# DeepHealth: A Self-Attention Based Method for Instant Intelligent Predictive Maintenance in Industrial Internet of Things

Weiting Zhang, *Student Member, IEEE*, Dong Yang, *Member, IEEE*, Youzhi Xu, Xuefeng Huang, Jun Zhang, and Mikael Gidlund, *Senior Member, IEEE*

***Abstract*—With the rapid development of artificial intelligence and industrial Internet of Things (IIoT) technologies, intelligent predictive maintenance (IPdM) has received considerable attention from researchers and practitioners. To efficiently predict impending failures and mitigate unexpected downtime, while satisfying the instant maintenance demands of industrial facilities is very important for improving the production efficiency. In this article, a self-attention based "Perception and Prediction" framework, called DeepHealth, is proposed for the instant IPdM. Specifically, the framework is composed of two submodels (i.e., DH-1 and DH-2), which are respectively utilized to perform the health perception and sequence prediction. By operating the framework, the proposed models can predict the health conditions via predicting the future signal samples, thereby completing the instant IPdM. Considering the potential temporal correlation in time series, we deploy an enhanced attention mechanism to capture global dependencies from the vibration signals, and leverage the long- and short-term sequence prediction of sensor signals to support instant maintenance decision-making. On this basis, we conduct a destructive experiment based on the IIoT-enabled rotating machinery and construct a balanced industrial dataset for model evaluations. Extensive experiment results show that the proposed solution achieves good prediction accuracy for instant IPdM on the automatic washing equipment and Case Western Reserve University datasets.**

***Index Terms*—Global dependencies, health perception, industrial Internet of Things (IIoT), instant intelligent predictive maintenance (IPdM), self-attention, sequence prediction.**

## I. Introduction

TRADITIONAL methods of health monitoring have some intractable defects, such as severe hysteresis, high time consumption, and over-reliance on expertise, which always result in the delayed maintenance measures and unexpected downtime. With the enhancement of industrial Internet of Things (IIoT) and artificial intelligence (AI) technologies [1], valuable insights can be captured from the tremendous industrial data and impending failures can be accurately predicted via the AI-empowered approaches. Based on these information, appropriate maintenance decisions can be provided for ensuring the efficient operation of industrial production systems, which is referred to as intelligent predictive maintenance (IPdM) [2], [3].

In recent years, many efforts [4]–[7] have emerged that focus on the fault diagnosis and lifespan prognosis of industrial facilities or key components, and numerous intelligent algorithms have been proposed [8]–[19] to expedite the realization of IPdM. In the initial stage, traditional machine learning algorithms received considerable attention and became extremely effective solutions. Saidi *et al.* [8] presented a novel pattern classification approach for the condition monitoring of bearings, which combined spectral features with a support vector machine for a multiclass task. Martin *et al.* [9] deployed an oblique random forest to build multivariate trees for an induction motor. With the increasing data volume and computing power, neural networks (NNs) became a mainstream learning paradigm and have played an increasingly significant role in IPdM. For instance, Wang *et al.* [10] presented a three-layer feedforward deep neural network (DNN) to perform the failure identification of wind turbine gearboxes. Qin *et al.* [11] developed an integrated model for planetary gearbox fault diagnosis based on deep belief networks. In addition, Shao *et al.* [12] proposed an ensemble model using deep autoencoders, which is a promising learning paradigm for unlabeled data, to conduct rolling bearing fault diagnosis.

Moreover, considering that the time-series characteristics of sensor signals are consistent with the functionality of NNs equipped with recurrent structures, Cheng *et al.* [13] developed

Manuscript received July 23, 2020; revised August 29, 2020 and September 28, 2020; accepted October 5, 2020. Date of publication October 7, 2020; date of current version May 3, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1702000, and in part by the National Natural Science Foundation of China under Grant 61771040. Paper no. TII-20-3545. *(Corresponding author: Dong Yang.)*

Weiting Zhang and Dong Yang are with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: 17111018@bjtu.edu.cn; dyang@bjtu.edu.cn).

Youzhi Xu and Mikael Gidlund are with the Mid Sweden University, 85170 Sundsvall, Sweden (e-mail: Youzhi.Xu@ju.se; mikael.gidlund@miun.se).

Xuefeng Huang and Jun Zhang are with the Beijing Sheenline Group Co., Ltd., Beijing 100044, China (e-mail: huangxuefeng@shenzhou-gaotie.com; zhangjun@shenzhou-gaotie.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2020.3029551

1551-3203 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

a prediction model using long short-term memory (LSTM) networks to predict the failure time of industrial facilities. Zhao *et al.* [14] presented a hybrid model to carry out machine health monitoring, which combines handcrafted features with automatic feature extraction via a bidirectional gated recurrent unit (GRU) network. In addition, spatial features hidden in the sensor sequences are also an effective measure for representing fault features. Accordingly, to extract the spatial correlation of sensor signals, Wen *et al.* [15] deployed an end-to-end learning model using the convolutional neural network (CNN) to extract hidden spatial information from 2-D vibration signals. Chen *et al.* [16] proposed a novel diagnosis method integrating single-layer CNN with an extreme learning machine, and the continuous wavelet transform technique was employed to preprocess the raw vibration signals. Li *et al.* [17] integrated a two-layer CNN with a bidirectional LSTM to respectively represent the spatial and temporal correlation of vibration signals and deployed a common attention module between the LSTM layer and softmax classifier to capture the informative information; thus, limited resources can be focused on those important areas. Furthermore, to address the unbalanced dataset, Guo *et al.* [18] developed a novel framework for the augmentation of faulty data using the auxiliary classifier generative adversarial network, which not only satisfies the urgent requirements of fault discrimination, but also provides effective measures for data generation.

In conclusion, AI techniques have made considerable contributions to the progress of IPdM and have achieved remarkable success in industrial health monitoring, such as fault detection and abnormal identification [19]. However, there are also some intractable challenges that need to be overcome to promote more AI-empowered IIoT applications for industrial intelligence [20], especially for IPdM.

First, it is difficult to collect a balanced dataset with enough labeled data from the continuous degradation process, especially from real industrial scenarios. As a result, most of the current studies are mainly focused on algorithm optimization according to several public datasets, which are mostly acquired from test rigs or simulation platforms, such as those from Case Western Reserve University (CWRU) [21], Intelligent Maintenance Systems [22], and National Aeronautics and Space Administration [23].

Then, it is a great challenge to develop an effective predictive model to provide complex machinery with high-accuracy decisions for instant maintenance. Particularly, most of the existing studies are mainly focused on the health monitoring of current status, which is generally carried out on a given dataset. However, satisfying the instant maintenance of industrial facilities during the run-to-failure process can provide more valuable services for health monitoring, especially for facilities with long lifecycles, which enables the predictive model to identify the health status for the current moment and predict the health status for the future moment simultaneously.

In this article, we mainly focus on addressing the above challenges. First, we conduct a destructive experiment for automatic washing equipment (AWE, which is a dual-bearing rotating machinery for rail vehicle body cleaning, such as a high-speed railway) to construct a vibration signal based industrial dataset, in which each condition will possess enough data volume, including normal condition and fault conditions, and provide predictive models with enough labeled data for model training. Then, we propose a self-attention based [24] "perception and prediction" framework for the instant IPdM, named DeepHealth, which is shown in Fig. 1. Specifically, the framework consists of two submodels (i.e., DH-1 and DH-2), which are respectively deployed to conduct health perception (i.e., step ①) and sequence prediction (i.e., step ②) [25], [26], according to the vibration signals. The predicted sequences of DH-2 then can be fed into DH-1 to carry out the preperceived procedure (i.e., step ③), which enables the DeepHealth to perceive the current condition of the facilities and predict the future condition in advance, thus constructing a close-formed scheme for instant IPdM decisions. The main contributions of this article are summarized as follows.

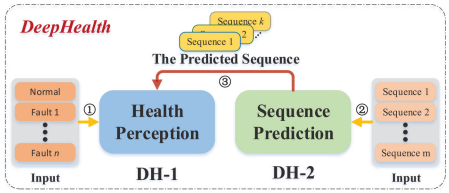



Fig. 1. Framework of the proposed DeepHealth.

