

Article

Stage-Aware Multi-Task Learning with Causal Degradation-Prior Fusion for Remaining Useful Life Prediction

Lei Song ¹, Chuanhao Zheng ², Shengkai Zhao ², Qin Bie ¹, Zhixiang Dai ¹, Feng Wang ¹, Jinjie Zhang ² and Jiachen Liu ^{2,*} 

¹ PetroChina Southwest Oil & Gasfield Surface Engineering Design Center, Chengdu 610000, China

² Diagnosis and Self-Recovery Engineering Research Center, Beijing University of Chemical Technology, Beijing 100029, China; 2016500030@buct.edu.cn (J.Z.)

* Correspondence: ymcslj@163.com

Abstract

Remaining useful life (RUL) prediction of wind-turbine bearings is challenged by non-stationary wind loads, multistage degradation, substantial lifetime dispersion, and strict deployment constraints. Conventional single-task regressors apply a unified feature-to-RUL mapping over the entire life cycle and therefore struggle to characterize stage transitions and bearing-specific degradation progress. To address these issues, this paper proposes a stage-aware multi-task RUL prediction method with causal degradation-prior fusion and collaborative distillation. During training, a high-capacity reference representation path transfers inter-sample relational structures and task-level degradation knowledge to a compact feature encoding path, while only the compact path is retained for inference. Based on the compact representation, a multi-task module jointly performs four-stage classification, stage-conditioned RUL regression, and continuous remaining-life estimation; predicted stage probabilities softly fuse the stage-conditioned outputs. A causal prior-fusion module further integrates a bearing-specific healthy-state anchor, causally identified first prediction time, cumulative damage, and the stage-aware prediction to adapt the RUL trajectory to individual degradation processes. Experiments on the IEEE PHM 2012 and XJTU-SY datasets demonstrate that the proposed method provides accurate and robust RUL prediction across different bearing degradation processes. Moreover, the compact inference path maintains efficient implementation, supporting its potential use in practical wind-turbine condition-monitoring applications.

Keywords: wind-turbine bearing; remaining useful life prediction; stage-aware; multi-task learning



Academic Editor: Dimitrios Giapopoulos

Received: 5 August 2026

Revised: 3 September 2026

Accepted: 7 September 2026

Published: 9 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Modern industrial equipment increasingly relies on intelligent fault diagnosis and physics-informed condition modeling to improve operational reliability [1,2]. Rolling bearings are critical components of rotating machinery, and their degradation directly affects equipment availability and maintenance decisions [3]. This issue is particularly relevant to wind-turbine bearing systems, where fluctuating wind conditions continuously induce variations in speed and load, resulting in strongly nonstationary operating environments [4,5]. Under such conditions, bearing degradation does not follow a uniform and stationary trajectory but evolves through multiple stages with different signal characteristics and degradation rates [6]. Moreover, variations in manufacturing, installation, lubrication,

loading history, and environmental exposure introduce substantial inter-bearing variability, such that even bearings of the same type may exhibit different degradation onset times and lifetime trajectories [5,7]. Therefore, practical RUL prediction requires both the identification of stage-dependent degradation patterns and adaptation to bearing-specific degradation processes.

Classical bearing life assessment methods, such as SKF rating-life models [8,9], estimate bearing life using engineering-prior information, including operating load, reliability requirements, lubrication conditions, and contamination level. These models are valuable for bearing design and maintenance planning when such information is available. In contrast, this study focuses on vibration-based, data-driven RUL prediction using publicly available run-to-failure datasets, which do not provide complete lubrication, contamination, and reliability-related information for each sample. Therefore, the proposed method is complementary to, rather than a replacement for, SKF-type life models.

Data-driven methods, particularly deep learning, have been extensively applied to remaining useful life (RUL) prediction owing to their powerful feature representation capabilities [10–12]. Many studies have improved prediction performance through hybrid modeling, temporal dependency learning, and adaptive degradation modeling. For example, Wang et al. [13] proposed an MFF-HI and used a weighted TCN (WTCN) to improve prediction accuracy. Song et al. [10] proposed a GCN-PR network using a spatiotemporal graph to extract spatial and temporal information for RUL prediction. Lin et al. [14] combined Contrastive Predictive Coding (CPC) with a transformer decoder for self-supervised pretraining on non-full life cycle data. Han et al. [15] highlighted the lab-to-field gap in health estimation and combined scientific AI, self-supervised learning, and transfer learning to establish a standardized data sharing mechanism. Fan et al. [16] developed an adaptive multistage sparse Gaussian process regression framework to model stage-dependent degradation dynamics. These models have demonstrated excellent feature extraction and prediction performance. However, increasingly complex architectures lead to larger parameter counts and higher computational overhead, which restrict their deployment in resource-constrained monitoring scenarios.

Beyond RUL prediction, recent studies in intelligent bearing diagnosis have also explored effective feature representations for rolling-bearing health monitoring. Spirito et al. [17] combined symmetrized dot-pattern indices with a feedforward neural network for rolling-bearing fault detection, whereas Chen et al. [18] developed an interpretable wavelet Kolmogorov–Arnold convolutional LSTM for spatiotemporal feature extraction and intelligent fault diagnosis. Although these studies focus on fault diagnosis rather than RUL prognostics, they further highlight the importance of discriminative and interpretable degradation representations for bearing health monitoring.

By reducing model size while preserving degradation-related knowledge, knowledge distillation offers a promising approach to balancing prediction accuracy and inference efficiency. In the field of RUL prediction, Xu et al. [19] developed KDnet-RUL to transfer degradation knowledge from a complex temporal model to a lightweight convolutional model for edge deployment. Ma et al. [20] proposed KDCN for solid-state lithium-ion batteries, transferring multilayer representations into a compact network to reduce parameter and computational overhead. Ye et al. [21] introduced DCI-KANs to improve the noise robustness and inference efficiency of battery RUL prediction through knowledge-guided model compression. Ren et al. [22] further proposed a lightweight and adaptive knowledge distillation framework to improve RUL prediction accuracy while facilitating deployment on resource-constrained devices. These studies demonstrate the potential of knowledge distillation to achieve efficient RUL prediction while retaining informative degradation representations.

However, applying knowledge distillation to bearing RUL prediction still faces some interrelated challenges. First, bearing degradation exhibits distinct stage-dependent signal characteristics and degradation rates. Although jointly modeling health states and RUL or constructing stage-specific prediction relationships can improve prognostic performance [23,24], conventional single-task regression generally employs a unified feature-to-RUL mapping over the entire life cycle without explicitly distinguishing degradation stages. Consequently, samples from different degradation stages are processed by the same regression function, which may lead to RUL overestimation, underestimation, or irregular prediction fluctuations near stage transitions.

