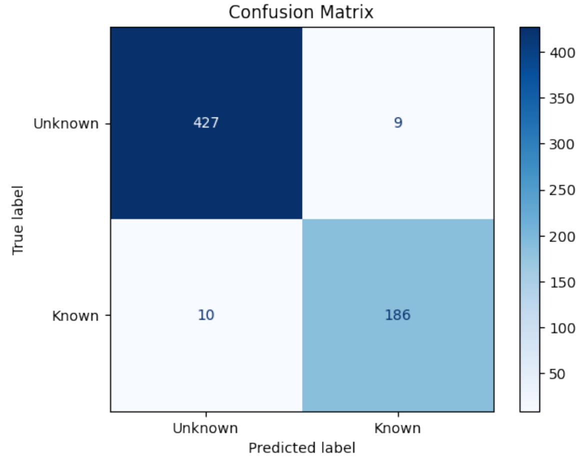
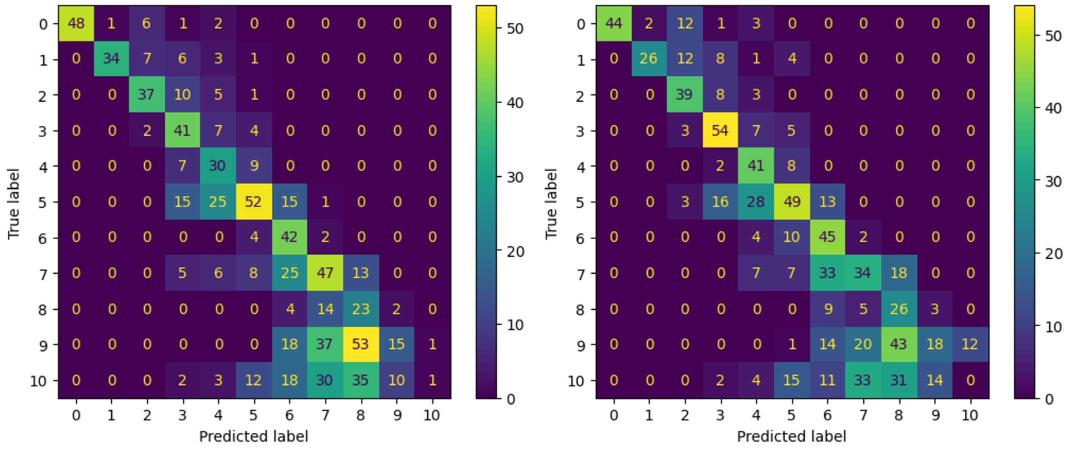


Figure 3. The confusion matrix for known labels and unknown labels classification.

In the first stage, a similarity-based model is developed to



(A) Classification for testing data

(B) Classification for validation data

Figure 4. The confusion matrix for the residual-based CNN model. (A) Classification for testing data. (B) Classification for validation data.

classify testing data into known, unknown, and undetermined classes. The findings, illustrated in Figure 3, indicate that out of the 800 test data points, 168 are assigned as the undetermined class due to the absence of corresponding speed values in the training set or the similarity-based model's inability to classify the data as either known or unknown. For the remaining 632 data points, we achieve a remarkable classification accuracy of 96.99%. This result underscores the effectiveness of utilizing similarity comparisons as a reliable approach for discerning known and unknown classes in ML classification tasks, highlighting the model's robustness and precision.

In this study, we choose Euclidean distance to measure sample similarity over cosine similarity or Manhattan distance due to its suitability for our specific application. While cosine similarity captures the angle between vectors and ignores magnitude, and Manhattan distance provides robustness to outliers but can be less effective in high-dimensional spaces, Euclidean distance excels in capturing absolute differences between feature values. Furthermore, Euclidean distance offers a straightforward geometric interpretation and computational efficiency, making it well-suited to the dimensionality and nature of our dataset. Therefore, this choice ensures precise classification of degradation levels and enhances the reliability of our diagnostic model.