1) We propose the DeepHealth framework and design a three-step scheme to achieve instant IPdM, which can predict the future health conditions such that can provide maintenance decisions in advance to avoid machine failures. Specifically, the DeepHealth is composed of two key components (i.e., DH-1 and DH-2), which are deployed to synchronously perform the health monitoring for the current and future moment. Different from the existing works, the proposed scheme predicts the health condition of the future moment by predicting the future signal sample, thereby completing the instant IPdM.
2) To address the long-range dependence between sampling points, we enhance the submodels of DeepHealth with the self-attention mechanism to efficiently capture global dependencies from the high-frequency vibration signals. Particularly, the proposed submodels can directly establish the association between sampling points of the signal sequence through the attention weights, thus the distance between long-range correlated features can be greatly shortened, which is beneficial to the efficient utilization of informative signal features.
3) We collect massive labeled data for model training via the destructive experiments on real industrial facilities, and we have opened the AWE dataset for public access.[1] Based on the dataset, we verify the effectiveness of the proposed self-attention based learning algorithms for the IIoT-enabled instant IPdM, and evaluate the generalization ability of the DeepHealth on the CWRU dataset. Notably, the proposed algorithms ultimately achieve instant

[1][Online]. Available: https://github.com/Intelligent-AWE/DeepHealth

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

prediction accuracies of 84.116% and 99.145% (for the instant IPdM) on both datasets.

The remainder of this article is organized as follows. In Section II, a concentricity deviation based destructive experiment is conducted to collect a balanced dataset for model training. In Section III, the DeepHealth framework and instant IPdM scheme are proposed using the self-attention based techniques. In Section IV, performance evaluations of the proposed algorithms are conducted using the AWE and CWRU datasets. Finally, Section V concludes this article.

## II. Destructive Experiment and Dataset Construction

The previous section noted that building a balanced dataset for model training is one of the great challenges of AI-empowered IPdM. To address this challenge, there is no doubt that the data acquisition system (DAS) and DES play a fundamental role in dataset construction. Accordingly, this section will describe these core components in detail.

### A. Data Acquisition System

To support IIoT applications with different data requirements, a multifunctional DAS is essential for data collection because of the diverse transmission protocols and complex communication environments of industrial scenarios. We, therefore, develop a set of customized DAS to address these heterogeneous demands. The core part is an industrial intelligent gateway (IIG), which integrates the functions of data acquisition, transmission, and storage, such that it provides a fundamental tool and guarantees reliable data collection.

The functionality modules are briefly described as follows.

*1) Acquisition*: The IIG provides four types of communication interfaces, including serial ports (RS232 and RS485), USB interfaces, Ethernet interfaces, and analog quantity (voltage signal and current signal) interfaces.

*2) Transmission*: The IIG integrates 4G, WLAN, and LAN modules to support wireless and wired transmission modes.

*3) Storage*: The IIG possesses 64 GB of storage space, and the limited storage resources support short-term data storage in local databases to provide emergency protection in cases of diminished communication conditions.

In addition, the hardware circuit contains two modules, i.e., the data acquisition board and the core control board. The former is designed for data acquisition and transmission, while the latter is designed for data aggregation and forwarding.

### B. Destruction Scheme

In IIoT-enabled industry scenarios, industrial communication, and networking technologies, such as industrial wireless sensor networks [27], have generated profound impacts on the availability of industrial big data over the past few years and enable high-reliability low-latency collection and transmission of on-site data. However, optimizing the utilization of these large-scale industrial data, extracting the potential information and further providing preperceived services (e.g., IPdM) via AI-empowered methods is the essence of the data value. In addition, except for a small number of research fields, more areas have problems with limited data volume and uneven quality, which are insufficient to support the training of ML models, especially DNN models. If only public data are used, the quantity of data is far from sufficient, resulting in poor system performance. Thus, building a large-scale dataset with sufficient annotations for the collected data becomes a considerable and inevitable challenge. Traditional annotation methods, unlike some computer vision applications, can be completed by ordinary people or data companies, while in IIoT-enabled fields, the requirements for data annotation are stringent. If executing the annotation step after data acquisition and only according to engineering experience, it is time-consuming and laborious and extremely unsuitable for enormous sensor data streams. There are two main reasons.

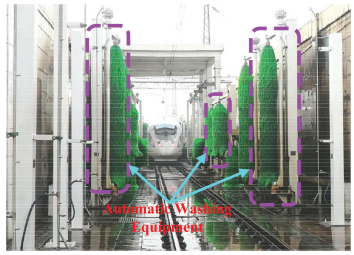



Fig. 2. Application scenario of automatic washing equipment.

*1) Varied Vulnerability Degree of Components:* Components, such as blades, usually have shorter life-cycles such that the full-life-cycle data are easy to collect. However, components, such as bearings, that wear at low speeds require a relatively slow process. As a result, the condition transition from normal to faulty is a long-term process, and it becomes unrealistic to accurately identify different categories of data from a large database. Moreover, the differences between two types of data are reflected in multiple dimensions, such as their amplitude or density, so it is impractical to distinguish them manually according to the experience of engineers.

*2) Expertise Alone is not Enough:* The traditional method of "collect first, annotate later" encounters difficulties in the sensor sequence. Since the sequence contains a strong correlation, it is cannot always determine the specific health condition based on a single sampling value but needs the intercorrelations of a relatively long sequence. As a result, the sensor sampling values cannot be correctly annotated even with solid expertise.

With the comprehensive deployment of networks of high-speed railways, inspection for rail vehicles and their fundamental facilities is an important component. Particularly, the AWE (an IIoT-enabled dual-bearing rotating machinery for rail vehicle body cleaning) is the first checkpoint of rail vehicles during the whole inspection process, and it plays a significant role in ensuring a safe and efficient maintenance process. As shown in Fig. 2, one washing rail contains dozens of AWEs of different types, and each AWE is empowered with different functionalities. As a consequence, keeping each vehicle running reliably and further guaranteeing the stability of the whole washing rail is faced with

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

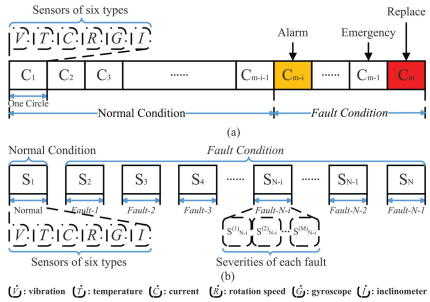



Fig. 3. Destructive scheme. (a) Run to Fail. (b) Destructive experiment.

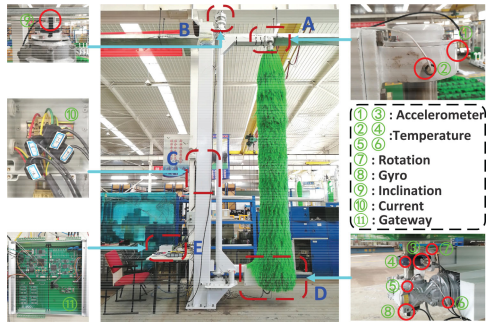



Fig. 4. Destructive experimental platform.

great challenges. For a single AWE, the critical part is located at the front end, which is composed of a central shaft that supports the rotation of the brush set (called the brush shaft) as well as two rolling bearings at the upper and lower ends of this shaft. However, during actual operations, the brush shaft is subjected to the limitations from the load resistance and the self-rotation driving force, thereby causing the concentricity to deviate between the upper and lower bearings. This phenomenon leads directly to a diminished cleaning service and even causes safety accidents due to shaft fractures. Based on these situations, this section introduces a DES (i.e., an industrial data annotation method) for a dual-bearing rotating machine, as shown in Fig. 3, to collect enough labeled data and construct a balanced dataset for model training. Notably, we perform verifications using the AWE (a real situation shown in Fig. 2) and propose a system-level IPdM strategy in the next section.

*1) Run to Fail:* Fig. 3(a) presents a complete "run to fail" process for the objects. For convenience, the condition transformation of a specific object is divided into normal and fault categories in the whole life cycle. It is worth noting that we tend to focus on the data under fault conditions. Here, $C$ is defined as a time slot in which the AWE devices complete an operation (i.e., a working circle), which can be continuously performed or at an interval. The equipment sequentially operates one time $(C_1)$, two times $(C_2)$, until many times $(C_m)$. During the entire operation process, we cannot know at what moment the health status of the machine will change. For this reason, it is challenging to collect a balanced dataset for each health condition. In addition, as mentioned above, the occurrence of any incipient faults requires a long waiting time, and continuous data collection involves alerts with respect to the storage and efficiency of the DAS.

*2) Destructive Scheme:* Fig. 3(b) shows the destructive experiment scheme, where $S$ represents the health status of the equipment, including one normal class $S_1$ and $(N-1)$ fault classes $[S_2, S_N]$. Each condition is independent of other conditions; in some cases, it may also contain $M$ quantified degrees of wear for each condition. In this scheme, the continuous degradation process is discretized into limited conditions; additionally, the nonstationary random process is transformed into a stationary random process by means of human intervention damages, which enables the equipment of various fault conditions to rapidly collect sufficient data for the corresponding conditions. If the measurement differences (e.g., environment noises) under each condition are refined enough, the data collection will have two ideal characteristics.

