Received 4 August 2022; accepted 26 August 2022. Date of publication 12 September 2022; date of current version 29 September 2022.
The review of this article was arranged by Associate Editor Min Xia.

Digital Object Identifier 10.1109/OJIM.2022.3205649

A Distributional Perspective on Remaining Useful Life Prediction With Deep Learning and Quantile Regression

MING ZHANG[1] (Member, IEEE), DUO WANG[2], NASSER AMAITIK[1], AND YUCHUN XU[1]

[1]College of Engineering and Physical Sciences, Aston University, B4 7ET Birmingham, U.K.

[2]Department of Automation, Tsinghua University, Beijing 100084, China

CORRESPONDING AUTHORS: M. ZHANG AND Y. XU (e-mail: m.zhang21@aston.ac.uk; y.xu16@aston.ac.uk)

This work was supported by the European Commission Horizon 2020 Research and Innovation Programme through the RECLAIM Project, named as "Remanufacturing and Refurbishment Large Industrial Equipment," under Grant 869884.

ABSTRACT With the rapid development of information and sensor technology, the data-driven remaining useful lifetime (RUL) prediction methods have been acquired a successful development. Nowadays, the data-driven RUL methods are focused on estimating the RUL value. However, it is more important to quantify the uncertainty associated with the RUL value. This is because increasingly complex industrial systems would arise various sources of uncertainty. This article proposes a novel distributional RUL prediction method, which aims at quantifying the RUL uncertainty by identifying the confidence interval with the cumulative distribution function (CDF). The proposed learning method has been built based on quantile regression and implemented from a distributional perspective under the deep neural network framework. The results of the run-to-failure degradation experiments of rolling bearing demonstrate the effectiveness and good performance of the proposed method compared to other state-of-the-art methods. The visualization results obtained by *t-SNE* technology have been investigated to further verify the effectiveness and generalization ability of the proposed method.

INDEX TERMS Deep learning, distributional remaining useful lifetime (RUL) prediction, quantile regression, rolling bearing, uncertainty.

I. INTRODUCTION

PROGNOSTICS and health management (PHM) techniques play a vital role in the condition-based maintenance of large industrial equipment, which could prevent unexpected failure and reduce downtime to achieve the purposes of saving the maintenance cost, maximizing the working time, safety, and reliability [1], [2]. The remaining useful lifetime (RUL) prediction is generally known as estimating the time before the machine completely fails and is used to support PHM in producing reasonable maintenance plans and strategies [3], [4], [5], [6]. Therefore, it is imperative that the RUL prediction method provides a precise estimation. However, the prediction of future conditions and precise RUL are significantly challenging from the massive data obtained from the operating systems due to the increasing complexity of industrial equipment.

As one of the advanced RUL methods, the model-based methods have proved their effectiveness in many studies [7], [8], [9]. They rely on accurate physical laws and knowledge of the degradation process for the machine system. However, it is very difficult to build an accurate physical degradation model in many industrial applications. The performance, robustness, and generalization of the model-based prediction model would be significantly decreased because of the more and more complexity of the modern industry [2], [4].

Meanwhile, the current real-world industrial system has been turned into a data-rich environment with the development of the Internet of Things (IoT). The data-driven methods of RUL prediction have been widely investigated and applied based on massive monitoring data, for instance, vibration, current, temperature, etc. As one of the powerful data-driven approaches, deep learning has been proposed to develop RUL prediction algorithm in recent years and achieve great results, for instance, deep neural network (DNN) [10], deep belief network (DBN) [11], [12], convolutional neural network (CNN) [13], [14], [15], [16], long short-term memory (LSTM) [17], [18], [19], [20], and their combinations [21], [22]. Moreover, some special deep models, such as the deep adversarial neural networks [23] and normalizing flow-embedded sequence-to-sequence model [24], have been developed for estimating RUL. Differently from the model-based approaches, the degradation models within these data-driven methods are independent of the prior knowledge and they could be learned from the available data. Therefore, they are easier to use and capable of handling the large industrial machine prediction problem, whose degradation model is too complex to establish.

Nevertheless, these methods focus on estimating the RUL value without considering the various types of uncertainty inherent in the degradation process. These uncertainties need to be carefully considered in order to arrive at reliable and accurate predictions of the RUL values. This is always the case for RUL prediction in industrial applications [25]. Therefore, quantifying the uncertainty of RUL would be as important as estimating the RUL value. Moreover, predicting RUL with confidence interval could support human-expert making maintenance decisions more comprehensively. Recently, there are limited data-driven approaches considering the uncertainty in the RUL prediction process. Wang *et al.* [26] proposed to use variational inference for quantifying the uncertainty of RUL prediction after training the recurrent CNN. Zhao *et al.* [27] built a probabilistic RUL prediction framework and used the probability density function (PDF) as the quantified uncertainty. Pang *et al.* [28] proposed a Bayesian inference model for updating the posterior distributions of model parameters and calculating the confidence interval for uncertainty.

However, the existing approaches cannot directly quantify the uncertainty of RUL, and they still need to implement further processing in order to identify the uncertainty interval. In traditional data-driven RUL learning algorithms, the primary focus is to train the model based on the RUL value as the label, but it discards the label information of the RUL uncertainty. The main motivation of this article is to introduce a new data-driven method that takes full advantage of utilizing RUL uncertain information to support the learning procedure. Inspired by the idea that there are more benefits to learning an approximate distribution rather than an approximate value [29], we propose a new distributional remaining useful life prediction method with a DNN and quantile regression. The proposed method could directly output the cumulative distribution function (CDF) and calculate the confidence interval for estimating the uncertainty of RUL. The main contributions are summarized as follows.

1) The distributional learning method has been proposed for distributional RUL prediction and implemented by quantile regression optimization. The quantile regression loss has been deduced following the theory of the Wasserstein metric, which is designed to calculate the divergence of inverse CDF between parameterized and target distribution.
2) The proposed distributional RUL prediction method has been implemented by using the deep learning framework with the learning method from a distributional perspective. With the support of the quantile distribution and Dirac delta function, it could directly quantify uncertainty and calculate the confidence interval.
3) The novel quantile Huber loss (QH-loss) function has been designed and utilized to optimize the proposed deep model of distributional RUL prediction by the stochastic gradient descent (SGD) method. It combines the advantages of quantile regression loss and Huber loss. The comparison experiment demonstrates that QH-loss is outperforming the typical MSE and mean absolute error (MAE).
4) The effectiveness and performance of the proposed method have been verified using the run-to-failure degradation experiment of rolling bearings under different working conditions. The visualization results demonstrate the high generalization ability for the proposed method with different feature mapping.

II. THEORETICAL FRAMEWORK OF PROPOSED METHOD

First, we define the new distributional RUL prediction problem. To solve this problem from the data-driven perspective, we deduce the learning method from a distributional perspective according to quantile regression optimization. The Wasserstein metric is the theoretical basis for constructing the quantile regression.

A. DEFINITION OF DISTRIBUTIONAL RUL PREDICTION

For the classical RUL prediction, the theoretical formula is

$$Y = f(X|\theta) \tag{1}$$

where Y denotes the ground-truth RUL value, X denotes the observation, and f denotes the prediction function with the parameter θ. The output of (1) is the RUL value. However, we expect to output the quantified uncertainty in this work. Therefore, we redefine a new distributional RUL prediction, which could directly output RUL uncertainty as a distribution function. The specific definition is to replace the RUL value Y with a certain distribution of Z whose expectation is the value Y

$$Y := \mathbb{E}[Z(X)] = \mathbb{E}[F(X|\theta)]. \tag{2}$$

This equation also defines that such distribution Z could be characterized by the conditional distribution function F with parameter θ.

B. WASSERSTEIN METRIC

The Wasserstein metric is used as the evaluation method for two different distributions, which has the mathematic property of continuous and differentiable almost everywhere [30], [31], [32]. It is the basic theory for deducing the quantile regression loss function. Müller [33] took it as the metric of L^p on inverse CDF (inverse CDF). Therefore, the Wasserstein metric W_p can be defined as

$$W_p(Y,U) = \left(\int_0^1 |F_Y^{-1}(q) - F_U^{-1}(q)|^p d\omega \right)^{1/p} \tag{3}$$

where for a random variable Y, the inverse CDF F_Y^{-1} of q is expressed as follows:

$$F_Y^{-1}(q) := \inf\{y \in \mathbb{R} : q \le F_Y(y)\} \tag{4}$$

where $F_Y(y) = \Pr(Y \le y)$ is the CDF of random variable Y.