4.2. Performance of stage 2 residual-based 1D-CNN regression model

In the second stage, we focus on training a residual-based CNN regression model to predict known degradation levels. The division for training data resulted in 1219 data points for training (”training train data”), 399 for validation (”training validation data”), and 398 for testing (”training test data”). The regression model provided predictions for each of the 398 testing data points, which were then averaged across their respective 22 segments. The R^2 value for the regression model is **0.9947**. To further determine the degradation level label, we rounded these averages to the nearest integer. The classification results for the training test subset (398 data points) are illustrated in the confusion matrix in Figure 5. This figure demonstrates that the classification accuracy is 97.99%, indicating that our model can accurately predict known degradation levels.

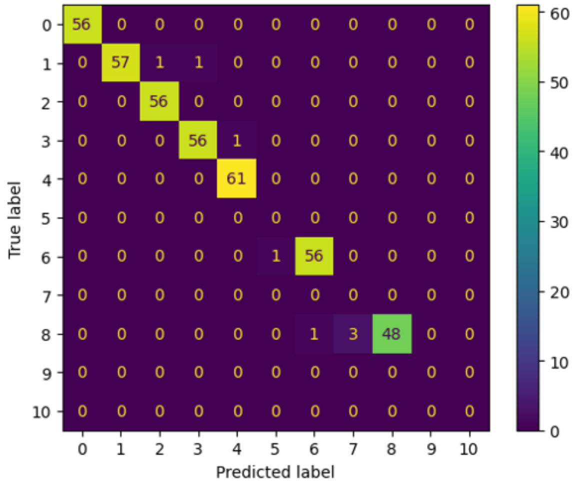
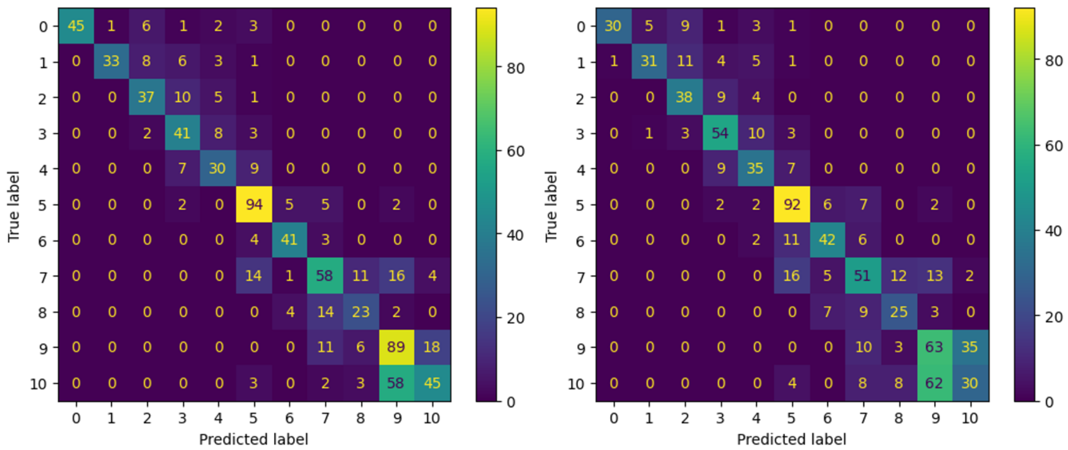


Figure 5. Classification for training test data using the stage 2 model.

However, the model's performance decreases when applied to testing and validation data, as depicted in Table 3. The average classification accuracy for both testing and validation data is 44.3%. The best model achieves an accuracy



(A) Classification for testing data

(B) Classification for validation data

Figure 6. The confusion matrix for the residual-based CNN classification model. (A) Classification for testing data. (B) Classification for validation data.

of 46.25% for both datasets. Figure 4 shows the confusion matrix to represent classification results for the testing and validation data. This figure reveals that the stage 2 model struggles to accurately predict unknown and undetermined classes, particularly for unknown degradation levels 5, 7, 9, and 10. This indicates that while the model effectively predicts known classes, it fails to perform well for unknown and undetermined classes. Even for known degradation levels, some predictions are also not satisfactory. Upon closer examination, we find that the poor predictions predominantly originate from data with unknown operating conditions. This discrepancy underscores the model's limitations and highlights the need for further refinement to enhance its predictive capabilities across a broader spectrum of degradation levels.