1) The health conditions of the equipment can be reduced to a finite condition space from the original infinite condition space.
2) The discrete stationary stochastic process can be used to approximate the continuous nonstationary stochastic process due to the finite condition space defined in the destructive scheme and the stable measurement environment during the data collection.

In summary, through destructive experiments on the actual equipment, a large quantity of valuable annotated data can be generated for each condition, which means that the status of the equipment can be abstracted from an infinite space to a finite space to create a quasi-stationary stochastic process.

### C. Dataset Construction

To guarantee the safety and feasibility of the dataset construction, we conduct a destructive experiment using a fire-new machine in the factory environment. The experimental platform of the AWE and the deployment of DAS are shown in Fig. 4. There are five types of sensors attached to the AWE prototype, including accelerometers, temperature sensors, hall rings, gyros, and rotation speed sensors, to perceive the operating conditions of this machine in real time.

*1) Destruction mode*: As shown in Fig. 4, the bearings on the upper and lower ends of the brush shaft are connected to the support arm through 4 screws. At the joint location of the upper bearing (i.e., dotted box **A** in Fig. 4), four gaskets are added to these four screws and remain unchanged to the lower bearing. In this way, two bearings deviate from each other, and the original concentricity changes through this manual intervention. Specifically, Fig. 5 explicitly demonstrates the implementation process, where Fig. 5(a) shows the size of each gasket (approximately 3 mm), which is added to the bolt and results in the progressive increase in the concentricity deviation of the dual-bearings. Thus, we take 3 mm as one health condition and add the gasket to the bolt from one chip to five chips [i.e., Fig. 5(b)–(f)]. Here, the maximum of five gaskets can be added, which is prudently determined according to the concentricity deviation degrees of the AWE in real situations.

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

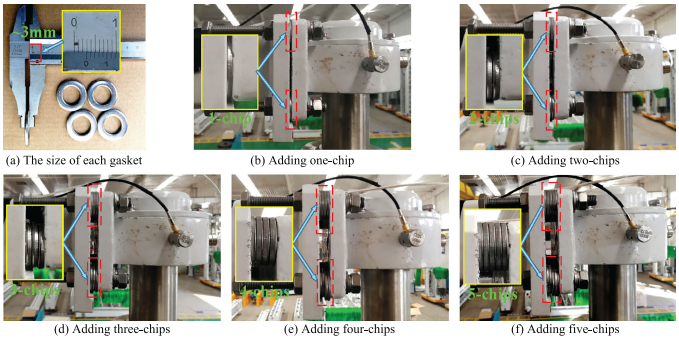



Fig. 5. Implementation of the destructive experiments. (a) Size of each gasket. (b) Adding one-chip. (c) Adding two-chips. (d) Adding three-chips. (e) Adding four-chips. (f) Adding five-chips.

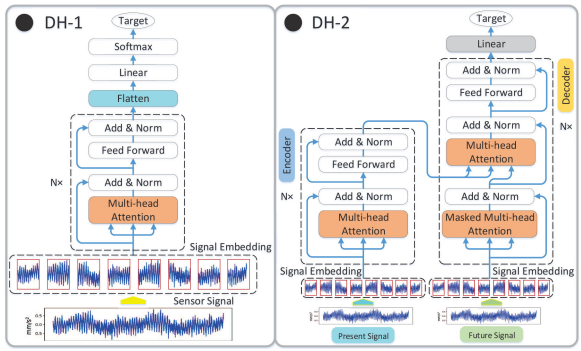



Fig. 6. Architecture of DeepHealth, composed of DH-1 and DH-2.

*2) AWE dataset*: Notably, two single-axis accelerometers are placed on the metal surfaces of the upper (i.e., dotted box **A** in Fig. 4) and lower (i.e., dotted box **D** in Fig. 4) bearings, respectively, with a sampling frequency of 4 kHz. The operating data of the AWE in each condition are collected separately, according to the established fault modes. On the basis of the actual working conditions, we manually control the start and stop of the equipment, and the cumulative collection time of each condition is approximately 2 h. Thus, we obtain a sufficient quantity of balanced annotation data for six health conditions (i.e., one normal condition and five fault conditions) while simultaneously satisfying the data requirements of being independently and identically distributed.

## III. Proposed "DeepHealth"

Most of the popular sequential models are based on the recurrent [28] structures, which is focused on the temporal-correlation of sequence modeling. Although the recurrent structure is helpful for the feature representation of sensor signals, the long short-term dependence of high-frequency vibration signals is hard to capture. The objective of this article is to achieve the instant IPdM, which requires the sequential models are equipped with the abilities of health perception (i.e., classification, for current health condition) and sequence prediction (i.e., regression, for the next/future signal sequence). Particularly, the health perception of future conditions depends heavily on the performance of sequence prediction. Thus, we need to design an effective algorithm to efficiently capture long-range dependencies for sequence prediction. Remarkably, the Transformer [29] provides a novel architecture for end-to-end prediction tasks, and its core components are composed of stacked self-attention modules, which is designed to capture the global dependencies between inputs and outputs while achieving high parallel computing power and operational efficiency, and has been proven to be better at sequence prediction tasks [30]. Therefore, inspired by the efficient learning mechanism of Transformer, this article is focused on the long short-term sensor signal prediction based on the self-attentional networks.

As shown in Fig. 6, to achieve accurate health perceptions and long short-term sensor sequence predictions, we divide this process into two procedures and propose an enhanced framework named "DeepHealth," which contains two complementary submodels based on the self-attention mechanism, i.e., DH-1 and DH-2, for these two tasks. Specifically, DH-1 is deployed to assess the health conditions of the industrial facilities, and DH-2 is utilized to predict the next sequence of sensor signals. Then, the signal sequence that is generated by DH-2 is fed into DH-1 to identify the corresponding health condition according to the predicted sensor sequences and further realize the instant IPdM. In this section, we describe the details of the proposed submodels.

### A. Self-Attention Based Health Perception Model

As shown in the left part of Fig. 6, the DH-1 model consists of two core modules, namely, the signal embedding module and the model component module.

*1) Signal Embedding Module-1*

We designed a signal embedding method for the sensor sequence, which can be applied to many types of sensor sequence data, such as vibration, current, and temperature. The method consists of two steps: dataset segmentation and signal segment embedding. This article takes the high-frequency vibration signals as an example.

a) First, the original sample set is divided into lightweight segmentation (for large-scale) and heavyweight segmentation (for small-scale), which are respectively realized using equal and sliding segmentation. Since a sufficient number of vibration signals from the AWE dataset were collected through the destructive experiments, the original dataset was therefore equally divided to form the sensor signal sequence.

b) Second, the equal-length segment is divided again to form a certain number of equal subsegments. The number of segments here is the same as the number in multihead attention. After that, the segment is then spliced to form an $n \times m$ embedded matrix $X_n^m$, where $n$ represents the number of heads and $m$ represents the length of each subsegment.

2) *DH-1 Model*

a) Scaled dot-product multisegment attention

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

The inputs of the Transformer block first flows through a scaled dot-product and multisegment attention layer, which is generally called the "self-attention layer." This layer helps the block to simultaneously focus on other subsegments in the same signal sequence when it represents a specific segment.