C. LEARNING FROM DISTRIBUTIONAL PERSPECTIVE

The proposed method is going to predict the quantiles of the target distribution, where $q_i = 1/N$, for $i = 1, \ldots, N$. So, it is called a quantile distribution $\mathcal{Z}_Q$, which has a fixed N. The discrete values derived from its CDF are $\tau_1 \ldots, \tau_N$, where $\tau_i = (i/N)$ for $i = 1, \ldots, N$, these also represent the cumulative probabilities with a certain distribution. Then, the quantile distribution $Z_\theta \in \mathcal{Z}_Q$ projects the observation x to a probability distribution supported by the parameterized model $\theta_i(x)$, which is defined as

$$Z_\theta(x) := \sum_{i=1}^{N} \delta_{\theta_i(x)} \tag{5}$$

where δ_z is a Dirac delta function at $z \in \mathbb{R}$. This reformulation allows us to learn the distributional prediction model by using the Wasserstein metric and implement it by quantile regression [34].

1) QUANTILE PROJECTION

The distributional learning is projected to a parameterized quantile distribution optimization, it quantifies the projection of a random distribution $Z \in \mathcal{Z}$ into $\mathcal{Z}_Q$, which is expressed as

$$\underset{Z_\theta \in \mathcal{Z}_Q}{\arg\min}\ W_1(Z, Z_\theta). \tag{6}$$

Assume that Y is the bounded target distribution and U is a quantile distribution based on the Dirac delta function, shown in (5), with the support $z_1, \ldots, z_N$. Accordingly, the Wasserstein metric is

$$W_1(Y,U) = \sum_{i=1}^{N} \int_{\tau_{i-1}}^{\tau_i} |F_Y^{-1}(\omega) - z_i| d\omega. \tag{7}$$

When $\tau_{i-1}, \tau_i \in [0,1]$ with $\tau_{i-1} < \tau_i$, if F^{-1} is the inverse CDF, then $F^{-1}((\tau_{i-1} + \tau_i)/2)$ is always a valid value; meanwhile, if F^{-1} is continuous at $(\tau_{i-1} + \tau_i)/2$, then $F^{-1}((\tau_{i-1} + \tau_i)/2)$ is the unique value. Therefore, we use the quantile midpoints which is $\hat{\tau}_i = (\tau_{i-1} + \tau_i)/2$ where $1 \le i \le N$, then minimizing $W_1(Y,U)$ is obtained by $z_i = F_Y^{-1}(\hat{\tau}_i)$

2) QUANTILE REGRESSION

Quantile regression or conditional quantile regression could approximate the quantile function of distribution or conditional distributions, which is an effective method to solve the distributional learning problem [35]. The parameterized distributional model would be trained by minimizing the quantile regression loss, which is defined as

$$L_{QR}^{\tau} := \mathbb{E}_{\hat{Z} \sim Z}\left[\rho_\tau\left(\hat{Z} - z\right)\right], \text{where}$$
$$\rho_\tau(u) = u\left(\tau - \delta_{\{u<0\}}\right)\ \forall u \in \mathbb{R}. \tag{8}$$

$z : \{z_1, z_2, \ldots, z_N\}$ denotes the values of the quantile function $F_Z^{-1}(\tau)$, where $\tau \in [0, 1]$. This objective function is convex and asymmetric, which could control overestimation errors with weight coefficient τ and underestimation errors with $1 - \tau$ during the training procedure [34].

III. IMPLEMENTATION OF PROPOSED METHOD WITH DEEP LEARNING FRAMEWORK

In this section, we implement the distributional RUL prediction method according to the theoretical framework of distributional learning which is proposed in Section II. The overall architecture of our proposed distributional RUL prediction model is shown in Fig. 1. The data-driven RUL prediction framework is generally divided into three steps after acquiring the data: 1) constructing the labels based on the health indicators; 2) training the prediction model of RUL with the labeled data; and 3) testing the performance of the optimized model by using the unseen data.

In the first step, the root mean square (RMS) feature is selected as the health indicator, and then the first prediction time (FPT) is determined. The labeling for the normal period is constant, and the linearly decreasing function is built for labeling the degradation period. The training step uses the data from the degradation period of vibration and their labels. The DNN is used as the estimating function to directly output the quantile distribution. The new QH-loss has been proposed to optimize the parameter of the DNN. After the training process, the total new unseen test data, including the normal and degradation period of vibration data are directly put into the deep model. As the outputs are the quantile distributions, we can directly obtain the RUL confidence interval for every time period.

A. FPT DETERMINATION AND RUL LABELING METHOD

In order to construct the labels of bearing data, the FPT should be first determined. In this work, we follow the simple and effective way proposed in [16] to calculate the FPT. The mean μ and standard deviation δ are calculated from the early normal period of the RMS of each vibration sample. The FPT is confirmed when the feature RMS value $f_t - \mu$ successive outside the 3δ, which can be expressed as follows:

$$|f_{t+i} - \mu| > 3\delta \tag{9}$$

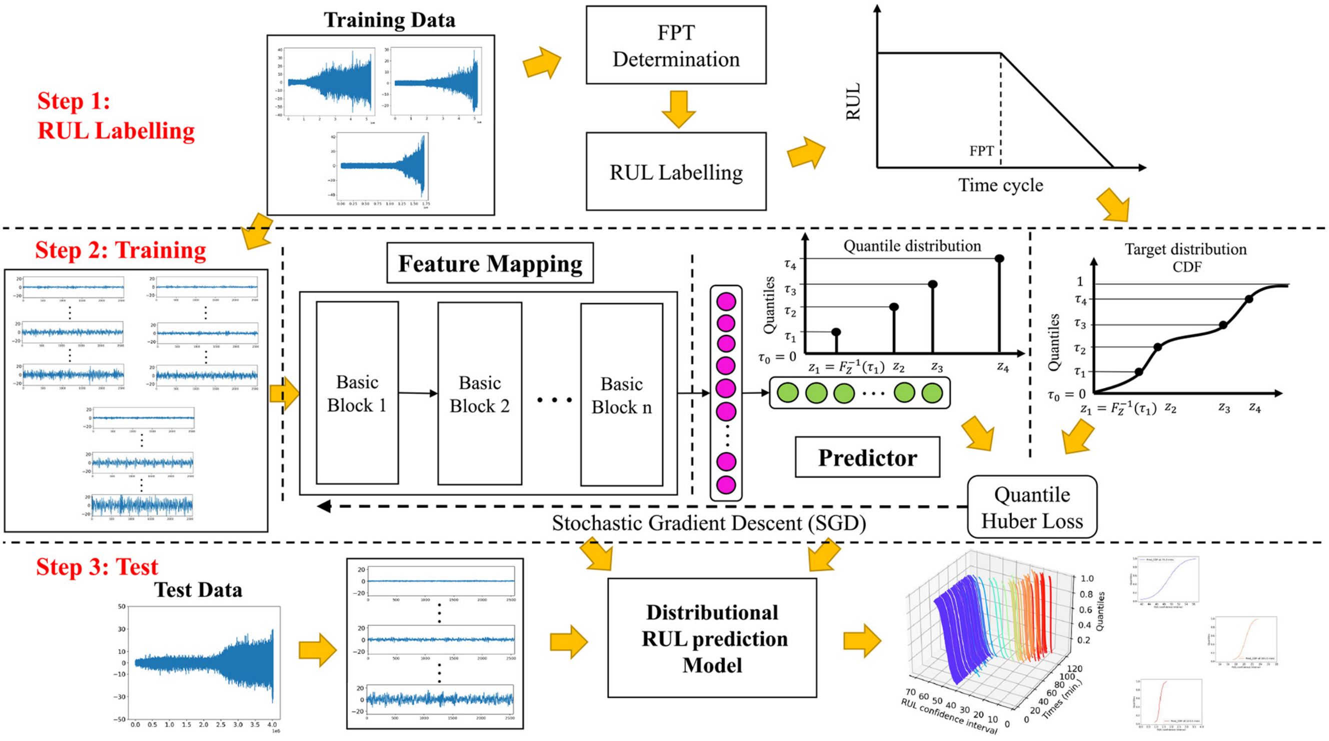


FIGURE 1. Architecture of distributional RUL prediction.

where $i = 0, 1, 2$ and only if all these sequence features f_{t+i} satisfied (9), the current time t would be set as FPT. As shown in Fig. 2(a) and (b), the FTP is determined to be 74 based on the proposed method for the original vibration signal of the overall period of the lifecycle.

Once the FPT has been decided, the ground-truth RUL label for the overall life cycle could be shaped in a segmental linear function, which is shown in Fig. 2(c). It consists of two periods: 1) the normal period is a constant and 2) the degradation period is a linearly decreasing function, which represents the life percentage of a machine. As one of the RUL labeling methods, it is a simple and effective way and has been proposed in many studies [13], [16], [36], [37]. For this kind of labeling method, only the labeled degradation date is used for training the prediction model. It should be noticed that the FPT could affect the performance of the prediction model. Therefore, comparing experiment has been carried out in Section IV-D for demonstrating the effects.