Second, bearings may differ considerably in their total lifetimes and actual degradation onset times. Although the first prediction time (FPT) can be adaptively identified from vibration features [25,26], it is commonly used only for data segmentation or label construction and is rarely incorporated into subsequent RUL prediction. Consequently, a fixed normalized-time prior may assign similar degradation progress to bearings that are actually at different degradation states, limiting its ability to characterize bearing-specific degradation trajectories and lifetime scales.

Third, the reference representation path and the compact feature encoding path differ substantially in structural depth, feature dimensionality, and temporal modeling capability, resulting in discrepancies between their degradation representations. Relying solely on output-response transfer or pointwise feature matching cannot adequately convey the relational degradation structures learned by the reference representation path, thereby limiting the discriminability and stability of the compact degradation representation [27,28].

These challenges are closely related. Stage-aware prediction requires a discriminative compact degradation representation, while the resulting stage-aware output must be further adapted to bearing-specific degradation onset and cumulative-damage information. Therefore, introducing explicit degradation-stage semantics, constructing temporally causal degradation priors from current and historical observations, and learning a consistent compact degradation representation remain key issues in accurate and efficient bearing RUL prediction.

As illustrated in Figure 1, the three challenges manifest as (a) degradation-stage ambiguity in single-task RUL regression, (b) lifetime-dependent stage misalignment caused by a fixed time prior, and (c) representation mismatch between the reference representation path and the compact feature encoding path. Collectively, these challenges limit stage discriminability, lifetime adaptability, and the effectiveness of degradation-knowledge transfer.

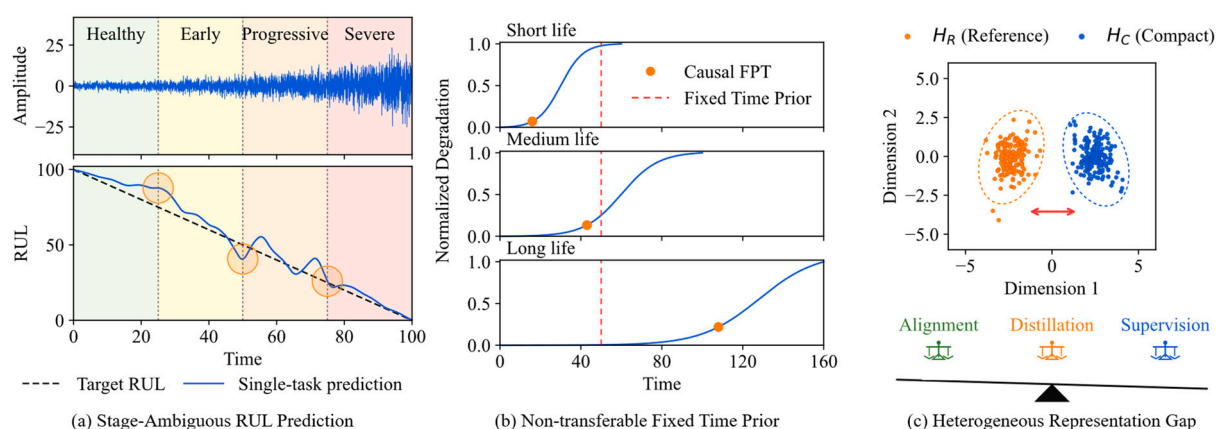


Figure 1. Three key challenges in stage-aware bearing RUL prediction.

To address these challenges, this paper proposes a stage-aware multi-task bearing RUL prediction method with causal degradation-prior fusion, as illustrated in Figure 2. The

compact feature encoding path first extracts a compact temporal degradation representation for subsequent stage-aware prediction. During training, the reference representation path provides relational and task-level guidance to improve the degradation discriminability of the compact representation, whereas only the compact feature encoding path is retained during inference. The main contributions are summarized as follows:

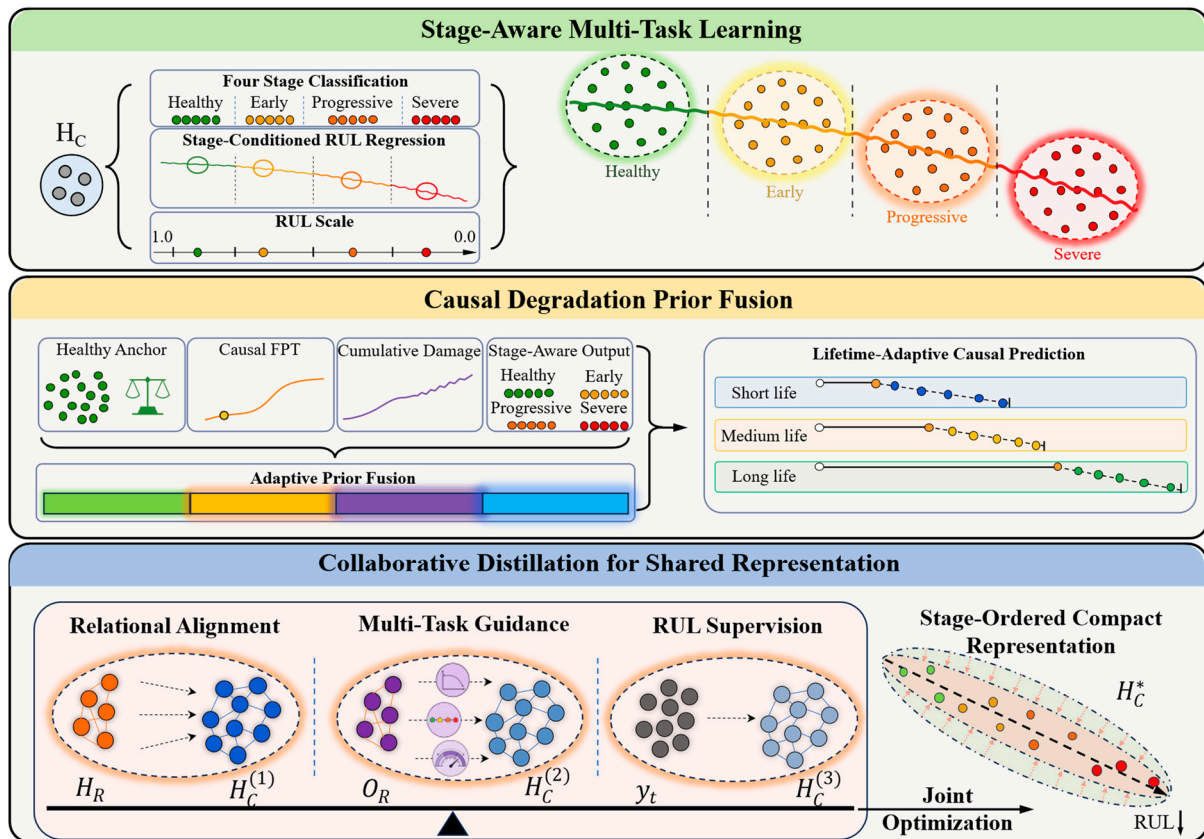


Figure 2. Problem-oriented illustration of the proposed stage-aware bearing RUL prediction method, with the downward arrow next to RUL indicating its decrease as degradation progresses.