4.3. Performance of stage 3 fine-tuned 1D-CNN classification model

The final stage involves the refinement of the classification model through the integration of both training test data with true labels, and testing data with pseudo labels. The integration of pseudo labels is instrumental in bridging the gap between known and previously unidentified or unknown degradation states, thus enriching the model's training dataset. The performance of the refined classification results are presented in Table 3. As demonstrated in Figure 6, there is a notable improvement in predicting unknown degradation levels for both testing and validation data. The classification accuracy of the fine-tuned model has increased by approximately 21%, achieving 67% on the testing data. Furthermore, the performance of the model on the validation data also shows significant enhancement, with an accuracy of 60.4%. This elevation in performance metrics underscores the efficacy of our approach, particularly the utilization of pseudo labels, in enhancing the model's predictive precision. Such an achievement not only demonstrates the model's capacity to adapt and learn from a composite dataset of true and pseudo labels but also highlights its potential applicability in real-world scenarios where distinctions between different states of degradation are critical.

Table 2. Health state score calculation standard.

Distance from true label (k)	Points (Q)	Reported prediction Probability (P)
0 (correct prediction)	1.0	P0
1	0.5	P1
2	0	P2
3	-0.5	P3
4	-1	P4
5	-1.5	P5
6	-2.0	P6
7	-2.5	P7
8	-3.0	P8
9	-3.5	P9
10	-4.0	P10

In the competition, a unique evaluation metric is provided to verify the classification results which is calculated using the following formula:

$$Score_{\text{total}} = \sum(\text{confidenceFactor} \times \sum(\text{predictionProbability} \times \text{healthStateScore}))$$

where the confidenceFactor takes a value of 0.2 for a low-confidence prediction and a value of 1 for a high-confidence prediction. The sum of the predictionProbability for all health states should be less than or equal to 1 for each observation. The healthStateScore is determined by a predefined point system based on the accuracy of the health state prediction. The health state score for each sample is calculated based on the distance difference between prediction and true labels. The

Table 3. Evaluation metric of validation and testing. (Average ± Standard deviation)

Evaluation Metric		Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score	Competition Score (best)
Decision Tree	Testing	0.303 ± 0.006	0.234 ± 0.012	0.303 ± 0.006	0.235 ± 0.008	**-9.5**
	Validation	0.319 ± 0.006	0.249 ± 0.009	0.319 ± 0.006	0.252 ± 0.006	**5**
Random Forest	Testing	0.317 ± 0.004	0.268 ± 0.005	0.317 ± 0.004	0.267 ± 0.005	**11.5**
	Validation	0.334 ± 0.001	0.287 ± 0.006	0.0.334 ± 0.001	0.283 ± 0.001	**41**
Stage 2 Model	Testing	**0.443 ± 0.016**	0.48 ± 0.034	0.443 ± 0.016	0.412 ± 0.017	**389**
	Validation	**0.443 ± 0.014**	0.461 ± 0.021	0.443 ± 0.014	0.415 ± 0.012	**402.5**
Stage 3 Model	Testing	**0.644 ± 0.011**	0.659 ± 0.034	0.644 ± 0.011	0.629 ± 0.020	**600**
	Validation	**0.607 ± 0.009**	0.618 ± 0.028	0.607 ± 0.009	0.591 ± 0.017	**574**
Notice: The first place in the competition: Testing (463.5 / 800) & Validation (472 / 813)						

detailed information is shown in Table 2. Our proposed models have yielded remarkable results in terms of scoring on the test set, **with scores of 600** and validation set **with scores of 574**, respectively. These outcomes are particularly noteworthy when benchmarked against the competitive landscape of the 2023 PHM Conference Data Challenge. Our best model outperformed the first-place score of the competition, which stood at 463.5 (testing data) and 472.0 (validation data), by a substantial margin of **136.5** and **102** points.