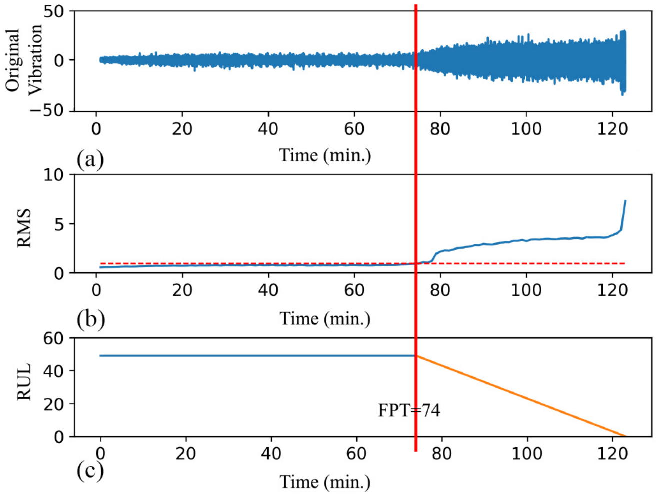


FIGURE 2. FPT determination and ground-truth RUL labeling with the proposed method. (a) Original vibration; (b) RMS; and (c) RUL.

B. PROPOSED NETWORK STRUCTURE

After obtaining the ground true RUL labels of training data, the DNN has been proposed as the foundation of the distributional RUL prediction model. As shown in Fig. 1, the labeled data is used for training the proposed model, then the model would be directly tested by the unseen data. There are two parts in the network structure, including featuring mapping with parameter θ_{FM} and predictor with parameter θ_{PD}. The feature mapping is designed to extract the essential features of original sequence data. These features are learned during the training process and presented in high-level representation in each layer. The basic block is the convolution layer [38] or its advanced variants [39], [40]. The performance and effects of different kinds of basic blocks have been compared and analyzed in Section IV-E. Meanwhile, the predictor is designed to output the RUL distribution directly, and it consists of two linear layers and the ReLU activation function. Different from the typical RUL structure based on a neural network taking one node in the final layer, we propose to take N nodes in the final layer,

Algorithm 1 Learning Procedure

```
Require: source data $X$, mini-batch size $m$, training steps $n$, learning rate $\alpha$.
Initialize the parameters $\theta_{FM}$ and $\theta_{PD}$.
for $epo = 1, \ldots, n$ do
    Sample mini-batch $\{x_i, y_i\}_{i=1}^{m}$ from $X$
    Build target distribution $F_Z(z|y, \sigma)$ based on $y$
    Calculate the inverse CDF $F_Z^{-1}(\hat{\tau}_i)$
    $\theta_{FM}, \theta_{PD} \leftarrow \theta_{FM}, \theta_{PD} - \alpha \nabla L_{QH}$
end for
```

which represent the N quantile midpoints $\hat{\tau}_i$ of the target CDF, where $1 \leq i \leq N$.

C. LEARNING PROCEDURE

During the training procedure, the SGD is used to optimize the parameters θ_{FM} and θ_{PD} in the proposed DNN structure. In this work, a new QH-loss is proposed and built based on the quantile regression loss of (8) and typical Huber loss [41]. The Huber loss is defined as

$$L_\kappa(u) = \begin{cases} \frac{1}{2}u^2, & \text{if}|u| \leq \kappa \\ \kappa\left(|u| - \frac{1}{2}\kappa\right), & \text{otherwise} \end{cases} \tag{10}$$

then the QH-loss has been designed as the combination of the Huber loss and quantile regression loss, that is

$$\rho_\tau^\kappa(u) = |\tau - \delta_{u<0}| L_\kappa(u). \tag{11}$$

However, the ground-truth RUL label y is a value, but a distribution function should be the target distribution for calculating the QH-loss. As a result, we design the CDF of target distribution $F_Z(z|y)$ based on y by using the Gaussian distribution function, which is

$$F_Z(z|y, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{(t-y)^2}{2\sigma^2}\right) dt. \tag{12}$$

Finally, QH-loss L_{QH} for the proposed distributional RUL prediction network can be calculated by

$$L_{QH} = \sum_{i=1}^{N} \mathbb{E}\left[\rho_{\hat{\tau}_i}^{\kappa}\left(F_Z^{-1}(\hat{\tau}_i) - P\left(M\left(\hat{\tau}_i|x\right)\right)\right)\right] \tag{13}$$

where M denotes the feature mapping, P denotes the predictor, and then $P(M(\hat{\tau}_i|x))$ denotes the output of quantile distribution at every quantile $\hat{\tau}_i$. The learning procedure of our distributional RUL prediction model has been summarized in Algorithm .

D. MEASURE INDICATORS

In this article, the prediction performance of the RUL model is quantitatively evaluated using the MAE, RMS error (RMSE), and R^2 score, which are

$$\text{MAE} = \frac{1}{Q}\sum_{i=1}^{Q} |\hat{y}_i - y_i| \tag{14}$$

$$\text{RMSE} = \sqrt{\frac{1}{Q}\sum_{i=1}^{Q} (\hat{y}_i - y_i)^2} \tag{15}$$

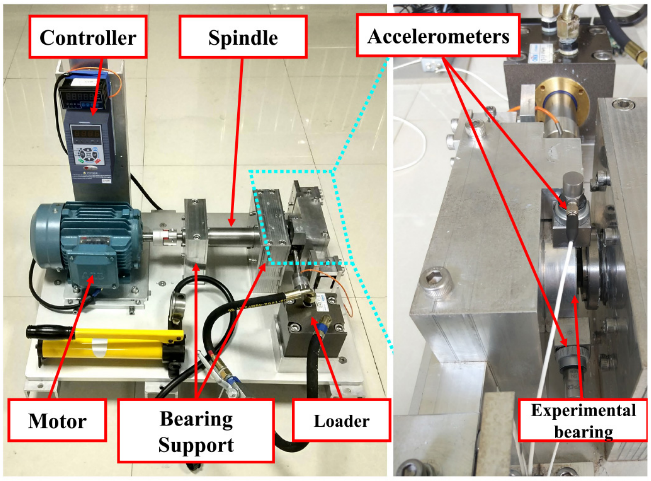


FIGURE 3. XJTU testbed of bearing degradation experiment.

$$R^2 = 1 - \frac{\sum_{i=1}^{Q}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{Q}(\bar{y} - y_i)^2} \tag{16}$$

where Q is the number of testing samples, $\hat{y}_i$ denotes the predicted RUL value, y_i denotes the true value of RUL (label), and $\bar{y}$ denotes the mean of all the true RUL value.

IV. EXPERIMENTS

A. DATA SET DESCRIPTION

The data set was acquired from the testbed built by the Xi'an Jiaotong University (XJTU), which is designed for the degradation experiment of rolling bearing [42]. As shown in Fig. 3, the testbed is composed of a controller for motor speed and loader force and a test bench body with motor, spindle, support, experimental bearing, and loader. The test bearing is installed on the outside of the test bench and the loader is directly forced on the outer race of the bearing. The vibration signal across the whole life cycle of the test bearing is collected by two accelerometers (PCB 352C33) which are located on the house of the bearing at 12 and 9 o'clock positions. The experiments have been conducted under three different conditions, the bearing type is LDK UER204 ball bearing. The detailed pieces of information of all experiments and their actual lifetimes are summarized in Table 1. The vibration data of each run-to-failure bearing experiment has been recorded every minute, the sampling frequency is 25.6 kHz and the length of each sample is equal to 32 768 (approximate 1.28 s). Three bearings of the overall life cycle vibration signal for three different conditions are shown in Fig. 4.