(1) A stage-aware multi-task module jointly performs four-stage classification, stage-conditioned RUL regression, and continuous degradation-progress estimation. Soft fusion of the stage-conditioned predictions enables differentiated RUL mappings and reduces prediction discontinuities near stage transitions.

(2) A causal degradation-prior fusion module identifies the FPT using only current and historical observations and constructs bearing-specific priors from the healthy-state anchor, degradation onset, and cumulative damage. These priors are adaptively integrated with the stage-aware RUL output to accommodate bearing-specific degradation onset and lifetime variations.

(3) A collaborative distillation-based compact representation-learning mechanism is developed. By aligning inter-sample correlation structures and multi-task degradation outputs between the reference representation path and the compact feature encoding path, the proposed mechanism enables the compact path to learn discriminative degradation representations while maintaining low inference cost.

Figure 2 provides a problem-oriented illustration of how the three components address the challenges identified in Figure 1. First, stage-aware multi-task learning introduces explicit degradation-stage semantics and differentiated RUL mappings to reduce ambiguity near stage transitions. Second, causal degradation-prior fusion combines the stage-aware output with the healthy anchor, a causally identified FPT, and cumulative damage, thereby

adapting the prediction trajectory to different lifetime scales. Third, collaborative distillation transfers relational and multi-task degradation knowledge from the reference representation path to the compact feature encoding path during training. Consequently, the compact path learns a discriminative degradation representation while maintaining efficient inference without the reference path.

The remainder of this paper is organized as follows. Section 2 describes the proposed method, experimental datasets, implementation details, and evaluation metrics. Section 3 presents the comparative performance, computational-efficiency, and ablation results. Section 4 discusses the implications and limitations of the results. Section 5 concludes the paper and summarizes the main findings.

2. Materials and Methods

2.1. Overall Framework

As shown in Figure 3, the proposed method consists of a compact feature encoding path, a stage-aware multi-task degradation modeling module, and a temporally causal degradation-prior fusion module. During training, a reference representation path provides relational and task-level guidance for learning the compact degradation representation.

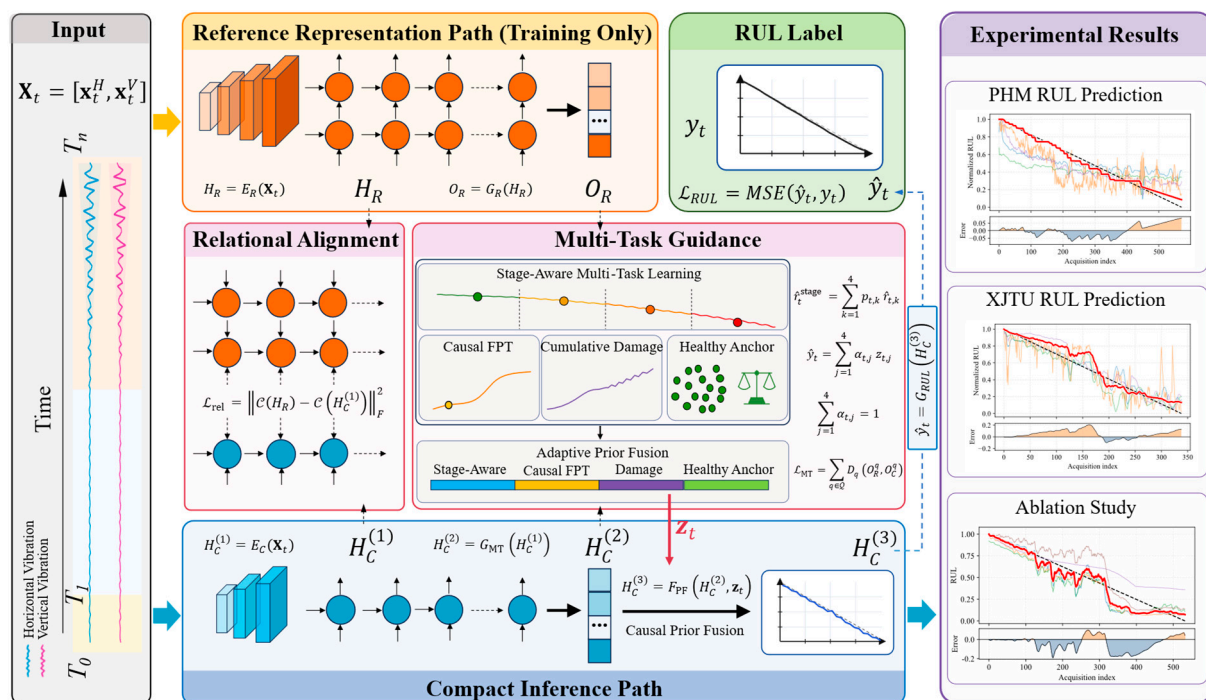


Figure 3. Overall architecture of the proposed stage-aware bearing RUL prediction method with causal degradation-prior fusion.

First, the compact feature encoding path extracts the compact degradation representation from the horizontal and vertical vibration signals. Second, the stage-aware multi-task module jointly estimates degradation-stage probabilities, stage-conditioned RUL responses, and continuous degradation progress. Third, the stage-aware RUL result is fused with bearing-specific priors constructed from the healthy-state anchor, causal FPT, and cumulative damage to obtain the final prediction.

The collaborative representation-learning mechanism is activated only during training. During inference, the reference representation path and the associated collaborative objectives are removed, and only the compact feature encoding path and the subsequent prediction modules are retained.

The stage-aware multi-task degradation modeling module, temporally causal degradation-prior fusion module, and collaborative compact representation-learning mechanism are described in Sections 2.2–2.4, respectively.

2.2. Stage-Aware Multi-Task Degradation Modeling

Bearing degradation exhibits stage-dependent signal characteristics and degradation rates. A unified single-task RUL regression function may therefore fail to characterize the differentiated feature-to-RUL mappings across degradation stages, particularly near stage-transition regions.

To address this issue, a stage-aware multi-task degradation modeling module is constructed based on the compact temporal degradation representation, as illustrated in Figure 3. The module jointly performs degradation-stage classification, stage-conditioned RUL regression, and continuous remaining-life estimation. In particular, the predicted stage probabilities are used not only as auxiliary degradation-state information but also as soft gating weights for the stage-conditioned regression branches. The module therefore integrates discrete stage semantics with continuous degradation-progress information to generate a stage-aware RUL prediction.