According to [29], an attention function can be defined as mapping a query and a set of key-value pairs to an output, and each of these elements is composed of a vector. First, three weight matrices $W^Q$, $W^K$, and $W^V$ are randomly initialized, and then they are iteratively trained during the learning process. Next, we multiply each input vector (in this article, which indicates the embedding of sensor sequences) by these three matrices to create the query vectors, key vectors, and value vectors, respectively. Afterward, to traverse the input sequence and score each subsegment of the input sequence against this subsegment, thereby determining how much focus to place on other subsegments in order to completely represent a certain subsegment. Specifically, the score is computed using the dot-products of the query with all keys. Then, each component is divided by $\sqrt{d_k}$ (the scaling factor, which is designed to have more stable gradients), and fed into a softmax function to acquire the attention weights of the value vector. Thus, the matrix of outputs is computed as follows:

$$\begin{aligned}\text{segment}_i &= \text{Attention}(Q^*W_i^Q, K^*W_i^K, V^*W_i^V)\\ &= \text{softmax}\left[\frac{(Q^*W_i^Q)(K^*W_i^K)^T}{\sqrt{d_k}}\right](V^*W_i^V)\end{aligned} \quad (1)$$

where the parameter matrices are $W_i^Q \in \mathbb{R}^{d_{\text{sequence}}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{sequence}}\times d_k}$ and $W_i^V \in \mathbb{R}^{d_{\text{sequence}}\times d_v}$. Notably, although $Q^*$, $K^*$, and $V^*$ are the three matrices that we defined, they will be replaced by an identical object (i.e., the input sample that is formed by the signal embedding). As mentioned above, we split the equal-length sequence into a set of equal subsegments. Instead of deploying a single attention function, multisegment attention allows the model to learn information from these subsegments in a parallelized way. The association between sampling points can be established via the attention weights, which contributes to the efficient utilization of informative signal features. Then, a representation matrix is formulated by concatenating these segments and multiplying them by an additional weight matrix $W^o$

$$\begin{aligned}&\text{MultiSegment}(Q, K, V)\\ &\quad = \text{Concat}(\text{segment}_1, \cdots, \text{segment}_h)W^o\end{aligned} \quad (2)$$

where the parameter matrix is $W^o \in \mathbb{R}^{hd_v\times d_{\text{sequence}}\times d_k}$. In this article, we employ eight parallel self-attention modules (i.e., $h = 8$), and use $d_k = d_v = d_{\text{sequence}}/8$.

b) Feedforward NNs

In addition to the multisegment attention layer, the model also contains a fully connected feedforward network in the Transformer block, which is simultaneously deployed on each subsegment. Specifically, it consists of two linear transformations and a nonlinear activation function with a ReLU unit

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (3)$$

where $x$ refers to the output of the multisegment attention layer, and $W_1$, $W_2$, $b_1$, and $b_2$ refer to the weights matrices and biases vectors of this layer, respectively.

c) Residual connection and layer normalization

As shown in the left part of Fig. 6, an operation of residual connection [31] is adopted in each layer of the Transformer block. Formally, this operation can be defined as follows:

$$y = \Phi(x, W^i) + x \quad (4)$$

where $x$ and $y$ are the input and output of the layers that are considered, respectively, and $W^i$ is the corresponding weight matrix. The function $\Phi(\cdot)$ indicates the residual representation that is to be learned. Specifically, the residual has two formations in this article. One formation is $\text{MultiSegment}(x)$, and the other formation is $\text{FFN}(x)$. In addition, layer normalization is applied after the residual operation.

d) Linear transformation and cost function

After the input flows through the Transformer block, a multidimensional representation matrix is formulated. To complete the final classification of the $N$ classes, we first implement a "flatten" operation to convert this matrix into a 1-D vector. Then, we multiply this operation by a trainable matrix to generate the hidden representation $v$, which has the dimension of $[d_{ff}, N]$. The entire transformation process is described as follows:

$$v = \text{flatten}(x) \cdot W_l + b_l \quad (5)$$

where $x$ refers to the output vector of the Transformer block, and $W_l$ and $b_l$ refer to the weight matrix and bias vector of the linear layer, respectively.

Finally, the vector $v$ is fed into a softmax classifier to generate the final health condition $h_\theta(x)$

$$h_\theta(x) = \text{softmax}(v). \quad (6)$$

The cross-entropy cost function can be defined as follows:

$$\min_\theta J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} 1\{y_i = j\}\log h_\theta(x_i)\right] + \frac{\lambda}{2}\sum_{i=1}^{m}\sum_{l=1}^{L}\theta_{il}^2 \quad (7)$$

where we use L2 regularization to modify the loss function, and $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$ denote the training inputs and labels, respectively. $1\{\cdot\}$ is an indicative function that returns 1 if the value in the parentheses is true, and 0 otherwise. $\lambda$ is a decay term of the L2 regularization.

### B. Self-Attention Based Sensor Sequence Prediction Model

As shown in the right part of Fig. 6, the DH-2 model is made up of two core modules, namely, the encoder and decoder.

1) *Signal Embedding Module-2*

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

This signal embedding module is divided into two steps: sequence-pair formulation and segment partition.

The sequence generation [32] task generally focuses on the random generation ability of similar sequences, and there is not necessarily a correlation between the input and output sequences. In contrast, the sequence prediction task needs to predict the future segment based on the previous time-series signal, and there is inevitably a potential internal correlation between both sequences. In practice, there is a need to input the two respective sequences for both the encoder and decoder, and the input of the decoder $x_{\text{encoder}}$ is the predicted result of the encoder $x_{\text{decoder}}$. We therefore call the combination of $[x_{\text{encoder}}, x_{\text{decoder}}]$ the sequence-pair.

To formulate the sequence-pairs, we first split the original dataset into equal segments with different sequence lengths

$$\text{Dataset} = \{\text{Seg}_1, \text{Seg}_2, \cdots, \text{Seg}_{n-1}, \text{Seg}_n\}. \tag{8}$$

Next, the corresponding segments from the original dataset are shifted and intercepted to form the $x_{\text{encoder}}$ and $x_{\text{decoder}}$

$$\begin{aligned} x_{\text{encoder}} &= \text{Dataset}[1:(n-1)] \\ &= \{\text{Seg}_1, \text{Seg}_2, \cdots, \text{Seg}_{n-1}\} \\ x_{\text{decoder}} &= \text{Dataset}[2:n] \\ &= \{\text{Seg}_2, \text{Seg}_3, \cdots, \text{Seg}_n\}. \end{aligned} \tag{9}$$

Finally, to align both inputs, the combined input of DH-2 can be obtained, which is represented as follows:

$$x = \{(\text{Seg}_1; \text{Seg}_2), (\text{Seg}_2; \text{Seg}_3), \cdots, (\text{Seg}_{n-1}; \text{Seg}_n)\} \tag{10}$$

where $x$ denotes the final input formation of DH-2, and $n$ denotes the sequence index. The segment partition procedure is exactly the same as that described in Section III-A.

2) *DH-2 Model*

a) Encoder–Decoder

The structure of the encoder–decoder has been broadly applied in sequence generation scenarios. In practice, CNN and RNN models are frequently deployed as the encoder and decoder, respectively. The encoder represents the original sequence $\{x_1, \cdots, x_n\}$ as a context vector $c$, and then generates the target sequence $\{y_1, \cdots, y_n\}$ based on the decoder.

As is known, the CNN extracts the features layer-by-layer through a series of convolution kernels, and the RNN divides the sequence into many time steps that represent the original sequence step-by-step. Such approaches achieve good performance but result in high computational complexity and low operational efficiency. However, the Transformer employs stacked self-attention and fully connected layers for both the encoder and decoder with entire parallel computing patterns. Obviously, the Transformer model provides an efficient solution for sequence tasks.

As shown in Fig. 6, the encoder and decoder are composed of stacks of $N = 3$ identical blocks, as mentioned above. Unlike the encoder, a modified self-attention sublayer, called masked multihead attention, is integrated into the decoder block to ensure that the following part of position $i$ can be shielded when predicting position $i$.

In summary, the specific process of sequence prediction can be expressed as follows:

$$y = \text{Decoder}[\text{Encoder}(x_{\text{encoder}}); x_{\text{decoder}}] \tag{11}$$

where $\text{Encoder}(\cdot)$ and $\text{Decoder}(\cdot)$ respectively denote the derivation procedures in Section III-A, $x_{\text{encoder}}$ denotes the input of the encoder and $y$ denotes the output of the decoder.

b) Linear transformation

To generate the predicted sequence, a linear transformation [as shown in (5)] needs to be deployed to convert the dimension of the decoder output to $d_{\text{sequence}}$.

### C. Orchestration of Health Perception and Sequence Prediction Models

In general, to implement an AI model, we need two phases, i.e., offline training and online execution. In the offline training phase, we train the health perception (i.e., DH-1) and sequence prediction (i.e., DH-2) models for hundreds of epochs. After the models are well-trained (i.e., achieving convergence), they can be deployed in the IIG as the callable programs. In the online execution phase, the real-time data streams are acquired in a certain sampling period, and the models can be fetched to execute the specific tasks. If a signal sample is fed into the DeepHealth, the DH-1 can identify the health condition of the industrial facility based on the current sample, while the DH-2 can predict the sample of the next moment. Then, feed the predicted sample into the DH-1, the corresponding health condition can also be identified. Iteratively, the DeepHealth can monitor the health conditions continually and ensure the efficient operation of industrial facilities.

## IV. Experiments

The primary objective of this article is to build an AI-empowered instant IPdM system by exploring the health perception and sequence prediction in the case of IIoT-enabled dual-bearing rotating machinery. This section will introduce the details about the scheme verification and further evaluate the performance of IIoT-enabled instant IPdM.