B. DATA PREPROCESSING

1) NORMALIZATION

As shown in Table 1, the experimental actual lifetimes indicate that there are significant variations even under the same working conditions. To reduce a certain amount of difference, we first use the $z-score$ method to normalize the overall

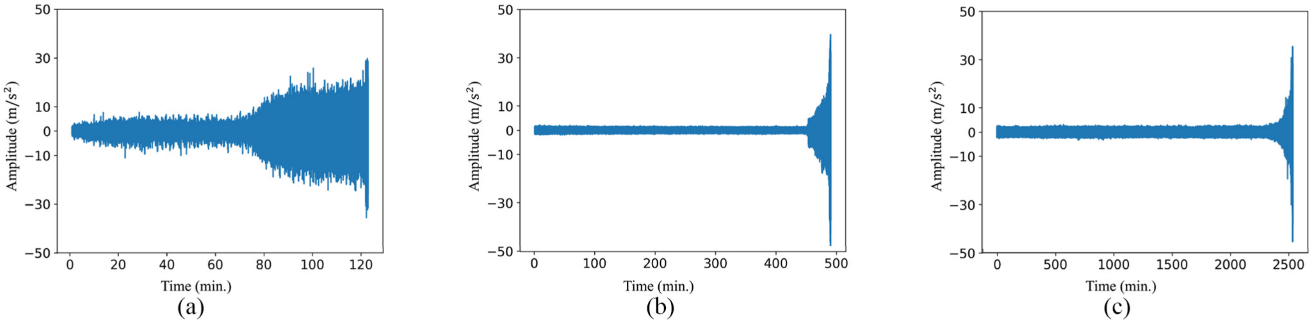


FIGURE 4. Original vibration signal of three run-to-failure bearing experiments during the overall life cycle. (a) Bearing-1 of Condition-1. (b) Bearing-1 of Condition-2. (c) Bearing-1 of Condition-3.

TABLE 1. Detailed information of the data set.

Condition	Bearing	Actual lifetime	Force	Speed
1	1	123 $min.$	12 kN	2100 rpm
	2	161 $min.$		
	3	158 $min.$		
	4	52 $min.$		
2	1	491 $min.$	11 kN	2250 rpm
	2	161 $min.$		
	3	533 $min.$		
	4	42 $min.$		
3	1	2496 $min.$	10 kN	2400 rpm
	2	371 $min.$		
	3	1515 $min.$		
	4	114 $min.$		

life cycle vibration signal for each bearing experiment, the normalization equation is defined as follows:

$$\hat{x}_i^t = \frac{x_i^t - \mu_i}{\delta_i} \tag{17}$$

where $\hat{x}_i^t$ and x_i^t are the normalized vibration signal and original vibration signal at time t of the ith set of experiment, respectively; μ_i and δ_i are the mean value and standard deviation of i^{th} set of experiment, respectively. Furthermore, it is obvious that we could not directly use the actual lifetime values as the target labels because the huge gap between the actual lifetimes varies from 42 to 2496 min. Therefore, we normalize the actual lifetime at every collection point in the range of [1, 0] by the following equation:

$$y_i^t = \begin{cases} 1, & \text{if } t \leq \text{FTP}_i \\ \frac{\text{ActLife}_i - t}{\text{RUL}_i}, & \text{if } \text{FTP}_i < t \leq \text{ActLife}_i \end{cases} \tag{18}$$

where y_i^t denotes the normalized ground-truth RUL value; and ActLife_i is the actual lifetime of i^{th} set of experiment. In this work, we use the normalized y_i^t as the label for training the proposed model. Then, the predicted RUL value PredRUL_i^t can be calculated by

$$\text{PredRUL}_i^t = \hat{y}_i^t \times \text{RUL}_i \tag{19}$$

where $\hat{y}_i^t$ is the predicted normalized RUL value from the deep model.

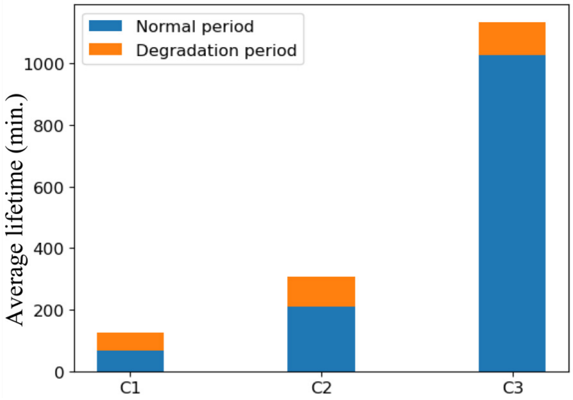


FIGURE 5. Actual lifetime statistic analysis.

TABLE 2. FPTs and RULs.

Condition		Bearing (min.)			
		1	2	3	4
1	FPT	74	55	108	34
	RUL	49	106	50	18
2	FPT	452	46	314	30
	RUL	39	115	219	12
3	FPT	2348	341	1417	5
	RUL	190	30	98	109

2) FPT DETERMINATION

Based on the proposed FPT determination method, we calculate the FPT for each bearing of three different conditions. The RUL is calculated by using the actual lifetime minus FPT. The results for all experiment scenarios are shown in Table 2, and the unit is minutes. Moreover, the average time of normal and degradation period for three conditions has been computed, the results are displayed in Fig. 5, which indicate the average lifetime of the experimental bearing will be significantly influenced by the normal period by the varying force and speed.

C. IMPLEMENTATION DETAILS

The proposed DNN structure is shown in Table 3, which consists of feature mapping and predictor. In the predictor, there are two layers, the first one is the linear layers with 128

TABLE 3. Details of network structure with CNN-base as feature mapping.

	Block	Layer type	Kernel	Stride	Channels	
					in	out
Feature Mapping	1	Convolution	64×1	16×1	1	16
		ReLU				
		Max Pooling	2×1	2×1		
	2	Convolution	3×1	1×1	16	32
		ReLU				
		Max Pooling	2×1	2×1		
	3	Convolution	3×1	1×1	32	64
		ReLU				
		Max Pooling	2×1	2×1		
	4	Convolution	3×1	1×1	64	64
		ReLU				
		Max Pooling	2×1	2×1		
	5	Convolution	3×1	1×1	64	64
		ReLU				
Predictor	Fully Connected	Flatten				
		Linear			576	128
		ReLU				
	Q-distribution	Linear			128	N_τ

TABLE 4. Default hyperparameters of the proposed method.

Hyperparameters	Value
Learning rate	0.001
Batch size	32
Maximum epochs	1000
Hidden size	128
N_τ	33
QH-loss factor κ	1.0
Target distribution factor σ	0.2

nodes following the ReLU function and another linear layer with N_τ nodes is the direct output of the quantile distribution of the Dirac delta function. The CNN-base is selected as the feature mapping, which has five convolution blocks proposed in [43]. This structure is used in the following experiment section for demonstrating the proposed distributional RUL prediction. Other kinds of feature mappings are included in Tables 7–9 of the Appendix and the state-of-the-art methods have been tested and analyzed in Section IV-E. In the present work, each sample has 2560 points drawn from the original vibration signal at each minute, which has been considered as the input of the proposed model. The sampling method of the vibration signal is following [44] to resegment each original vibration with certain overlapping. Meanwhile, only the labeled date of the degradation period is used for training, but the whole period of unseen data would be tested and analyzed.

All the experiments are tested on a computer with one Nvidia GeForce GTX 2060 GPU, one Intel Core i7-10750 H CPU of 2.60 GHz, and 16 GB of memory. The default values of hyperparameters are given in Table 4.

TABLE 5. Experiment scenarios for testing distribution RUL prediction.

Condition	Training Data	Test Data
1	Bearing-2 Bearing-3 Bearing-4	Bearing-1
2	Bearing-2 Bearing-3 Bearing-4	Bearing-1
3	Bearing-2 Bearing-3 Bearing-4	Bearing-1

D. DISTRIBUTIONAL RUL PREDICTION USING PROPOSED METHOD

In order to demonstrate the performance of the proposed method, three experiment scenarios have been decided to test the distribution RUL prediction model first. As shown in Table 5, Bearing-1 of each condition is selected as the unseen data for testing, while the rest of the Bearings 2, 3, and 4 are set to the training data.

1) RESULTS AND ANALYSIS

The results of prediction RUL distribution are shown in Fig. 6, and the quantified indicators of three scenarios can be found in Table 6. In Fig. 6, the color of each time step is determined by the output of the proposed model. The blue represents the normal condition of the bearing, while the red means the bearing is close to the end of its lifetime. The visualization results of distributional RUL indicate that the proposed method can predict not only the RUL value but also the uncertainty at each time step with high performance. All vibration data in the normal period are predicted with a high RUL confidence interval, especially for the second and third scenarios, which is shown in Fig. 6(b) and (c). For the degradation period of the unseen testing data, the learned deep model can directly identify the decreasing process without any support from data and its label within the testing scenarios. The prediction of the first scenarios shown in Fig. 6(a) has more fluctuations, the reason is that it was running in the harshest working condition. The prediction are more stable in Fig. 6(b) and (c). From the selected CDF prediction at the degradation process, it can be noticed that the confidence interval will reduce with the RUL value close to zero. Based on the results, it is obvious that the optimized model can predict the normal period accurately, the essential rule of degradation period has been learned from the other three training data sets, and the uncertainty reduces as the RUL decrease. The trained deep model did not see any data from the testing scenarios during the training process, but it still perform very well. This result demonstrates that the proposed method can learn the essential function under the same condition, which proves its generalization capability. Fig. 7 shows the QH-loss and the three measure indicators (MAE, RMSE, R^2) during the training procedure.