2.2.1. Degradation-Stage Classification

Let $H_{C,t}^{(2)}$ denote the compact temporal degradation representation generated by the temporal encoding layer of the compact feature encoding path at time t , as illustrated in Figure 3. According to the causal degradation indicators extracted from the current and historical observations, each training sample is assigned to one of four degradation stages. Let q_t , u_t , and D_t denote the FPT confidence, instantaneous shock, and normalized cumulative damage at time t , respectively. The degradation-stage label s_t is defined as:

$$s_t = \begin{cases} 4, & D_t \geq 0.80, \\ 3, & 0.25 \leq D_t < 0.80, \\ 2, & D_t < 0.25 \text{ and } (q_t \geq 0.50 \text{ or } u_t \geq 0.20), \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

The four labels correspond to the healthy, degradation-onset, progressive-degradation, and severe-degradation stages, respectively. Because q_t , u_t , and D_t are constructed using only observations available at or before time t , the stage labels do not require future vibration measurements.

The corresponding stage-probability distribution is estimated as:

$$\mathbf{p}_t = \text{Soft max}(W_{\text{stage}} H_{C,t}^{(2)} + b_{\text{stage}}) \quad (2)$$

where $\mathbf{p}_t = [p_{t,1}, p_{t,2}, p_{t,3}, p_{t,4}]$ and $\mathbf{p}_t$ denotes the predicted probability that the current sample belongs to the c -th degradation stage. Moreover, W_{stage} and b_{stage} denote the weight matrix and bias of the stage-classification head, respectively.

The predicted stage probabilities are subsequently used as soft gating weights for the stage-conditioned RUL regression branches.

2.2.2. Stage-Conditioned RUL Regression

The feature-to-RUL relationship varies across different degradation stages. Consequently, a single regression branch with a unified set of parameters may be insufficient to characterize the entire degradation process, particularly near stage-transition regions.

To model the differentiated feature-to-RUL relationships, four stage-conditioned RUL regression branches are constructed based on the compact temporal degradation representation. Each branch receives the same compact representation as input but employs an independent set of parameters to learn the RUL mapping associated with a specific degradation stage. The output of the c -th regression branch is expressed as:

$$\hat{y}_{t,c} = f_c\left(\mathbf{H}_{C,t}^{(2)}; \boldsymbol{\theta}_c\right), \quad c = 1, 2, 3, 4 \quad (3)$$

where $f_c(\cdot)$ denotes the c -th stage-conditioned RUL regression function, $\boldsymbol{\theta}_c$ denotes its learnable parameters, and $\hat{y}_{t,c}$ denotes the corresponding stage-conditioned RUL estimate.

After obtaining the outputs of the four regression branches, the predicted stage probabilities are employed to perform soft stage-aware fusion:

$$\hat{y}_t^{stage} = \sum_{c=1}^4 p_{t,c} \hat{y}_{t,c} \quad (4)$$

where $p_{t,c}$ denotes the predicted probability of the c -th degradation stage, $\hat{y}_{t,c}$ denotes the RUL estimate produced by the corresponding stage-conditioned regression branch, and $\hat{y}_t^{stage}$ denotes the resulting stage-aware RUL prediction.

Unlike hard stage assignment, the probability-based fusion mechanism allows multiple stage-conditioned regression branches to jointly contribute to RUL estimation. When a sample is located near the transition between adjacent degradation stages, the branch contributions vary continuously with the predicted stage probabilities, thereby reducing prediction discontinuities caused by hard stage switching. The resulting stage-aware prediction $\hat{y}_t^{stage}$ is subsequently provided to the degradation-prior fusion module in Section 2.3.

2.2.3. Continuous Degradation-Progress Estimation

In addition to degradation-stage classification and stage-conditioned RUL regression, a remaining-life estimation branch is introduced to provide continuous degradation-progress information. Unlike the stage-conditioned regression branches, which estimate RUL directly from degradation-sensitive vibration features, this branch first estimates the remaining number of acquisitions and then converts it into a normalized time-based RUL estimate.

Let e_t denote the number of acquisitions elapsed up to time t , and let N denote the total number of acquisitions of a training bearing. The ground-truth remaining number of acquisitions is defined as:

$$l_t = N - 1 - e_t \quad (5)$$

where l_t gradually decreases to zero as the bearing approaches failure. The total number of acquisitions N is used only to construct the supervision target for the training bearings and is not required during inference.

Based on the compact temporal degradation representation and the predicted stage probabilities, the remaining number of acquisitions is estimated as:

$$\hat{l}_t = m_c \text{Softplus}\left(f_r\left([H_{C,t}^{(2)}; p_t]\right)\right) \quad (6)$$

where $[H_{C,t}^{(2)}; p_t]$ denotes the concatenation of the compact temporal degradation representation and the four-dimensional stage-probability vector, $f_r(\cdot)$ denotes the learnable remaining-life estimation function, and m_c denotes the nominal lifetime determined from the training bearings under the same operating condition. The Softplus function ensures that $\hat{l}_t$ remains positive.

The estimated remaining number of acquisitions is subsequently converted into a normalized time-based RUL estimate:

$$\hat{y}_t^{time} = \frac{\hat{l}_t}{e_t + \hat{l}_t} \quad (7)$$

where $\hat{y}_t^{time} \in [0, 1]$ represents the continuous remaining-life estimate derived from the elapsed and predicted remaining acquisition numbers. During inference, its calculation requires only the elapsed acquisition number and the observations available up to the current time, without using the total lifetime or future measurements of the test bearing.

This branch does not independently determine the final RUL prediction. Instead, it provides complementary continuous degradation-progress information for adaptive fusion and serves as an auxiliary constraint on the temporal ordering of the compact degradation representation. Its supervision and corresponding task-level guidance are incorporated into the collaborative objective described in Section 2.4.3.

2.3. Causal Degradation-Prior Fusion

Bearings may differ considerably in their total lifetimes and actual degradation onset times. Consequently, a fixed normalized-time prior may not accurately represent the actual degradation states of bearings with different lifetime scales. Although the stage-aware RUL prediction introduced in Section 2.2 accounts for stage-dependent feature-to-RUL relationships, it does not explicitly incorporate the bearing-specific healthy baseline, degradation onset, or accumulated degradation degree.

To address this limitation, a causal degradation-prior fusion module is constructed using the current and historical vibration observations. As illustrated in the overall framework in Figure 3, the module first constructs a healthy-state anchor from the initial healthy observations. The deviation of the current compact degradation representation from this anchor is then used to identify the first prediction time (FPT) and calculate the cumulative damage after the FPT. Finally, the healthy-state information, causal FPT, cumulative damage, and stage-aware RUL prediction are adaptively fused and further integrated with the compact degradation representation to obtain the final RUL prediction.