### A. Descriptions of the Datasets

*Dataset 1*: As described in Section II, we developed a DAS for data collection, designed a DES for data annotation, and constructed the AWE dataset for the IIoT-enabled dual-bearing rotating machinery. Notably, the DES is based on concentricity deviation, which is implemented by manual destructions; thus, a balanced dataset with enough labeled data for six health conditions is collected, including one normal condition and five fault conditions.

Admittedly, vibration signals have been widely used in the health perception of bearings, which is always considered the optimal means for representing the real-time operation status of industrial machinery. As a result, the vibration signal that is

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

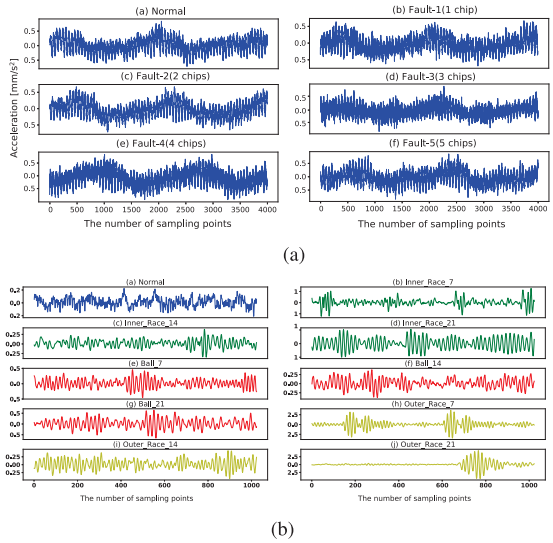



Fig. 7. Original signal segments of (a) AWE and (b) CWRU dataset.

collected via the sensors ① (i.e., the accelerometers) in Fig. 4 is selected for the model learning, and the original signal is shown in Fig. 7(a). Here, it only shows a one-second segment of sampling points for each condition because the sampling frequency is 4 kHz. Actually, we also collect four other types of operational data, but this article is focused on the vibration signal for the IPdM so that we only demonstrate the segments of vibration signals. In addition, considering the fairness and rationality for model training, we select 1000 s of signals for each condition; thus, the total sampling volume of the constructed signal set used for experiments includes 2.4e+07 sampling points.

*Dataset 2*: To sufficiently evaluate the effectiveness and generalization ability of the DeepHealth, we also carry out additional verifications using a public dataset, which is provided by the bearing data center at CWRU [21]. The original signal segments are shown in Fig. 7(b), which contains ten health conditions: one is normal and the other nine are faulty.

Specifically, we use the vibration signals of the "normal baseline data" and "48k drive end bearing fault data," which represent normal data and single-point fault data, respectively (including three fault types and three severities for each type, i.e., fault diameters of 7 mils, 14 mils, and 21 mils). Both of these subdatasets are collected via a test stand under a motor load of 3 hp. The sampling frequency of the vibration signal is 48 kHz, and the total number of sampling points is approximately equal to 4.8e+06. Obviously, there is an intractable bottleneck in that the sample volume cannot guarantee efficient parameter updating and convergence. As a result, a data augmentation method called equitable sliding stride segmentation (ESSS) described in our previous work [33] is used to augment the sample scale of the CWRU dataset.

## B. Experimental Setup

In this section, we instantiate the DeepHealth and evaluate the model performance of DH-1 and DH-2 on both datasets. To build the sample sets for model training, valuating, and testing, both datasets are segmented in a certain sequence length. To ensure the randomness of samples, we shuffle the whole sample sets and randomly sampling in the ratio of 7:2:1 to build the corresponding subsets. Thus, 70%, 20%, and 10% samples are utilized to train, valuate, and test the learning performance, respectively. Specifically, we train the DH-1 and DH-2 models on the training sets for 300 epochs. After each epoch, we valuate the models on the valuating sets, thereby the learning effects can be monitored. After the models are well-trained, the testing sets are utilized to evaluate the model performance. As shown in Fig. 1, the DeepHealth framework is composed of three procedures. Accordingly, to complete the instant IPdM, we setup a three-step experiment to verify the effectiveness of the proposed solution.

First, we evaluate the health perception performance of DH-1 on both datasets. As shown in Fig. 6, the input of DH-1 is a vibration signal sample with a certain sequence length, while the output is a discrete value (0, 1, or others), which represents a corresponding health condition. As described in the previous section, both datasets contain several types of health conditions. DH-1 is specifically employed to represent the various types of signal samples in the datasets and accurately identify the health conditions that are reflected by the real-time signal samples. In other words, when feeding a signal sample into DH-1, it needs to identify the current health condition according to this sample, which is also the fundamental function of this submodel. In addition, to explore the representation ability for sensor sequences with different lengths, both datasets are segmented into six sequence lengths (i.e., $[128, 256, 512, 1024, 2048, 4096]$) to train the DH-1 model. As a result, six well-trained DH-1 submodels are obtained. To comprehensively verify the model performance, the prediction accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curve and area under curve (AUC) value are used as evaluation metrics. Specifically, for the classification tasks, the calculation of these metrics is generally according to the classification results of positive and negative samples, which can be represented as a confusion matrix containing four parts, i.e., true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The corresponding formulas are given as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - \text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

$$\text{AUC} = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1}) \tag{16}$$

where $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$ denotes the sequential connection points on the ROC curve, and $x_1 = 0$, $x_m = 1$. In addition, the time consumption, including the training time and testing time, is also considered.

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DESCRIPTION OF THE TRAINING, EVALUATING, AND TESTING PROCESS OF DEEPHEALTH

| Dataset | Sequence length | Health Perception (DH-1) | | | | | | | Sequence Prediction (DH-2) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Data volume | Time consumption | | Metrics (Testing) | | | | Dataset descriptions | | Time consumption | |
| | | | Train | Test | Accuracy | Precision | Recall | F1-score | Train/Valuate/Test | Total volume | Train | Test |
| ***Our (AWE)*** | 128 | 187500 | 5980.5 s | 1.292 s | 69.52% | 70.30% | 69.52% | 69.54% | 172500/7499/7499 | 187498 | 5594.7 s | 1.719 s |
| | 256 | 93750 | 4505.9 s | 1.165 s | 95.87% | 95.86% | 95.87% | 95.86% | 84374/4687/4687 | 93748 | 6971.1 s | 1.573 s |
| | 512 | 46872 | 4109.6 s | 0.983 s | 98.78% | 98.77% | 98.78% | 98.77% | 42184/2343/2343 | 46870 | **5565. 5 s** | 1.371 s |
| | 1024 | 23436 | **4019.6 s** | **0.887 s** | **98.81%** | **98.80%** | **98.80%** | **98.80%** | 21092/1171/1171 | 23434 | 5602.2 s | **1.275 s** |
| | 2048 | 11718 | 11874.8 s | 1.047 s | 98.29% | 98.29% | 98.29% | 98.28% | 10546/585/585 | 11716 | 17589.8 s | 1.473 s |
| | 4096 | 5859 | 21606.2 s | 1.238 s | 91.96% | 91.95% | 91.95% | 91.96% | 5270/292/292 | 5854 | 25411.1 s | 1.794 s |
| ***CWRU*** | 128 | 37760 | 1484.1 s | 0.450 s | 81.25% | 81.45% | 81.25% | 81.26% | 33984/1887/1887 | 37758 | 2520.9 s | 0.774 s |
| | 256 | 18880 | **906.2 s** | **0.352 s** | 85.80% | 86.40% | 85.80% | 85.85% | 16992/943/943 | 18878 | 1285.9 s | 0.660 s |
| | 512 | 16100 | 1399.2 s | 0.441 s | 93.41% | 93.61% | 93.41% | 93.38% | 8496/471/471 | 9438 | **1093.9 s** | 0.650 s |
| | 1024 | 16080 | 2801.7 s | 0.672 s | 98.32% | 98.32% | 98.32% | 98.32% | 4248/235/235 | 4718 | 1149.6 s | **0.615 s** |
| | 2048 | 16050 | 15816.8 s | 1.225 s | 99.56% | 99.57% | 99.56% | 99.56% | 2124/117/117 | 2358 | 6122.7 s | 0.647 s |
| | 4096 | 15980 | 55860.2 s | 3.253 s | **99.88%** | **99.88%** | **99.87%** | **99.87%** | 1062/58/58 | 1178 | 10533.9 s | 0.704 s |

Note: Due to the limited data volume of the CWRU dataset, we therefore deploy a data augmentation method called equitable sliding stride segmentation (ESSS) [33] for the sequence lengths of 512, 1024, 2048, and 4096.