TABLE 6. Quantitative results of other data-driven methods and different feature mapping structures.

		Condition 1				Condition 2				Condition 3			
Bearing		1	2	3	4	1	2	3	4	1	2	3	4
DNN	MAE	19.588	35.519	21.010	7.421	18.428	36.845	86.805	5.118	90.761	14.182	46.972	28.462
	RMSE	20.958	39.655	22.169	7.873	18.671	41.453	93.072	5.375	91.898	14.373	47.479	32.846
	R2	−0.914	−0.247	−1.379	−1.084	−7.868	−0.166	−0.895	−1.402	−8.856	−7.639	−10.270	−0.005
LSTM	MAE	19.729	35.522	20.955	7.378	18.456	36.682	85.784	5.077	90.480	14.219	46.630	28.411
	RMSE	21.125	39.621	22.104	7.829	18.698	41.077	91.785	5.329	91.613	14.412	47.129	32.769
	R2	−0.944	−0.245	−1.365	−1.060	−7.894	−0.144	−0.843	−1.361	−8.795	−7.685	−10.105	−0.001
TFR-CNN [14]	MAE	21.075	41.900	23.879	7.801	18.961	37.100	89.404	5.252	94.510	14.311	48.402	26.816
	RMSE	22.658	46.718	25.298	8.301	19.217	42.288	96.226	5.498	95.723	14.504	48.922	31.931
	R2	−1.237	−0.731	−2.098	−1.316	−8.394	−0.213	−1.025	−1.513	−9.693	−7.797	−10.966	0.050
MFE-CNN [16]	MAE	19.105	35.142	21.167	7.525	18.846	37.321	84.193	4.988	90.854	14.478	47.511	28.224
	RMSE	20.393	39.130	22.327	7.998	19.100	42.129	90.010	5.233	91.995	14.675	48.047	32.485
	R2	−0.812	−0.215	−1.413	−1.150	−8.281	−0.204	−0.772	−1.276	−8.877	−8.005	−10.542	0.016
CNN-3	MAE	5.911	23.454	5.525	6.664	0.677	12.989	17.902	1.013	11.141	1.655	3.625	24.698
	RMSE	7.391	27.221	8.042	6.968	1.634	17.725	28.924	1.661	19.650	2.355	5.002	29.497
	R2	0.762	0.412	0.687	−0.633	0.932	0.787	0.817	0.770	0.549	0.768	0.875	0.189
CNN-base	MAE	4.269	16.270	4.901	**2.336**	**0.857**	12.371	21.067	0.851	**5.478**	**0.470**	2.119	20.848
	RMSE	6.161	23.838	8.251	**3.222**	**1.488**	16.709	30.553	1.520	**15.737**	**1.021**	3.617	23.314
	R2	0.835	0.549	0.670	**0.651**	**0.944**	0.811	0.796	0.807	**0.711**	**0.956**	0.935	0.494
RseNet-5	MAE	3.643	16.551	**2.662**	2.926	0.814	11.708	18.557	**0.711**	5.908	0.929	3.584	**17.700**
	RMSE	5.532	21.747	**4.698**	3.782	1.743	15.150	30.467	**1.405**	18.272	1.401	5.273	**21.162**
	R2	0.867	0.624	**0.893**	0.519	0.923	0.844	0.797	**0.835**	0.610	0.918	0.861	**0.583**
DenseNet-5	MAE	**3.128**	**13.792**	3.200	4.054	0.863	**10.363**	**17.678**	0.855	4.578	0.725	**1.549**	19.265
	RMSE	**5.220**	**18.801**	5.837	4.570	1.804	**13.050**	**28.441**	1.550	17.109	1.257	**3.396**	21.795
	R2	**0.881**	**0.719**	0.835	0.298	0.917	**0.884**	**0.823**	0.800	0.658	0.934	**0.942**	0.557

TABLE 7. Details of the feature mapping structure of CNN-3 [44].

	Block	Layer type	Kernel	Stride	Channels	
					in	out
Feature Extractor	1	Convolution	5×1	2×1	1	16
		ReLU				
		Max Pooling	2×1	2×1		
	2	Convolution	3×1	2×1	16	32
		ReLU				
		Max Pooling	2×1	2×1		
	3	Convolution	3×1	2×1	32	64
		ReLU				
		Max Pooling	2×1	2×1		

TABLE 8. Details of the feature mapping structure of ResNet-5.

	Block	Layer type	Kernel	Stride	Channels	
					in	out
Feature Extractor	1	Convolution	7×1	2×1	1	16
		ReLU				
		Max Pooling	3×1	2×1		
	2	ResNet Block			16	32
	3	ResNet Block			32	64
	4	ResNet Block			64	64
	5	ResNet Block			64	64
ResNet Block	Base layer	Convolution	3×1	1×1	N_{in}	N_{out}
		BatchNorm				
		ReLU				
		Convolution	3×1	1×1	N_{out}	N_{out}
		BatchNorm				
	Match Layer	Convolution	1×1	1×1	N_{in}	N_{out}
		BatchNorm				

	Shortcut Connections	if $N_{in}=N_{out}$: $y=F(x,W_i)+x$
		else: $y=F(x,W_i)+W_sx$
		Where: W_i is Base Block and W_s is the Match Block				

		ReLU	$y=ReLU(y)$			

It indicates that the new QH-loss can effectively converge and the plausibility of using this loss as the evaluation criterion.

2) COMPARISON WITH DIFFERENT LOSS FUNCTIONS

In order to further analyze the proposed QH-loss, the typical MSE and MAE loss and QH-loss with different N_τ are selected for comparison. When using MSE, MAE, and QH-loss (1), only one linear node is set to the last layer, while the QH-loss (33) and (97) mean that 33 nodes and 97 nodes are in the last layer, respectively. We use R^2 as the comparison index because it is a normalized index only reflecting the performance but would not vary with the number of the actual lifetime. The results are shown in Fig. 8, which demonstrate that the QH-loss performs significantly better than MSE and MAE under the same structure, and more N_τ can not only predict the distributional RUL but also lead to better performance.

3) COMPARISON WITH DIFFERENT FPTS

The proposed method could be obviously affected by the FPT. The reason is that we only use the degradation period data to optimize the deep model and the FTP determines the start point of the degradation period. Therefore, the data of three unseen test scenarios are selected as examples for studying the impact of FPT. The current FPT for this situation is following Table 2. Based on the current FPT, we move the FPT proportionally between the normal period and the degradation period, then retrain the model and calculate the R^2 for the test data. The results are shown in Fig. 9(b) and (c), which indicate that the proposed FPT determination is reasonable and the performance of the proposed model will drop dramatically with moving away from the current FPT. Meanwhile, the FPT shows no major effect on the performance during the normal period of the first scenarios, shown in Fig. 9(a). However, the FPT is still sensitive during the degradation process.

TABLE 9. Details of the feature mapping structure of DenseNet-5.

	Block	Layer type	Kernel	Stride	Channels	
					in	out
Feature Extractor	1	Convolution	7×1	2×1	1	64
		BatchNorm				
		ReLU				
		Max Pooling	3×1	2×1		
	2	DenseNet Block			64	96
		Transition layer			96	48
	3	DenseNet Block			48	80
		Transition layer			80	40
	4	DenseNet Block			40	72
		Transition layer			72	36
	5	DenseNet Block			36	68
DenseNet Block	Dense Layer	BatchNorm			N_{in}	$N_{gr} \times N_{bn}$
		ReLU				
		Convolution	1×1	1×1		
		BatchNorm			$N_{gr} \times N_{bn}$	N_{gr}
		ReLU				
		Convolution	3×1	1×1		
	Dense Connections	$x_l = H_l([x_0, x_1, ..., x_{l-1}])$ where $[x_0, x_1, ..., x_{l-1}]$ refers to the concatenation of the feature-maps produced in each layer				
Transition layer		BatchNorm			$N_{in} + lN_{gr}$	$\frac{N_{in}+lN_{gr}}{2}$
		ReLU				
		Convolution	1×1	1×1		
		Average Pooling	2×1	2×1		

E. COMPARISON WITH OTHER DATA-DRIVEN METHODS AND DIFFERENT FEATURE MAPPING STRUCTURES

In this section, more experiments have been conducted for comparison and analysis. While one bearing is selected for testing, the rest three bearings' data under the same condition are used for training.