4.4. Performance comparison results with other machine learning methods

In addition, we also evaluate tree-based methods, including decision tree and random forest models, using the training data as a baseline and tested their performance on both the validation and testing datasets to compare with our proposed method. Direct application of these models to the original high-dimensional input is impractical due to the substantial feature space. To address this, we first apply principal component analysis (PCA) to reduce the dimensionality to 100 features before training the decision tree and random forest models. As shown in Table 3, the classification accuracy of both tree-based models is approximately 30% lower than that of our best model. Furthermore, these models struggle significantly with predicting unknown classes, particularly classes 9 and 10, often misclassifying them as one of the classes (0 to 8), resulting in almost no correct predictions for these unknown classes. Moreover, many of these misclassifications involve predicting smaller classes when the ground truth is class 9 or 10. This results in a large discrepancy between the predicted and true classes, with a distance often exceeding 6, which significantly reduces the competition score for the tree-based models (refer to Table 2).

4.5. Limitations and Future Work

This study, while contributing valuable insights into gearbox diagnosis, still has several limitations and directions for future research. First, the methodology for classifying data as known, unknown, or undetermined based on similarity comparisons, particularly when considering rotational speed and torque, presents a foundational approach. Yet, the capacity to differentiate among various degradation levels within the unknown classes, given their potentially different similarity degree, poses a challenge for the first stage of analysis. This raises the question of whether it's feasible to distinguish between different degradation levels from the outset. Furthermore, while the CNN architecture employed has demonstrated promising results, exploring more architectures such as inception module (Szegedy et al., 2015), and hybrid residual-inception module, could potentially yield improvements. Additionally, investigating the application of transformer models for this task presents an exciting opportunity for innovation (Vaswani et al., 2017). Currently, our ML model is trained exclusively on frequency domain features. Incorporating diverse data modalities—such as the original time domain, Short-Time Fourier Transform (STFT), and Power Spectral Density (PSD), among others—into the training process could be beneficial. Evidence from previous studies indicates that multimodal machine learning models have the potential to significantly enhance model capabilities (Jiang et al., 2019; D. Wang, Li, Jia, Song, & Liu, 2021; Su, Song, & Ahmed, 2023). Moving forward, there is a critical need to explore more effective and efficient deep learning techniques. Advancements in these areas are crucial for developing robust and comprehensive diagnostic tools for gearbox health monitoring. These improvements can serve as foundational resources for future Industrial Large Knowledge Models (ILKM) (J. Lee & Su, 2024), paving the way for significant enhancements in predictive geabox diagnosis.

5. CONCLUSION

In this study, we develop and implement a three-stage diagnostic approach that represents a significant advancement in the field of gear fault diagnosis. By integrating a similarity-based model, a residual-based CNN regression model, and a refined classification model utilizing pseudo-labeling, we have addressed the inherent challenges of diagnosing gearbox degradation under uncertain and various operating conditions. Our methodology not only showcases the potential of ML models to transcend the limitations of traditional diagnostic methods but also highlights the importance of deep learning and pseudo-label techniques in enhancing model robustness and generalization capabilities. The outstanding performance of our models, as evidenced by the scores (600 (testing data) and 574 (validation data)) in the 2023 PHM Conference Data Challenge, underscores the viability of our approach in contributing to the broader applications and enhanced reliability of mechanical systems.

REFERENCES

Bettahar, T., Rahmoune, C., Benazzouz, D., & Merainani, B. (2020). New method for gear fault diagnosis using empirical wavelet transform, hilbert transform, and cosine similarity metric. *Advances in Mechanical Engineering*, *12*(6), 1687814020927208.

Cao, J., He, Z., Wang, J., & Yu, P. (2020). An antinoise fault diagnosis method based on multiscale 1dcnn. *Shock and Vibration*, *2020*, 1–10.

Chu, F. (2023). Gear pitting fault diagnosis using domain generalizations and specialization techniques. In *Annual conference of the phm society* (Vol. 15).