Second, we evaluate the sequence prediction performance of DH-2 on both datasets. As depicted in Section III-B, the DH-2 model is designed to perform the sequence-to-sequence prediction that maps a predicted equal-length sequence from a given equal-length sequence. As shown in Fig. 6, the input of DH-2 is a signal sample-pair including a sample (the left side of the input) of current sampling period and a sample (the right side of the input) of next sampling period, and these two samples are sequentially continuous. The training stage requires feeding these two samples into the model simultaneously, and then DH-2 learns the mapping relationships between the two samples. The testing stage requires feeding a current sample into the model, and then it needs predict a next/future sample according to the current sample. Consequently, to explore the fitting capacity of signal sequences, the sequence pairs are also segmented into six sequence lengths (i.e., $[128, 256, 512, 1024, 2048, 4096]$) through the signal embedding method that was described in Section III-B. Notably, the selection of sequence lengths is consistent with the first experiment. Similarly, each sequence length corresponds to a specific submodel of sequence prediction, and each submodel possesses the ability to predict sequences with potential correlations. The metrics of root mean squared error (RMSE) and mean absolute error (MAE) are used to evaluate the prediction performance. The equations are given as follows:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \\ \text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \end{aligned} \tag{17}$$

where $y_i$ and $\hat{y}_i$ denote the true value and predicted value of the samples, respectively. $N$ indicates the sequence length.

Third, we synthetically evaluate the performance of instant maintenance decisions for the AI-empowered IPdM. Based on the above evaluations, it should be evaluated that the DH-1 model is capable of identifying the current operating status for dual-bearing rotating machinery according to real-time sequence flows, while the DH-2 model possesses the capability of predicting the next signal sequence according to the current sequence.

TABLE II
COMPARISONS OF SEVERAL ALGORITHMS ON THE AWE DATASET

| Method | Length of Input | Accuracy |
|---|---|---|
| LSTM | 256 | 91.55% |
| GRU | 256 | 92.46% |
| LeNet | 1024 | 95.58% |
| DH-1 | 1024 | 98.81% |

If the DH-1 and DH-2 are operating together, the submodels can simultaneously perform the specific tasks. When a signal sample is fed into the DeepHealth system, the health condition of the future moment should be predicted in advance according to the predicted samples of DH-2. As a result, the third step comprehensively evaluates the instant perception performance of DeepHealth by feeding the predicted sequences of DH-2 into DH-1. The accuracy, precision, recall, F1-score, ROC, and AUC are used as evaluation metrics to compare the instant prediction performance, and the time consumption is also considered.

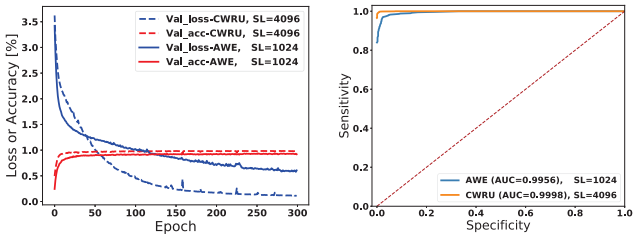
## C. Experimental Results

For the first experiment, evaluations of the perception capacity of DH-1 were completed. Table I demonstrates the various performance metrics for both datasets during its training and testing processes. Overall, when the sequence length is set to 1024 and 4096, the proposed model achieves the best performance in terms of accuracy, precision, recall and F1-score. Although both datasets are composed of vibration signals, the optimal performance of this model still exhibits considerable differences. Obviously, DH-1 obtains the best identification performance for the CWRU samples with a sequence length of 4096, while the best results for the AWE samples appear when the sequence length tends to 1024. The reason is mainly caused by the complexity of the vibration signals and the high similarity of signal samples between different conditions. Table II shows the results of several algorithms, including LSTM, GRU, and LeNet, on the AWE dataset. It can be seen that the DH-1 model achieves the best accuracy; additionally, it also proves that the self-attention mechanism is an effective method for capturing global dependencies of sensor sequences. Table III

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPARISONS OF DH-1 WITH RELATED WORKS ON THE CWRU DATASET

| Method | Model Input | Task Complexity | Accuracy |
|---|---|---|---|
| biGRU [14] | Local features | 4-class | 99.60% |
| CNN+ELM [16] | Wavelet scalogram | 10-class | 99.92% |
| CNN+biLSTM +Attention [17] | Times-series | 10-class | 99.74% |
| DeepHealth (DH-1) | Times-series | 10-class | 99.88%±0.06 |

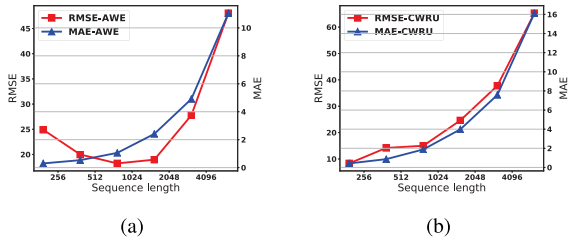



(a) (b)

Fig. 8. Metrics of DH-1 in the evaluating and testing processes. (a) Evaluating accuracy and loss. (b) Testing TPR, FPR, and AUC.

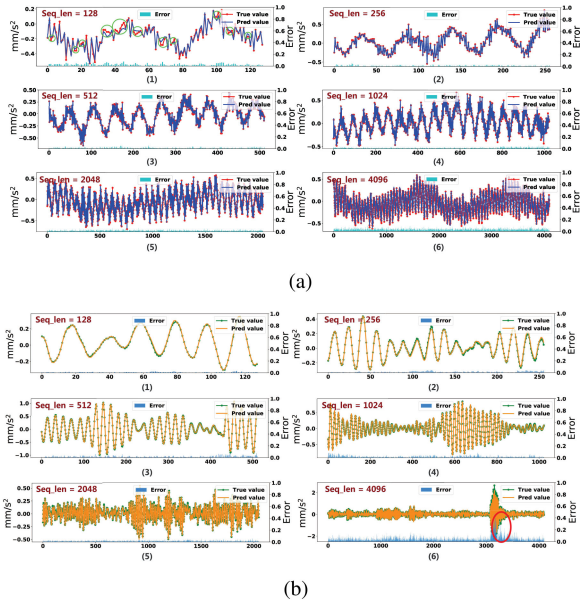
presents the comparison of related works on the CWRU dataset, which contains the state-of-the-art performance based on the structures of recurrent, convolution, and attention, to the best of our knowledge. It can be seen that all of these works achieve a competitive and satisfactory performance with a double-nine level. But actually, Chen *et al.*[16] also needs to employ a continuous wavelet transform to obtain preprocessed representations of the original vibration signal, which may lead to a more complex modeling process. However, the model in [17] and the DH-1 model achieve end-to-end prediction without feature engineering, which greatly simplifies the modeling process and contributes to accelerating the online inference considering the model deployment. It also proves that the self-attention mechanism can be a promising method for the IPdM, which not only guarantees prediction performance, but also makes full use of potential global correlations for an efficient prediction. In addition, Fig. 8 shows the training process of the DH-1 model with sequence lengths of 4096 and 1024. Fig. 8(a) shows the evaluation loss and accuracy on both datasets. Intuitively, the model converges slowly with the AWE dataset, which is mainly caused by the complicated data distribution, so it is difficult for the model to capture the distinguishing information from the original signals. Fig. 8(b) shows that the AUC values reach 0.9956 and 0.9998 on both datasets, respectively. Of note is that there is a slightly higher perception effect for the CWRU dataset.

For the second experiment, the evaluations of the predictive ability of DH-2 for the time-series signals were also completed. Some of the overall comparison metrics for the DH-2 model are listed in Table I. In summary, DH-2 predicts sequence samples for both datasets with volumes of [7499, 4687, 2343, 1171, 585, 292] and [1887, 943, 471, 235, 117, 58]. This means that such a volume of predicted samples can be generated via the second experiment. However, due to the great difference in sample scales,




(a) (b)

Fig. 9. Sequence length versus RMSE and MAE. (a) AWE. (b) CWRU.