For comparison with the state-of-the-art data-driven RUL prediction methods, TFR-CNN [14] combines the wavelet transform (WT) processing the degradation data with multiscale CNN and MEF-CNN [16] combining short-time Fourier transform (STFT) with multiscale CNN, have been selected. Meanwhile, following the typical data-driven RUL prediction research [13], we also select DNN and LSTM as the benchmark method for a fair comparison. Comparing the results shown in Table 6 between the proposed model of the CNN-base structure and the benchmark methods, it is obvious that our proposed method is significantly superior according to the three quantified indicators.

To further study the advantages and disadvantages of the proposed method, other three different feature mapping structures have been selected compared with the CNN-base structure. The detailed structures of CNN-3, ResNet-5, and DenseNet-5 are summarized in Tables 7–9, respectively. CNN-3 has three convolution basics, two blocks less than CNN-base. ResNet-5 and DenseNet-5 have five blocks similar to CNN-base, but they are following the structure of ResNet [39] and DenseNet [40]. The quantified results are shown in Table 6, which illustrates the proposed method need feature mapping to have enough complexity of high-level representation so that the essential features could be learned during the training process and the obtained model can perform outstandingly. The results demonstrate that CNN-base has obtained good performance. Compared with ResNet-5 and DenseNet-5, it has a more simple structure and fewer parameters. Therefore, the training speed will be much quicker using the same computer hardware.

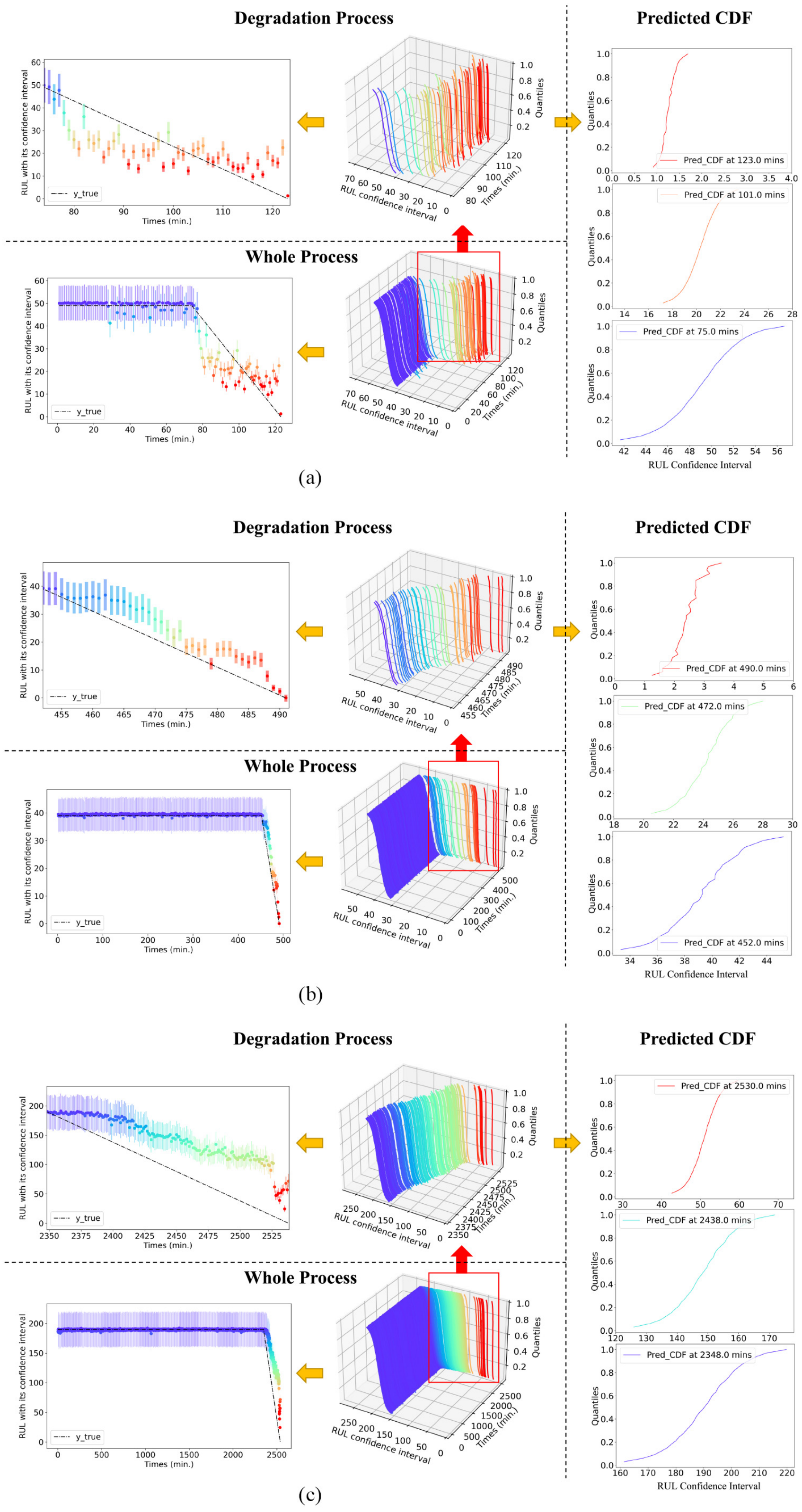
F. FEATURE VISUALIZATION ANALYSIS

For further investigating the effectiveness of our proposed method and explaining the performance of various feature mappings on the run-to-failure bearing experiments, the features of the linear layer before the output layer are used for visualization. We use the *t-SNE* [45] technique to compress these high-dimensional features into 2-D visual images.

The test scenario of bearing-1 of condition-2 has been chosen as the supporting example for investigation. The visualization results are shown in Fig. 10, the color bar is changing from 0 (red) to 100 (blue), which represents the normalized ground-truth RUL label y_i^t but it has been adjusted from [0,1] to [0,100] by multiplied 100 and rounded to integer in order to better view. The various feature visualization results of training data, shown in Fig. 10(a), (c), (e), and (g), demonstrate that the models have been trained well enough because the labeled features are changing from 100 to 0 regularly and there is no obvious misgrading. The high generalization ability of the proposed method has been proved by the results of the unseen test scenario, shown in Fig. 10(b), (d), (f), and (h). Although the learned model never sees the test data during the training process, it can still recognize the various degree of degradation clearly. These visualization results confirm the quantified results in Table 6, verifying the outstanding effectiveness and generalization ability of the proposed method. The visualization figures of training show the rule of gradual change, which indicates that the deep model is capable of learning the essential features, otherwise, the feature will be mixed together irregularly. This proves the effectiveness of the proposed method. The test visualization figures demonstrate a similar rule to the training figures, which verifies that the learned model has good generalization ability and can effectively predict unseen testing scenarios.

V. CONCLUSION AND DISCUSSION

In this article, a distributional RUL prediction method has been proposed, which is capable of directly estimating the RUL uncertainty using the DNN framework and quantile regression loss of distributional learning. The proposed approach has been verified in the real



(a)

(b)

(c)

FIGURE 6. Results of distributional RUL prediction for unseen test scenarios. (a) Bearing-1 of Condition-1. (b) Bearing-1 of Condition-2. (c) Bearing-1 of Condition-3.

run-to-failure bearing experiments under three different working conditions. The main conclusions are summarized as follows.

1) After the training procedure, the proposed model can directly predict the RUL value and uncertainty at each time step for the unseen test scenario. Meanwhile,

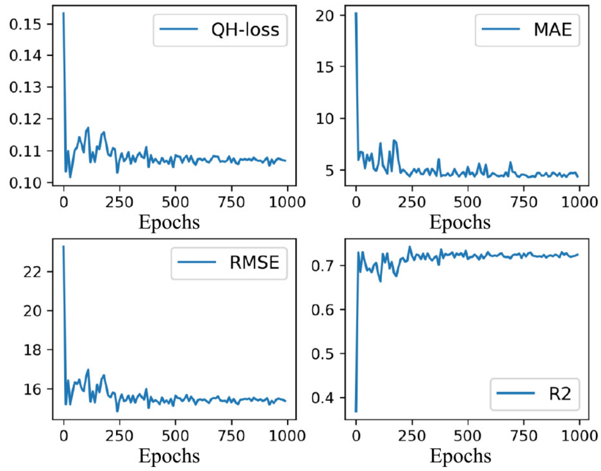


FIGURE 7. QH-loss and three measure indicators during the training procedure.