Here, “causal” indicates that all degradation priors at time t are constructed exclusively from observations available at or before t , without accessing future measurements.

2.3.1. Healthy-State Anchor Construction

During the initial operating period, a bearing generally remains in a relatively stable healthy state. Therefore, the compact degradation representations extracted from the initial healthy observations are used to construct a bearing-specific healthy-state anchor:

$$\mathbf{a}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} H_{C,i}^{(2)} \quad (8)$$

where $H_{C,i}^{(2)}$ denotes the compact degradation representation of the i -th initial healthy observation, N_h denotes the number of initial observations used for anchor construction, and $\mathbf{a}_h$ denotes the resulting healthy-state anchor.

The deviation of the current compact degradation representation from the healthy-state anchor is calculated as:

$$\mathbf{d}_t = ||H_{C,i}^{(2)} - \mathbf{a}_h||_2 \quad (9)$$

where $\mathbf{d}_t$ denotes the healthy-state deviation at time t . A small value of $\mathbf{d}_t$ indicates that the current bearing state remains close to the initial healthy baseline, whereas an increasing $\mathbf{d}_t$

indicates a growing departure from the healthy state. The bearing-specific anchor therefore provides a reference for subsequent FPT identification and cumulative-damage calculation.

2.3.2. Causal FPT Identification and Cumulative Damage

Based on the historical evolution of the healthy-state deviation d_t , the FPT is identified as the first time at which d_t persistently exceeds a deviation threshold determined from the historical healthy observations:

$$t_{\text{FPT}} = \min\{t : d_t > \delta_h \text{ for three consecutive acquisitions}\} \quad (10)$$

where δ_h denotes the deviation threshold determined from the historical healthy-state observations, and t_{FPT} denotes the time at which the persistent threshold exceedance is first confirmed.

The persistence requirement prevents an isolated fluctuation or transient disturbance from being incorrectly identified as the onset of degradation. Because the identification process uses only the deviations observed up to the current time, the resulting FPT can adapt to the actual degradation onset of an individual bearing without accessing future observations.

After the FPT, the degradation increments exceeding the deviation level at the FPT are accumulated to obtain the cumulative damage:

$$D_t = \sum_{i=t_{\text{FPT}}}^t \max(0, d_i - d_{\text{FPT}}) \quad (11)$$

where d_{FPT} denotes the healthy-state deviation at the identified FPT and D_t denotes the cumulative damage up to time t . Before the FPT, D_t is set to zero.

The FPT characterizes the bearing-specific degradation onset, whereas the cumulative damage reflects the degradation degree accumulated after that onset. Together, they provide complementary temporal and state-dependent information for subsequent RUL prediction.

2.3.3. Adaptive Degradation-Prior Fusion

The healthy-state anchor, causal FPT, and cumulative damage characterize the healthy baseline, degradation onset, and accumulated degradation degree of the current bearing, respectively. These degradation priors are adaptively combined with the stage-aware RUL prediction obtained in Section 2.2:

$$\mathbf{z}_t = F_{\text{gate}}(\hat{y}_t^{\text{stage}}, a_h, t_{\text{FPT}}, D_t) \quad (12)$$

where $\hat{y}_t^{\text{stage}}$ denotes the stage-aware RUL prediction, $F_{\text{gate}}(\cdot)$ denotes the adaptive prior-fusion function, and $\mathbf{z}_t$ denotes the fused degradation-prior representation. The learnable gating mechanism determines the relative contributions of the stage-aware prediction and causal degradation priors, while the implementation parameters associated with the causal priors are determined exclusively from the training or development bearings.

The adaptive gating function learns the relative contributions of the stage-aware prediction and the bearing-specific degradation priors according to the current degradation state. During the healthy period, the model can rely more strongly on the relatively stable stage-aware prediction and healthy-state information. After persistent degradation is detected, the contributions of the FPT and cumulative-damage information can be increased to adapt the prediction to the actual degradation onset and degradation degree.

As illustrated in Figure 3, the fused degradation-prior representation is further combined with the compact degradation representation through the feature-prior fusion function:

$$H_{C,t}^{(3)} = F_{FPF}(H_{C,t}^{(2)}, \mathbf{z}_t) \quad (13)$$

where $F_{FPF}(\cdot)$ denotes the feature-prior fusion function and $H_{C,t}^{(3)}$ denotes the prior-enhanced compact degradation representation.

Finally, the RUL prediction head generates the final RUL estimate:

$$\hat{y}_t = G_{RUL}(H_{C,t}^{(3)}) \quad (14)$$

where $G_{RUL}(\cdot)$ denotes the final RUL prediction head and $\hat{y}_t$ denotes the final predicted RUL.

Through the proposed fusion mechanism, the final prediction incorporates the stage-dependent RUL mapping, bearing-specific healthy baseline, causal degradation onset, and accumulated degradation degree. This reduces the prior-state misalignment caused by fixed normalized-time information and improves adaptation to bearings with different degradation onset times and lifetime scales. The corresponding final RUL supervision is incorporated into the collaborative training objective described in Section 2.4.3.

2.4. Collaborative Distillation for Compact Degradation Representation Learning

Collaborative distillation guides the compact feature encoding path using the relational and task-level degradation knowledge learned by the reference representation path. As shown in Figure 3, relational alignment constrains the compact degradation representation, whereas multi-task guidance transfers the stage-aware outputs defined in Section 2.2.

2.4.1. Reference Multi-Scale Degradation Representation Network

To extract multi-scale degradation patterns and temporal evolution information from vibration signals, a Multi-Scale Degradation Representation Network is constructed as the reference representation path. The horizontal and vertical vibration signals are processed through successive multi-scale feature extraction blocks, in which multiple receptive fields capture complementary local degradation patterns at different temporal scales.

The scale-specific features are adaptively aggregated using learnable normalized weights and subsequently summarized through dual pooling, thereby preserving both global degradation trends and salient local responses with negligible additional parameter overhead. The pooled waveform features are then fused with auxiliary degradation descriptors containing operating-condition and temporal information.

The fused acquisition-level features within a historical window are temporally encoded to capture degradation dependencies across consecutive acquisitions. The resulting reference degradation representation is expressed as:

$$H_R = E_R(X_t) \quad (15)$$

where X_t denotes the input vibration sequence, $E_R(\cdot)$ denotes the Multi-Scale Degradation Representation Network, and H_R denotes the resulting reference degradation representation.