Duan, J., Shi, T., Zhou, H., Xuan, J., & Wang, S. (2021). A novel resnet-based model structure and its applications in machine health monitoring. *Journal of Vibration and Control*, *27*(9-10), 1036–1050.

Durbhaka, G. K., Selvaraj, B., Mittal, M., Saba, T., Rehman, A., & Goyal, L. M. (2021). Swarm-lstm: Condition monitoring of gearbox fault diagnosis based on hybrid lstm deep neural network optimized by swarm intelligence algorithms. *CMC-Comput. Mater. Continua*, *66*(2), 2041–2059.

Feng, K., Ni, Q., Beer, M., Du, H., & Li, C. (2022). A novel similarity-based status characterization methodology for gear surface wear propagation monitoring. *Tribology International*, *174*, 107765.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

He, Z., Shao, H., Wang, P., Lin, J. J., Cheng, J., & Yang, Y. (2020). Deep transfer multi-wavelet auto-encoder for intelligent fault diagnosis of gearbox with few target training samples. *Knowledge-Based Systems*, *191*, 105313.

Jiang, G., Zhao, J., Jia, C., He, Q., Xie, P., & Meng, Z. (2019). Intelligent fault diagnosis of gearbox based on vibration and current signals: a multimodal deep learning approach. In *2019 prognostics and system health management conference (phm-qingdao)* (pp. 1–6).

Kreuzer, M., & Kellermann, W. (2023). 1-d residual convolutional neural network coupled with data augmentation and regularization for the icphm 2023 data challenge. In *2023 ieee international conference on prognostics and health management (icphm)* (pp. 186–191).

Kumar, A., Gandhi, C., Zhou, Y., Kumar, R., & Xiang, J. (2020). Latest developments in gear defect diagnosis and prognosis: A review. *Measurement*, *158*, 107735.

Lee, D.-H., et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, icml* (Vol. 3, p. 896).

Lee, J., & Su, H. (2024). A unified industrial large knowledge model framework in industry 4.0 and smart manufacturing. *International Journal of AI for Materials and Design*, *3681*.

Li, J., Li, X., He, D., & Qu, Y. (2019). A novel method for early gear pitting fault diagnosis using stacked sae and gbrbm. *Sensors*, *19*(4), 758.

Liu, P. (2023). Interpolate and extrapolate machine learning models using an unsupervised method: An approach for 2023 phm north america data challenge. In *Annual conference of the phm society* (Vol. 15).

Mohamad, T. H., Abbasi, A., Kim, E., & Nataraj, C. (2021). Application of deep cnn-lstm network to gear fault diagnostics. In *2021 ieee international conference on prognostics and health management (icphm)* (pp. 1–6).

Oh, H. J., Yoo, J., Lee, S., Chae, M., Park, J., & Youn, B. D. (2023). A hybrid approach combining data-driven and signal-processing-based methods for fault diagnosis of a hydraulic rock drill. *International Journal of Prognostics and Health Management*, *14*(1).

Saufi, S. R., Ahmad, Z. A. B., Leong, M. S., & Lim, M. H. (2020). Gearbox fault diagnosis using a deep learning model with limited data sample. *IEEE Transactions on Industrial Informatics*, *16*(10), 6263–6271.

Shao, X., Ra, I., Kim, C.-S., et al. (2024). Dsmt-1dcnn: Densely supervised multitask 1dcnn for fault diagnosis. *Knowledge-Based Systems*, *292*, 111609.

Song, Y., Li, Y., Jia, L., & Qiu, M. (2019). Retraining strategy-based domain adaption network for intelligent fault diagnosis. *IEEE Transactions on Industrial Informatics*, *16*(9), 6163–6171.

Su, H., & Lee, J. (2024). Machine learning approaches for diagnostics and prognostics of industrial systems using industrial open source data: A review. *International Journal of Prognostics and Health Management*, *15*(2).