(a)

(b)

Fig. 10. Predicted sequences on both datasets. (a) Predicted sequences on the AWE dataset. (b) Predicted sequences on the CWRU dataset.

the objectivity is lost if directly comparing the performance metrics of the same sequence length for both datasets. Therefore, the analysis is separately performed for both datasets. Intuitively, as shown in Fig. 9(a), the data curve of MAE shows a monotonically increasing trend, while the RMSE reaches the optimal result (i.e., the minimum) when the sequence length is 512. Combining both metrics, the proposed DH-2 achieves the best prediction performance when the sequence length of the AWE dataset is equal to 512. As shown in Fig. 9(b), both the MAE and RMSE show monotonically increasing trends on the CWRU dataset. Obviously, the DH-2 obtains the best prediction performance when the sequence length is equal to 128. Based on the above results, we selected six samples with different sequence lengths from the predicted samples of DH-2 and compared them with the original corresponding samples to intuitively demonstrate the performance of sequence prediction. The results are shown in Fig. 10. Overall, the red and green curves represent the real signals from the original datasets, and the blue and yellow curves represent the signals predicted by the DH-2 model, while the bright blue and light blue pillars on the bottom of each subfigure indicate the errors between the predicted signals and real signals. Obviously, the proposed model exhibits different

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

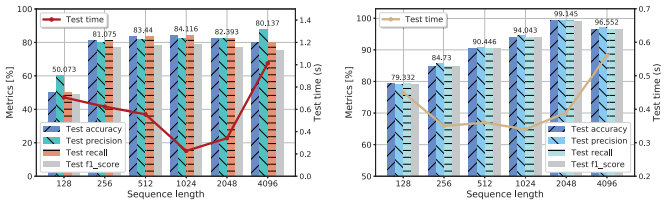



Fig. 11. Metrics of the "perception and prediction" system for both datasets, the number on both bar charts denote the specific values of accuracy. (a) Metrics on the AWE dataset. (b) Metrics on the CWRU dataset.

characteristics on both datasets. However, it is worth noting that the signal sampling frequency of the AWE dataset is set to 4 kHz, while that of the CWRU dataset is set to 48 kHz. The signals with high sampling frequency generally contain more abundant information, which leads to great challenges in learning the potential long-term correlations. As a result, the model can extend the predicted length of signal sequences to 512 on the AWE dataset with a relatively low sampling frequency. Here, the definitions of the short- and long-term are relative, which considers the sampling frequency. For time series, the attention-based model is more concerned with the time correlations than the absolute model. As shown in the green circles of Fig. 10(a-1), the swaying of AWE during its operation is asymmetric, and the data collection is low-frequency sampling. Accordingly, the local areas correspond with the nonmonotonic trend and exhibit excessive jitter, which directly leads to poor prediction effects for the short sequences. Similarly, as shown in the red circles of Fig. 10(b-6), due to the high sampling frequency, some local areas are dense and have large amplitudes, which prevents the DH-2 model from precisely predicting future sequences. In summary, the results of the second experiment demonstrate the predictive ability of DH-2 model for the sensor sequences, which indicates that the proposed DH-2 can generate the sensor sequences of the next moment according to the sensor sequences of the current moment, and it is also the fundamental capability for the completion of instant IPdM.

For the third experiment, the predicted samples from DH-2 (mentioned in the second experiment and several predicted samples are shown in Fig. 10) are fed into the corresponding submodels of DH-1 in terms of the sequence length. Thus, the effectiveness and feasibility of the instant IPdM solution can be evaluated through this experiment, which is also the goal of the entire operation scheme we designed. As shown in Fig. 11(a), the DeepHealth obtains the optimal prediction metrics on both datasets in terms of accuracy, recall, and F1-score, when the sequence length is equal to 1024. However, the precision metric achieves the best when the sequence length is equal to 4096. As shown in Fig. 11(b), the DeepHealth obtains the optimal prediction metrics on both datasets in terms of all four metrics, when the sequence length is equal to 2048. Although a short sequence possesses better sequence prediction effect (as shown in Fig. 9), the identification accuracy of the machine failures with short sequences can also be worse (as shown in Table I).

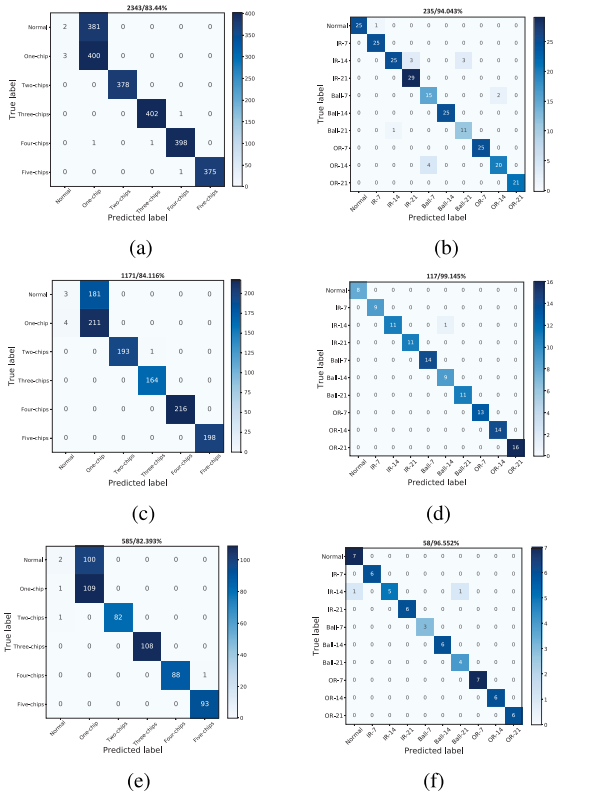
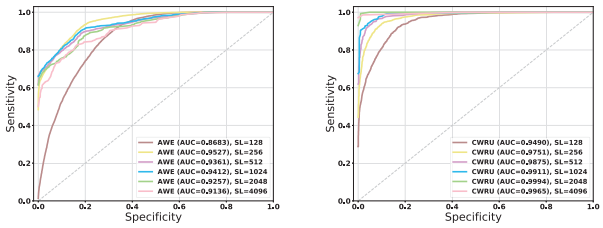



Fig. 12. Confusion matrices on both datasets when the sequence length is equal to [512, 1024, 2048] and [1024, 2048, 4096], respectively. (a) AWE-512. (b) CWRU-1024. (c) AWE-1024. (d) CWRU-2048. (e) AWE-2048. (f) CWRU-4096.

On the one hand, it is mainly caused by the similarity of short sequences in different health conditions. On the other hand, short sequences are easy to be generated, but the corresponding conditions are difficult to be accurately identified due to few valuable information can be provided. Thus, the sequence length should be reasonably chosen according to the specific task (e.g., classification or regression), dataset (e.g., CWRU or AWE), or system requirement (e.g., low-latency transmission demand). In summary, to build a complete "perception and prediction" system, we respectively choose the value of 1024 and 2048 as the sequence length for both datasets to obtain the optimal instant IPdM performance. In addition, Fig. 12 shows the instant IPdM distribution of those predicted samples on both datasets when the sequence length is equal to [512, 1024, 2048] and [1024, 2048, 4096], respectively. Here, the sequence length corresponds to the instant prediction accuracy of the top three. In specific, for each subfigure, the vertical denotes the true labels of these samples, while the horizontal denotes the predicted labels. The diagonal line indicates that the predicted samples of DH-2 are correctly identified by DH-1; otherwise, they are misidentified. Fig. 12(a), (c), and (e) shows that the proposed solution predicts almost all of the "Normal" conditions to be "One-chip" condition. On the contrary, several predicted samples under "One-chip" conditions are also predicted to be "Normal" condition in each case. On the one hand, it is caused by the data similarity between "Normal" and "One-chip" conditions because the angle between

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.




(a) (b)

Fig. 13. TPR, FPR, and AUC metrics on both datasets. (a) AWE. (b) CWRU.

the upper bearing and lower bearing is very small for both conditions such that the well-trained models cannot accurately differentiate both conditions. On the other hand, this is also caused by the accumulated errors of predicted sequences of DH-2, which inevitably leads to more difficulties in the instant IPdM scheme. According to Fig. 12(b), (d), and (f), it can be seen that the proposed solution achieves satisfactory instant IPdM performance on the CWRU dataset. Particularly, the DeepHealth obtains the instant prediction accuracy of 99.145% when the sequence length is equal to 2048 and only one predicted sample is misidentified. Additionally, Fig. 13 shows the instant prediction ability of DeepHealth in the case of imbalanced samples. It can be seen from the ROC curves that the AUC values reach [0.8683, 0.9527, 0.9361, 0.9412, 0.9257, 0.9136] and [0.9490, 0.9751, 0.9875, 0.9911, 0.9994, 0.9965] on both datasets, and the optimal learning performance of DeepHealth is obtained when the sequence length is equal to 256 and 2048, respectively.