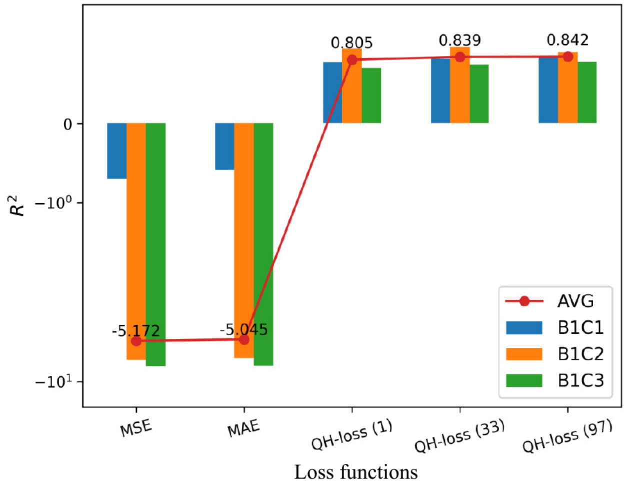


FIGURE 8. R^2 of the different loss functions and different quantile numbers of the proposed QL-loss.

the novel QH-loss can quickly converge and lead the optimization to obtain the best model.

2) In comparison with the state-of-the-art data-driven RUL prediction methods, the proposed model shows better performance and generalization capabilities.
3) In order to obtain better performance, more convolution blocks or the advanced ResNet block and DenseNet block should be used to construct the DNN framework.
4) The feature visualization can prove that the model has been trained well. These results also verify the outstanding performance and generalization ability of the proposed method for the unseen test scenario.

The performance of the proposed method has been verified by a series of experiments and comparisons. However, there are still certain future directions that can be further investigated to promote and expand the current research, which is summarized as follows.

1) The FPT has been proved as one of the impact factors for the performance of the proposed method. However, the current work only uses a simple and effective way, and it can possibly be improved by more advanced FPT determination and RUL labeling method in the future to calculate a more reasonable FPT for the run-to-failure bearing experiment.
2) The impact of different kinds of feature mapping structures has been studied. However, the constructed way of building the optimal structure should be further investigated when the data information is unknown in advance.
3) Currently, our proposed method has been tested in the unseen test scenario, but the working condition of training and test data is still consistent. To further promote the generalization ability, we will extend it to different working conditions.

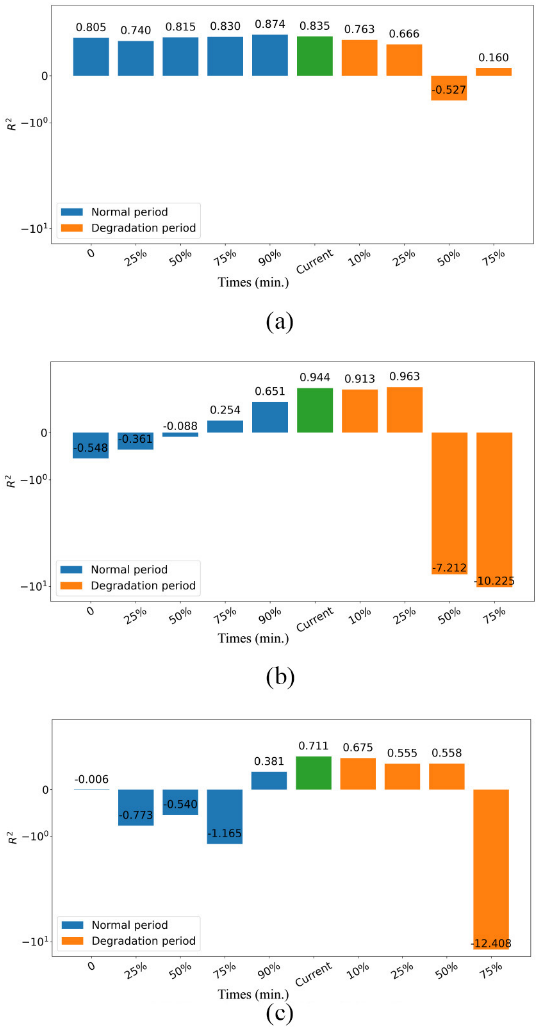


FIGURE 9. R^2 of different FPTs for three unseen test scenarios. (a) Bearing-1 of Condition-1. (b) Bearing-1 of Condition-2. (c) Bearing-1 of Condition-3.

APPENDIX

The structures of CNN-3, ResNet-5, and DenseNet-5 that have been designed for making a comparison with different feature mapping structures are shown in Tables 7–9,

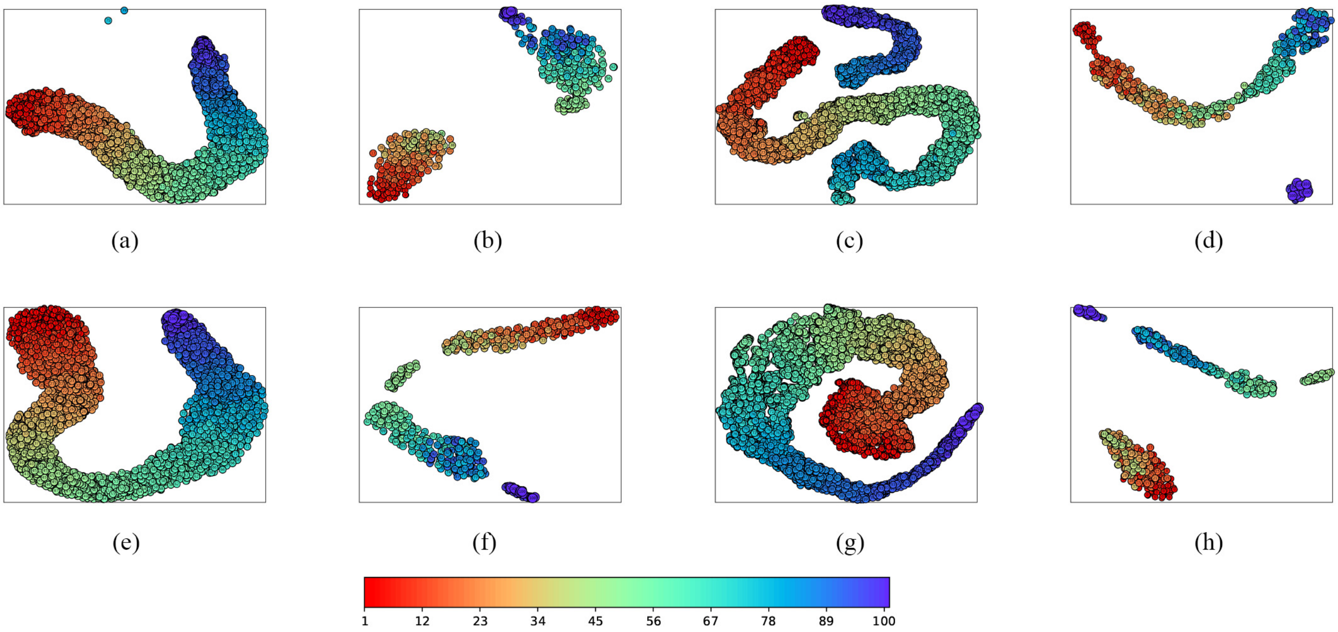


FIGURE 10. *t*-SNE visualization results of the proposed method with different structures of feature mapping under the test scenario of bearing-1 of condition-2. (a) Training of CNN-3. (b) Test of CNN-3. (c) Training of CNN-base. (d) Test of CNN-base. (e) Training of ResNet-5. (f) Test of ResNet-5. (g) Training of DenseNet-5. (h) Test of DenseNet-5.

respectively. The comparison results and analysis are summarized in Section IV-E.

REFERENCES

[1] S. Ramezani, A. Moini, and M. Riahi, "Prognostics and health management in machinery: A review of methodologies for RUL prediction and roadmap," *Int. J. Ind. Eng. Manage. Sci.*, vol. 6, no. 1, pp. 38–61, 2019.

[2] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, May 2018.

[3] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation—A review on the statistical data driven approaches," *Eur. J. Oper. Res.*, vol. 213, no. 1, pp. 1–14, 2011.

[4] L. Liao and F. Köttig, "Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction," *IEEE Trans. Rel.*, vol. 63, no. 1, pp. 191–207, Mar. 2014.

[5] Y. Wang, Y. Zhao, and S. Addepalli, "Remaining useful life prediction using deep learning approaches: A review," *Procedia Manuf.*, vol. 49, pp. 81–88, Oct. 2020.

[6] M. Zhang *et al.*, "Predictive maintenance for remanufacturing based on hybrid-driven remaining useful life prediction," *Appl. Sci.*, vol. 12, no. 7, p. 3218, 2022.

[7] Y. Qian and R. Yan, "Remaining useful life prediction of rolling bearings using an enhanced particle filter," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 10, pp. 2696–2707, Oct. 2015.