Based on H_R , the reference prediction heads generate multi-task degradation outputs as:

$$O_R = G_R(H_R) \quad (16)$$

where $G_R(\cdot)$ denotes the reference multi-task prediction heads and O_R contains the degradation-stage probabilities, stage-conditioned RUL responses, and continuous

degradation-progress estimates defined in Section 2.2. During collaborative training, H_R provides relational guidance, while O_R provides task-level guidance for the compact feature encoding path.

Before collaborative training, the reference representation path was independently pretrained using the ground-truth RUL, degradation-stage, and remaining-life labels of the training bearings. After pretraining, its parameters were frozen, and the resulting reference representations and multi-task outputs were used as fixed guidance for compact representation learning.

2.4.2. Compact Degradation Feature Encoding Network

To obtain a compact degradation representation for subsequent RUL prediction, a Compact Degradation Feature Encoding Network is constructed. Rather than reproducing the high-capacity reference path, it adopts a streamlined hierarchical encoding strategy to retain degradation-sensitive information with fewer parameters and lower computational cost.

The compact network progressively extracts local degradation patterns from the vibration signals and aggregates the resulting waveform features through pooling. The pooled waveform features are fused with the same auxiliary degradation descriptors used by the reference path, ensuring input-information consistency between the two paths. The fused features are subsequently encoded over a historical window to capture temporal degradation dependencies.

The resulting compact degradation representation is expressed as:

$$H_C^{(1)} = E_C(X_t) \quad (17)$$

where $E_C(\cdot)$ denotes the Compact Degradation Feature Encoding Network and $H_C^{(1)}$ denotes the compact degradation representation. During collaborative training, $H_C^{(1)}$ receives relational guidance from H_R and task-level guidance from O_R . It is subsequently provided to the stage-aware multi-task module to obtain $H_C^{(2)}$.

2.4.3. Correlation-Guided Collaborative Distillation

The collaborative training objective consists of relational alignment, multi-task output guidance, and final RUL supervision.

First, because the reference and compact paths employ heterogeneous representation mechanisms, directly matching their feature values may impose an excessively strict constraint. Therefore, the correlation structures of H_R and $H_C^{(1)}$ are aligned instead of enforcing element-wise feature equality. The relational alignment loss is defined as:

$$L_{rel} = \|C(H_R) - C(H_C^{(1)})\|_F^2 \quad (18)$$

where $C(\cdot)$ denotes the correlation-structure mapping and $\|\cdot\|_F$ denotes the Frobenius norm. This loss transfers the relationships among degradation samples from H_R to $H_C^{(1)}$, thereby reducing the representation gap between the two heterogeneous paths.

Second, the reference multi-task outputs are employed as complementary soft targets for the corresponding outputs of the compact path. The multi-task guidance loss is formulated as:

$$L_{MT} = \sum_{q \in Q} D_q(O_R^q, O_C^q) \quad (19)$$

where Q denotes the set of transferred tasks, O_R^q and O_C^q denote the outputs of the reference and compact paths for task q , and $D_q(\cdot)$ denotes the corresponding task-specific

discrepancy function. Distribution-based outputs are transferred using divergence constraints, whereas continuous RUL and degradation-progress outputs are transferred using regression constraints.

Third, the ground-truth RUL label is employed to directly supervise the final prediction generated by the compact path. The RUL supervision loss is defined as:

$$L_{RUL} = MSE(\hat{y}_t, y_t) \quad (20)$$

where $\hat{y}_t$ denotes the final RUL prediction obtained after causal degradation-prior fusion and y_t denotes the corresponding ground-truth RUL label. This loss ensures that the collaborative knowledge-transfer process remains consistent with the actual RUL prediction objective.

In addition to the final RUL supervision, the ground-truth degradation-stage, stage-aware RUL, and remaining-life labels are used to supervise the corresponding auxiliary tasks. The auxiliary-task supervision loss is defined as:

$$L_{aux} = \beta_1 L_{cls} + \beta_2 L_{stage-RUL} + \beta_3 L_{rem} \quad (21)$$

where L_{cls} denotes the degradation-stage classification loss, $L_{stage-RUL}$ denotes the stage-aware RUL regression loss, and L_{rem} denotes the remaining-life estimation loss. In all reported experiments, $\beta_1, \beta_2, \beta_3$ were set to 0.25, 0.55, and 0.18 for the degradation-stage classification loss, stage-aware RUL regression loss, and remaining-life estimation loss, respectively. These coefficients were fixed throughout training.

Finally, the overall collaborative training objective is formulated as:

$$L_{total} = L_{rel} + L_{MT} + L_{RUL} + L_{aux} \quad (22)$$

The component-level loss coefficients were fixed in all reported experiments. Specifically, the relational alignment loss was weighted by 0.08. For multi-task output guidance, the signal-RUL, final-RUL, stage-distribution, and remaining-life distillation terms were weighted by 0.35, 0.25, 0.10, and 0.08, respectively. For final RUL supervision, the final-RUL loss, sequence-level RUL loss, prediction-bias loss, and fusion-gate regularization were weighted by 1.00, 0.20, 0.20, and 0.01, respectively. The auxiliary-task coefficients are given in Equation (21).

The coefficients were fixed a priori according to the intended roles of the loss terms and were subsequently examined through the local sensitivity analysis described below. The final-RUL loss was assigned a coefficient of 1.00 as the primary optimization objective. The auxiliary-task coefficients were set to provide complementary supervision without overwhelming the final-RUL objective, whereas the distillation and relational-alignment coefficients were kept smaller to balance knowledge transfer against label-based supervision.

To assess the local robustness of these fixed coefficients, we performed a sensitivity analysis using three non-test bearings, without using the final test-bearing labels. The network architecture, optimization settings, training epochs, and random seed were kept unchanged. For each coefficient group, all coefficients in that group were separately scaled to $0.8\times$ and $1.2\times$ of their original values, while the other coefficient group was kept at $1.0\times$. The final test-bearing labels were not used in this analysis. As summarized in Table 1, the original setting achieved the lowest mean MAE (0.1725). Although the $1.2\times$ distillation/alignment setting slightly improved mean R2 and RMSE, no single perturbation consistently improved all metrics. Therefore, the reported coefficients were retained as a stable trade-off and were fixed throughout all reported experiments; no dynamically learned loss weighting was used.

Table 1. Local sensitivity analysis of the fixed loss coefficients on three non-test bearings, with final test-bearing labels excluded.