Su, H., Song, B., & Ahmed, F. (2023). Multi-modal machine learning for vehicle rating predictions using image, text, and parametric data. In *International design engineering technical conferences and comput-*

ers and information in engineering conference (Vol. 87295, p. V002T02A089).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Tao, Y., Wang, X., Sánchez, R.-V., Yang, S., & Bai, Y. (2019). Spur gear fault diagnosis using a multilayer gated recurrent unit approach with vibration signal. *IEEE Access*, *7*, 56880–56889.

Vaerenberg, R., Marx, D., Hosseinli, S. A., De Fabritiis, F., Wen, H., Zhu, R., & Gryllias, K. (2023). Predicting pitting severity in gearboxes under unseen operating conditions and fault severities using convolutional neural networks with power spectral density inputs. In *Proceedings of the annual conference of the phm society 2023* (Vol. 15).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vydana, H. K., & Vuppala, A. K. (2017). Residual neural networks for speech recognition. In *2017 25th european signal processing conference (eusipco)* (pp. 543–547).

Wang, D., Li, Y., Jia, L., Song, Y., & Liu, Y. (2021). Novel three-stage feature fusion method of multimodal data for bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–10.

Wang, X., Qin, Y., & Zhang, A. (2018). An intelligent fault diagnosis approach for planetary gearboxes based on deep belief networks and uniformed features. *Journal of Intelligent & Fuzzy Systems*, *34*(6), 3619–3634.

Yu, J., & Zhou, X. (2020). One-dimensional residual convolutional autoencoder based feature learning for gearbox fault diagnosis. *IEEE Transactions on Industrial Informatics*, *16*(10), 6347–6358.

Zhang, A., Li, S., Cui, Y., Yang, W., Dong, R., & Hu, J. (2019). Limited data rolling bearing fault diagnosis with few-shot learning. *Ieee Access*, *7*, 110895–110904.

Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2017). A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. *Sensors*, *17*(2), 425.

Zhao, M., Kang, M., Tang, B., & Pecht, M. (2017). Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes. *IEEE Transactions on Industrial Electronics*, *65*(5), 4290–4300.

Zhou, Y., Zhi, G., Chen, W., Qian, Q., He, D., Sun, B., & Sun, W. (2022). A new tool wear condition monitoring method based on deep learning under small samples. *Measurement*, *189*, 110622.

Zhu, Z., Lei, Y., Qi, G., Chai, Y., Mazur, N., An, Y., & Huang, X. (2023). A review of the application of deep learning in intelligent fault diagnosis of rotating machinery. *Measurement*, *206*, 112346.

BIOGRAPHIES

Hanqi Su received his B.Eng. degree in Robotics Engineering from Southern University of Science and Technology, Shenzhen, Guangdong, China, in 2023. Previously, he participated in a one-year Special Student Program at the Massachusetts Institute of Technology (MIT) during his fourth undergraduate year. He is currently pursuing his Ph.D. in Mechanical Engineering at the University of Maryland, College Park. His research is primarily focused on machine learning, deep learning, prognostics and health management, engineering design, and industrial artificial intelligence.

Jay Lee received his B.S. degree in mechanical engineering from Tamkang University, Taipei City, Taiwan, in 1979, M.S. degree in mechanical engineering from the University of Wisconsin-Madison, Madison, WI, USA in 1983, M.S. degree in industrial management from the State University of New York-Stony Brook, Stony Brook, NY, USA in 1987, and D.Sc. degree in mechanical engineering from George Washington University, Washington, DC, USA in 1992. Currently, he is a Clark Distinguished Chair Professor in Mechanical Engineering at the University of Maryland, College Park, and the founding director of the Industrial Artificial Intelligence Center to tackle the opportunities and unmet needs in the application of Artificial Intelligence to a wide range of industrial applications. Previously, he served as the Founding Director of the NSF I/UCRC on Intelligent Maintenance Systems (IMS) which has developed research memberships with over 100 global companies since 2000 and was selected as the most economically impactful I/UCRC in the NSF Economic Impact Study Report in 2012. He advised his students to win 1st place 5 times in the PHM Data Challenge during 2008-2016 as well as 1st Place in the PHM AP Data Challenge in 2023.