Based on the above experiments, the effectiveness of the DeepHealth has been verified. However, due to the heterogeneity among industrial facilities, data characteristics of different facilities exist potential discrepancy, which limits the proposed model to be directly generalized to other facilities. Hence, to improve the generalization ability of AI-empowered IPdM methods is an urgent demand for heterogeneous IIoT scenarios. Meanwhile, to implement the AI-empowered IPdM models in low-latency required IIoT, the prediction latency for massive industrial facilities should also be considered.

## V. Conclusion

In this article, an AI-empowered framework, namely DeepHealth, was proposed and applied to the health perception and sequence prediction of time-series, which was helpful in realizing the IIoT-enabled instant IPdM and enables making maintenance decisions in advance to avoid machine failures. Particularly, the proposed DeepHealth makes use of a self-attention mechanism to capture the global dependencies in high-frequency vibration signals. Additionally, customized destructive experiments on a dual-bearing rotating machine were conducted, and finally constructed the AWE industrial dataset. Sufficient experiments demonstrate the effectiveness of the proposed models for the health perception of the current and future moment. For our future work, we will focus on improving the generalization ability of AI-empowered IPdM models, and investigate the orchestration of IPdM models to ensure an efficient operation of industrial systems.

## References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[2] K. Wang, "Intelligent predictive maintenance (IPdM) system–Industry 4.0 scenario," *WIT Trans. Eng. Sci.*, vol. 113, pp. 259–268, 2016.
[3] W. Zhang, D. Yang, and H. Wang, "Data-driven methods for predictive maintenance of industrial equipment: A survey," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2213–2227, Sep. 2019.
[4] D. Liu, W. Cheng, and W. Wen, "Generalized demodulation with tunable E-Factor for rolling bearing diagnosis under time-varying rotational speed," *J. Sound Vib.*, vol. 430, pp. 59–74, Sep. 2018.
[5] F. Cheng, L. Qu, and W. Qiao, "Fault prognosis and remaining useful life prediction of wind turbine gearboxes using current signal analysis," *IEEE Trans. Sustain. Energy*, vol. 9, no. 1, pp. 157–167, Jan. 2018.
[6] M. Xia, T. Li, L. Xu, L. Liu, and C. W. De Silva, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatron.*, vol. 23, no. 1, pp. 101–110, Feb. 2018.
[7] M. Zhao, M. Kang, B. Tang, and M. Pecht, "Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4290–4300, May 2018.
[8] L. Saidi, J. B. Ali, and F. Fnaiech, "Application of higher order spectral features and support vector machines for bearing faults classification," *ISA Trans.*, vol. 54, pp. 193–206, Jan. 2015.
[9] I. Martin-Diaz, D. Morinigo-Sotelo, O. Duque-Perez, R. A. Osornio-Rios, and R. J. Romero-Troncoso, "Hybrid algorithmic approach oriented to incipient rotor fault diagnosis on induction motors," *ISA Trans.*, vol. 80, pp. 427–438, Sep. 2018.
[10] L. Wang, Z. Zhang, H. Long, J. Xu, and R. Liu, "Wind turbine gearbox failure identification with deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1360–1368, Jun. 2016.
[11] Y. Qin, X. Wang, and J. Zou, "The optimized deep belief networks with improved logistic Sigmoid units and their application in fault diagnosis for planetary gearboxes of wind turbines," *IEEE Trans. Ind. Electron.*, vol. 66, no. 5, pp. 3814–3824, May 2019.
[12] H. Shao, H. Jiang, Y. Lin, and X. Li, "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders," *Mech. Syst. Signal Proc.*, vol. 102, pp. 278–297, Mar. 2018.
[13] Y. Cheng, H. Zhu, J. Wu, and X. Shao, "Machine health monitoring using adaptive kernel spectral clustering and deep long short-term memory recurrent neural networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 987–997, Feb. 2019.
[14] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.
[15] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
[16] Z. Chen, K. Gryllias, and W. Li, "Mechanical fault diagnosis using convolutional neural networks and extreme learning machine," *Mech. Syst. Signal Proc.*, vol. 133, Nov. 2019, Art no. 106272.
[17] X. Li, W. Zhang, and Q. Ding, "Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism," *Signal Process.*, vol. 161, pp. 136–154, Aug. 2019.
[18] Q. Guo, Y. Li, Y. Song, D. Wang, and W. Chen, "Intelligent fault diagnosis method based on full 1D convolutional generative adversarial network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2044–2053, Mar. 2019.
[19] H. Liu, J. Zhou, Y. Zheng, W. Jiang, and Y. Zhang, "Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders," *ISA Trans.*, vol. 77, pp. 167–178, Jun. 2018.
[20] J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial Artificial Intelligence for industry 4.0-based manufacturing systems," *Manuf. Lett.*, vol. 18, pp. 20–23, Oct. 2018.
[21] Case Western Reserve University. Accessed on: 2018. [Online]. Available: https://csegroups.case.edu/bearingdatacenter/pages/download-data-file
[22] H. Qiu, J. Lee, J. Lin, and G. Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *J. Sound Vib.*, vol. 289, no. 4/5, pp. 1066–1090, Feb. 2006.

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.

[23] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *Proc. Int. Conf. Prognostics Health Manage.*, 2008, pp. 1–9.
[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR 2015)*, 2015, pp. 1–15.
[25] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "Geoman: Multi-level attention networks for geo-sensory time series prediction," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, 2018, pp. 3428–3434.
[26] Y. Qin *et al.*, "Hybrid forecasting model based on long short term memory network and deep learning neural network for wind signal," *Appl. Energy*, vol. 236, pp. 262–272, Feb. 2019.
[27] D. Yang *et al.*, "Assignment of segmented slots enabling reliable real-time transmission in industrial wireless sensor networks," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3966–3977, Jun. 2015.
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[30] G. Tang, M. Müller, G. A. Rios, and R. Sennrich, "Why self-attention? A targeted evaluation of neural machine translation architectures," in *Proc. 2018 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 4263–4272.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[32] S. Shao, P. Wang, and R. Yan, "Generative adversarial networks for data augmentation in machine fault diagnosis," *Comput. Ind.*, vol. 106, pp. 85–93, Apr. 2019.
[33] W. Zhang, D. Yang, H. Wang, X. Huang, and M. Gidlund, "CarNet: A dual correlation method for health perception of rotating machinery," *IEEE Sensors J.*, vol. 19, no. 16, pp. 7095–7106, Aug. 2019.

**Weiting Zhang** (Student Member, IEEE) is currently working toward the Ph.D. degree in communication and information systems with the Beijing Jiaotong University, Beijing, China.

His research interests include intelligent predictive maintenance, mobile edge computing, and application of artificial intelligence for industrial Internet of Things.

**Dong Yang** (Member, IEEE) received the B.S. degree in computer science and technology from Central South University, Hunan, China, in 2003, and the Ph.D. degree in communications and information science from Beijing Jiaotong University, Beijing, China, in 2009 .

From 2009 to 2010, he was a Postdoctoral Research Associate with Jonkoping University, Jonkoping, Sweden. In 2010, he joined the School of Electronic and Information Engineering, Beijing Jiaotong University. His research interests include network architecture, industrial network, and Internet of Things.

**Youzhi Xu** received the Graduate degree from Xian Jiaotong University, Xian, China, in 1968, the academic degrees of the Tech. Licentiate, the Ph.D. degree, and the Docent degree from Linkoping University, Sundsvall, Sweden, in 1989, 1991, and 2000, respectively, all in electrical engineering.

His research interests are in the areas of video compression, source coding, channel coding, and communication networks.

**Xuefeng Huang** received the B.S. degree in engineering from Wuhan University of Technology, Wuhan, China, in 1998.

He is currently the President of the Rail Transit Research Institute Sheenline. He has been engaged in the research of rail transit inspection and maintenance for more than 20 years. He holds more than 20 patents (granted and pending applications) in the area of rail transit.

**Jun Zhang** received the Graduate degree in automation from the Changchun Institute of Technology, Changchun, China, in 2006.

He joined Beijing Sheenline Group Company, Ltd., in 2007, and engaged in the development of automation equipment and automatic production line. From September 2014, he began to research factory information system and big data application system of equipment as well as equipment asset and health management system.

**Mikael Gidlund** received the M.Sc. and Ph.D. degrees in electrical engineering from Mid Sweden University, Sundsvall, Sweden, in 2000 and 2005, respectively.

He has pioneered the area of industrial wireless sensor network and he holds more than 20 patents (granted and pending applications) in the area of wireless communications. He has authored or coauthored more than 200 scientific publications in refereed fora. His current research interests include wireless communication and networks, access protocols, and security.

Authorized licensed use limited to: Queens University Belfast. Downloaded on August 11,2024 at 05:49:03 UTC from IEEE Xplore. Restrictions apply.