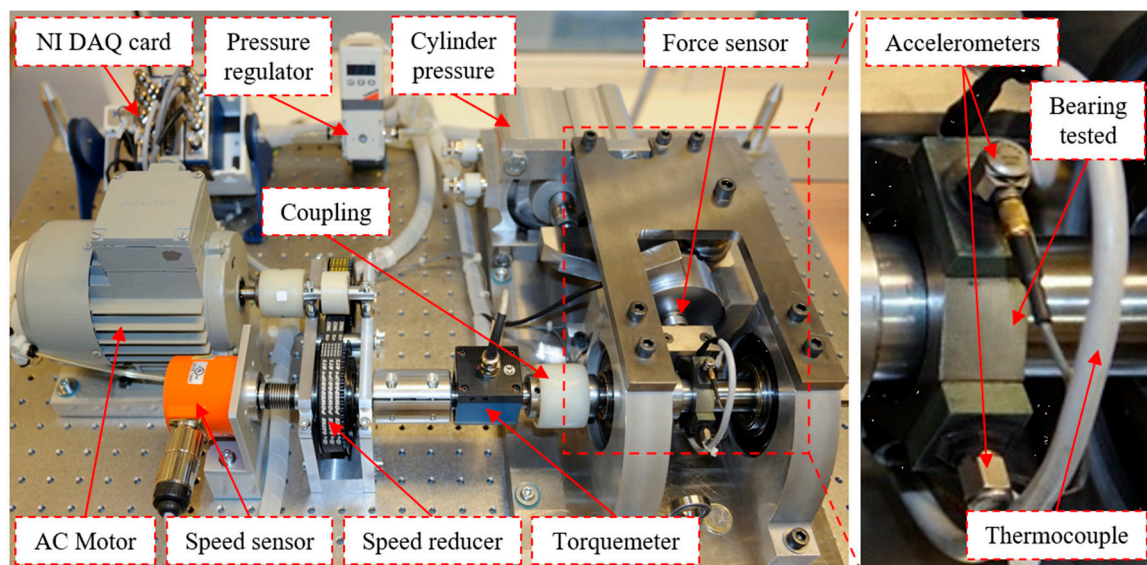
Setting	Auxiliary Scale	Distillation/Alignment Scale	Mean R ²	Mean RMSE	Mean MAE
Original	1.0	1.0	0.5176	0.1962	0.1725
Auxiliary reduced	0.8	1.0	0.5056	0.1988	0.1749
Auxiliary increased	1.2	1.0	0.5101	0.1979	<u>0.1732</u>
Distillation/alignment reduced	1.0	0.8	0.4902	0.2014	0.1744
Distillation/alignment increased	1.0	1.2	0.5231	0.1959	0.1740

Bold and underlined values indicate the best and second-best results, respectively.

During collaborative training, the parameters of the reference representation path remain unchanged, whereas the compact feature encoding path, stage-aware multi-task module, and temporally causal degradation-prior fusion module are jointly optimized. During inference, the reference representation path and the associated collaborative objectives are removed. Only the compact feature encoding path and the subsequent prediction modules are retained, thereby enabling computationally efficient RUL prediction.

2.5. Datasets and Implementation Details

Experiments were conducted on the XJTU-SY bearing dataset [29] and the IEEE PHM 2012 bearing dataset [30] to evaluate the RUL prediction capability of the proposed method across predefined fixed operating conditions. The corresponding bearing test rigs are illustrated in Figures 4 and 5.

**Figure 4.** PRONOSTIA test platform for the IEEE PHM 2012 bearing dataset.

The publicly available run-to-failure datasets used in this study represent degradation trajectories associated with the accumulation of rolling-contact fatigue during accelerated bearing tests. Horizontal and vertical vibration signals are used to characterize the corresponding health-state evolution. Although the final failure may manifest in different bearing components, the proposed method does not target fault-location diagnosis; instead, it models the stage evolution and remaining-life variation in fatigue-related degradation processes reflected in the vibration responses.

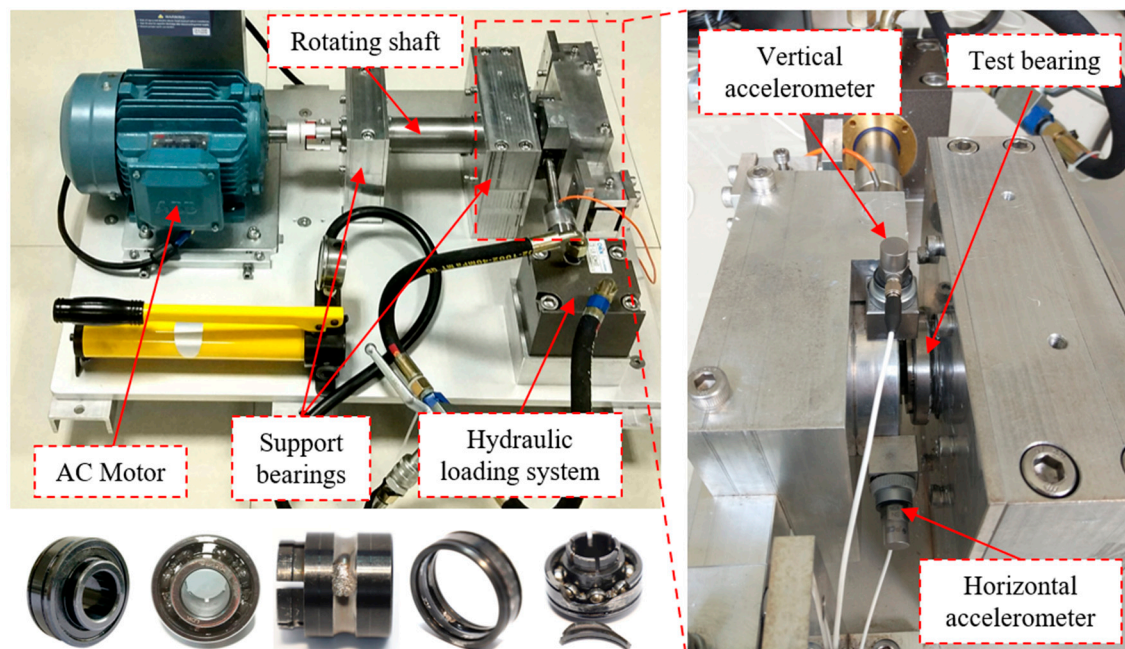


Figure 5. Accelerated bearing degradation test platform for the XJTU-SY bearing dataset.

Because the study reuses public datasets, the test conditions were specified by their original providers rather than controlled in the present work. For the IEEE PHM 2012 data used here, Condition 1 corresponds to 30 Hz (1800 rpm) and a 4.0 kN radial load, whereas Condition 2 corresponds to 27.5 Hz (1650 rpm) and a 4.2 kN radial load. The XJTU-SY dataset includes three prescribed operating conditions: 35 Hz/12 kN, 37.5 Hz/11 kN, and 40 Hz/10 kN. Each bearing run was acquired under one fixed nominal speed-load condition. The PRONOSTIA platform documentation reports a temperature sensor located near the outer bearing ring, with measurements sampled at 10 Hz [30]. In contrast, the XJTU-SY public dataset provides horizontal and vertical vibration signals for the released bearing runs [29]. Neither public source specifies a lubricant type or a controlled test-temperature setpoint for the data used in this study. To ensure a consistent input setting across the two datasets, the present analysis uses only horizontal and vertical vibration signals; temperature and lubricant information are not incorporated into the model inputs or subsequent analysis.