[8] Y. Lei, N. Li, S. Gontarz, J. Lin, S. Radkowski, and J. Dybala, "A model-based method for remaining useful life prediction of machinery," *IEEE Trans. Rel.*, vol. 65, no. 3, pp. 1314–1326, Sep. 2016.

[9] J. B. Ali, B. Chebel-Morello, L. Saidi, S. Malinowski, and F. Fnaiech, "Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network," *Mech. Syst. Signal Process.*, vol. 56, pp. 150–172, May 2015.

[10] M. Xia, T. Li, T. Shu, J. Wan, C. W. De Silva, and Z. Wang, "A two-stage approach for the remaining useful life prediction of bearings using deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3703–3711, Jun. 2019.

[11] A. L. Ellefsen, E. Bjørlykhaug, V. Æsøy, S. Ushakov, and H. Zhang, "Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture," *Rel. Eng. Syst. Safety*, vol. 183, pp. 240–251, Mar. 2019.

[12] H. Pei *et al.*, "An adaptive prognostics method for fusing CDBN and diffusion process: Application to bearing data," *Neurocomputing*, vol. 421, pp. 303–315, Jan. 2021.

[13] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," *Rel. Eng. Syst. Safety*, vol. 172, pp. 1–11, Apr. 2018.

[14] J. Zhu, N. Chen, and W. Peng, "Estimation of bearing remaining useful life based on multiscale convolutional neural network," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3208–3216, Apr. 2019.

[15] B. Wang, Y. Lei, N. Li, and T. Yan, "Deep separable convolutional network for remaining useful life prediction of machinery," *Mech. Syst. Signal Process.*, vol. 134, Dec. 2019, Art. no. 106330.

[16] X. Li, W. Zhang, and Q. Ding, "Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction," *Rel. Eng. Syst. Safety*, vol. 182, pp. 208–218, Feb. 2019.

[17] M. Xia, X. Zheng, M. Imran, and M. Shoaib, "Data-driven prognosis method using hybrid deep recurrent neural network," *Appl. Soft Comput.*, vol. 93, Aug. 2020, Art. no. 106351.

[18] Z. Chen, M. Wu, R. Zhao, F. Guretno, R. Yan, and X. Li, "Machine remaining useful life prediction via an attention-based deep learning approach," *IEEE Trans. Ind. Electron.*, vol. 68, no. 3, pp. 2521–2531, Mar. 2021.

[19] Y. Cheng, J. Wu, H. Zhu, S. W. Or, and X. Shao, "Remaining useful life prognosis based on ensemble long short-term memory neural network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, Oct. 2020.

[20] R. Guo, Y. Wang, H. Zhang, and G. Zhang, "Remaining useful life prediction for rolling bearings using EMD-RISI-LSTM," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, Jan. 2021.

[21] J. Li, X. Li, and D. He, "A directed acyclic graph network combined with CNN and LSTM for remaining useful life prediction," *IEEE Access*, vol. 7, pp. 75464–75475, 2019.

[22] T. Xia, Y. Song, Y. Zheng, E. Pan, and L. Xi, "An ensemble framework based on convolutional bi-directional LSTM with multiple time windows for remaining useful life estimation," *Comput. Ind.*, vol. 115, Feb. 2020, Art. no. 103182.

[23] X. Li, W. Zhang, H. Ma, Z. Luo, and X. Li, "Data alignments in machinery remaining useful life prediction using deep adversarial neural networks," *Knowl.-Based Syst.*, vol. 197, Jun. 2020, Art. no. 105843.

[24] H. Yang, K. Ding, R. C. Qiu, and T. Mi, "Remaining useful life prediction based on normalizing flow embedded sequence-to-sequence learning," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1342–1354, Dec. 2021.

[25] S. Sankararaman and K. Goebel, "Why is the remaining useful life prediction uncertain," in *Proc. Annu. Conf. Prognostics Health Manage. Soc.*, 2013, pp. 1–13.

[26] B. Wang, Y. Lei, T. Yan, N. Li, and L. Guo, "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery," *Neurocomputing*, vol. 379, pp. 117–129, Feb. 2020.

[27] Z. Zhao, J. Wu, D. Wong, C. Sun, and R. Yan, "Probabilistic remaining useful life prediction based on deep convolutional neural network," in *Proc. 9th Int. Conf. Through-Life Eng. Services*, 2020, p. 6.

[28] Z. Pang, X. Si, C. Hu, D. Du, and H. Pei, "A Bayesian inference for remaining useful life estimation by fusing accelerated degradation data and condition monitoring data," *Rel. Eng. Syst. Safety*, vol. 208, Apr. 2021, Art. no. 107341.

[29] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 449–458.

[30] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, vol. 2, 2001, pp. 251–256.

[31] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.* 2017, pp. 214–223.

[33] A. Müller, "Integral probability metrics and their generating classes of functions," *Adv. Appl. Probab.*, vol. 29, no. 2, pp. 429–443, 1997.

[34] R. Koenker and K. F. Hallock, "Quantile regression," *J. Econ. Perspect.*, vol. 15, no. 4, pp. 143–156, 2001.

[35] R. Koenker, S. Leorato, and F. Peracchi, "Distributional vs. quantile regression," EIEF Working Paper 1329, Einaudi Inst. Econ. Finan. (EIEF), Rome, Italy, 2013.

[36] F. O. Heimes, "Recurrent neural networks for remaining useful life estimation," in *Proc. Int. Conf. Prognost. Health Manage.*, 2008, pp. 1–6.

[37] H. Li, W. Zhao, Y. Zhang, and E. Zio, "Remaining useful life prediction using multi-scale deep convolutional neural network," *Appl. Soft Comput.*, vol. 89, Apr. 2020, Art. no. 106113.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[41] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 492–518.

[42] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020.

[43] D. Wang, M. Zhang, Y. Xu, W. Lu, J. Yang, and T. Zhang, "Metric-based meta-learning model for few-shot fault diagnosis under multiple limited data conditions," *Mech. Syst. Signal Process.*, vol. 155, Jun. 2021, Art. no. 107510.

[44] M. Zhang, D. Wang, W. Lu, J. Yang, Z. Li, and B. Liang, "A deep transfer model with wasserstein distance guided multi-adversarial networks for bearing fault diagnosis under different working conditions," *IEEE Access*, vol. 7, pp. 65303–65318, 2019.

[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

MING ZHANG (Member, IEEE) received the B.S. and Ph.D. degrees in mechanical engineering from the Beijing University of Chemical Technology, Beijing, China, in 2011 and 2017, respectively.

He was a Postdoctoral Fellow with the Department of Information Science and Technology, Shenzhen International Graduate School, Tsinghua University, Beijing, China. He is currently a Research Associate with the College of Engineering and Physical Sciences, Aston University, Birmingham, U.K. His research interests include predictive maintenance and condition monitoring, intelligent fault diagnosis and remaining useful life prediction, and deep learning in industrial scenarios, especially transfer learning and few-shot learning.

DUO WANG received the B.S. degree in automation from the Harbin Institute of Technology, Harbin, China, in 2015, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2022.

His current research interests include deep learning with incomplete data, multimodal learning, unsupervised learning, and their application in fault diagnosis and computer vision.

NASSER AMAITIK received the M.Sc. degree in information and business systems engineering from the University of Surrey, Guildford, U.K., in 2007, and the Ph.D. degree in computer science from Aston University, Birmingham, U.K., in 2018.

He has over 10 years of combined working experience in both industry and academia. During his Ph.D. research, he investigated intuitive models of assessment and decision making. He is currently working as a Research Associate with the College of Engineering and Physical Sciences, Aston University. His main area of research lies in the intelligent knowledge engineering for decision support systems, predictive modeling related to industrial applications, reasoning with uncertainty, and cost modeling for engineering domain.

YUCHUN XU received the B.S. degree in automotive engineering and the Ph.D. degree in manufacturing engineering from the Harbin Institute of Technology, Harbin, China, in 1993 and 1999, respectively.

He is the Chair in Manufacturing, the Head of the Smart Systems and Manufacturing Group, and the Leader of two interdisciplinary research themes Digital Engineering and Circular Economy with Aston University, Birmingham, U.K. His research is mainly focused on smart manufacturing, remanufacturing, life cycle engineering, and cost modeling/analysis. His research is sponsored by EPSRC, Innovate UK, Horizon EU, H2020, FP7, etc., with total profile over £46M. His research closely links with industry to address the interdisciplinary challenges associated with product life cycle, including product development, manufacturing, maintenance, remanfuacturing/life extension, and recyle/disposal. His research has strong synergy with digital manufacturing, sustainable manufacturing, life cycle engineering, asset management, and circular economy.