2.5.1. IEEE PHM 2012 Dataset

The IEEE PHM 2012 dataset was acquired using the PRONOSTIA test platform shown in Figure 4. The vibration signals were sampled at 25.6 kHz for 0.1 s every 10 s, resulting in 2560 sampling points per acquisition. The horizontal and vertical vibration signals were used as the two input channels. Following the conventional condition-wise division of the dataset, Bearing 1-1 and Bearing 1-2 were used for model training under Condition 1, whereas Bearing 1-3, Bearing 1-5, and Bearing 1-6 were used for testing. Under Condition 2, Bearing 2-1 and Bearing 2-2 were used for training, and Bearing 2-6 was used for testing. The detailed data division is presented in Table 2.

Table 2. IEEE PHM 2012 bearing dataset.

Conditions	Development Bearings	Test Bearings
1	B1-1, B1-2	B1-3, B1-5, B1-6
2	B2-1, B2-2	B2-6

2.5.2. XJTU-SY Dataset

The XJTU-SY dataset was collected using the accelerated bearing degradation test platform shown in Figure 5, where LDK UER204 rolling bearings were tested. The vibration signals were sampled at 25.6 kHz for 1.28 s at one-minute intervals. A same-condition leave-one-bearing-out protocol was adopted. Specifically, one bearing was selected for testing, while the remaining available bearings under the same operating condition were used for model training, parameter validation, and causal-prior calibration. B3-3 was excluded in advance because of its abnormal degradation trajectory. The detailed data division is presented in Table 3.

Table 3. XJTU-SY bearing dataset.

Conditions	Development Bearings	Test Bearings
1	B1-1, B1-2, B1-4, B1-5	B1-3
2	B2-1, B2-2, B2-4, B2-5	B2-3
	B2-1, B2-2, B2-3, B2-4	B2-5
3	B3-1, B3-2, B3-4	B3-5

2.5.3. Common Preprocessing and Implementation Details

For both datasets, the horizontal and vertical vibration signals were standardized separately, and the waveform entering the convolutional encoder was adaptively pooled to 256 sampling points. The normalized RUL label decreased linearly from 1 at the beginning of operation to 0 at failure. The degradation-stage labels were generated according to the causal indicator-based rule defined in Section 2.2.1.

For causal-prior construction, the first five acquisitions of each bearing were used to construct the initial healthy-state anchor. The FPT threshold was set to the larger of 0.10 and the 99th percentile of the early healthy-state deviations obtained from the first 20 acquisitions of the corresponding training bearings. All causal-prior and calibration parameters were determined exclusively from the training bearings and remained fixed during testing.

All experiments were implemented in PyTorch 2.0.0 and conducted on a computer equipped with an AMD Ryzen 9 7945HX with Radeon Graphics CPU and an NVIDIA GeForce RTX 4060 Laptop GPU. The AdamW optimizer was employed for model training. The reference representation path was trained with a learning rate of 2×10^{-4} , a batch size of 64, and a weight decay of 2×10^{-4} . The batch size of the compact feature encoding path was set to 32. A five-epoch linear warm-up followed by cosine learning-rate decay was adopted, and gradient clipping with a maximum norm of 5.0 was applied. The distillation temperature was set to 2.0. All loss coefficients were fixed throughout training, and no dynamically learned loss weighting was used.

As detailed in Table 4, the reference representation path consists of a multi-scale convolutional encoder, a bidirectional long short-term memory network, a degradation-stage classification branch, and stage-specific RUL regression branches. It is employed during training to extract high-capacity representations with strong degradation-discrimination capability. The compact feature encoding path employs a single-branch convolutional encoder and two gated recurrent units to learn a compact degradation representation under relational alignment and multi-task guidance. During deployment, only the compact feature encoding path is retained to reduce computational and storage costs.

Table 4. Detailed network architecture and parameter settings of the proposed method.

Path	Modules	Structure and Parameters
Reference representation path	Multi-scale convolutional encoder	Multi-scale Conv1d: 2→32→64→96
	Feature fusion	Global average/max pooling; MLP (32)
	Temporal modeling	BiLSTM, hidden size = 64
	Stage classification	Linear (128, 4)
	Stage-specific regression	4 × Linear (128, 1)
	Remaining-life estimation	Linear (132, 32); Linear (32, 1)
	Relational projection	Linear (128, 64)
Compact feature encoding path	Convolutional block 1	Conv1d (2, 32), kernel size = 7, stride = 4
	Convolutional block 2	Conv1d (32, 64), kernel size = 5, stride = 4
	Convolutional block 3	Conv1d (64, 96), kernel size = 3, stride = 4
	Feature fusion	Global average/max pooling; MLP (32)
	Temporal modeling	GRU1, hidden size = 64; GRU2, hidden size = 32
	Stage classification	Linear (32, 4)
	Stage-specific regression	4 × Linear (32, 1)
	Remaining-life estimation	Linear (36, 24); Linear (24, 1)

2.6. Evaluation Metrics

To quantitatively evaluate the RUL prediction performance, the root mean square error (RMSE) and coefficient of determination (R^2) were employed as evaluation metrics. RMSE measures the average deviation between the predicted and ground-truth RUL values and is defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (23)$$

where n denotes the total number of samples involved in the evaluation, y_i denotes the ground-truth value of the i -th sample, and $\hat{y}_i$ denotes the predicted value of the i -th sample.

The coefficient of determination measures the goodness of fit between the predicted and ground-truth RUL trajectories and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}}{\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n}} \quad (24)$$

where $\bar{y}$ denotes the mean of the ground-truth RUL values. An R^2 value closer to 1 indicates better agreement between the predicted and ground-truth trajectories. In addition, R^2 can be negative when the prediction performance is worse than that obtained by directly using the mean ground-truth value.

3. Results

3.1. Performance Evaluation on the PHM 2012 Dataset

To evaluate life-prediction performance of the proposed method, we selected bearing life data under Operating Condition 1 and Operating Condition 2 for testing. Figure 6 presents RUL estimation results of four baseline approaches and the proposed method for Bearings 1-3, 1-5, 1-6, and 2-6. As can be seen from Figure 6, the proposed method can effectively capture bearing degradation information and accurately predict its remaining useful life (RUL).

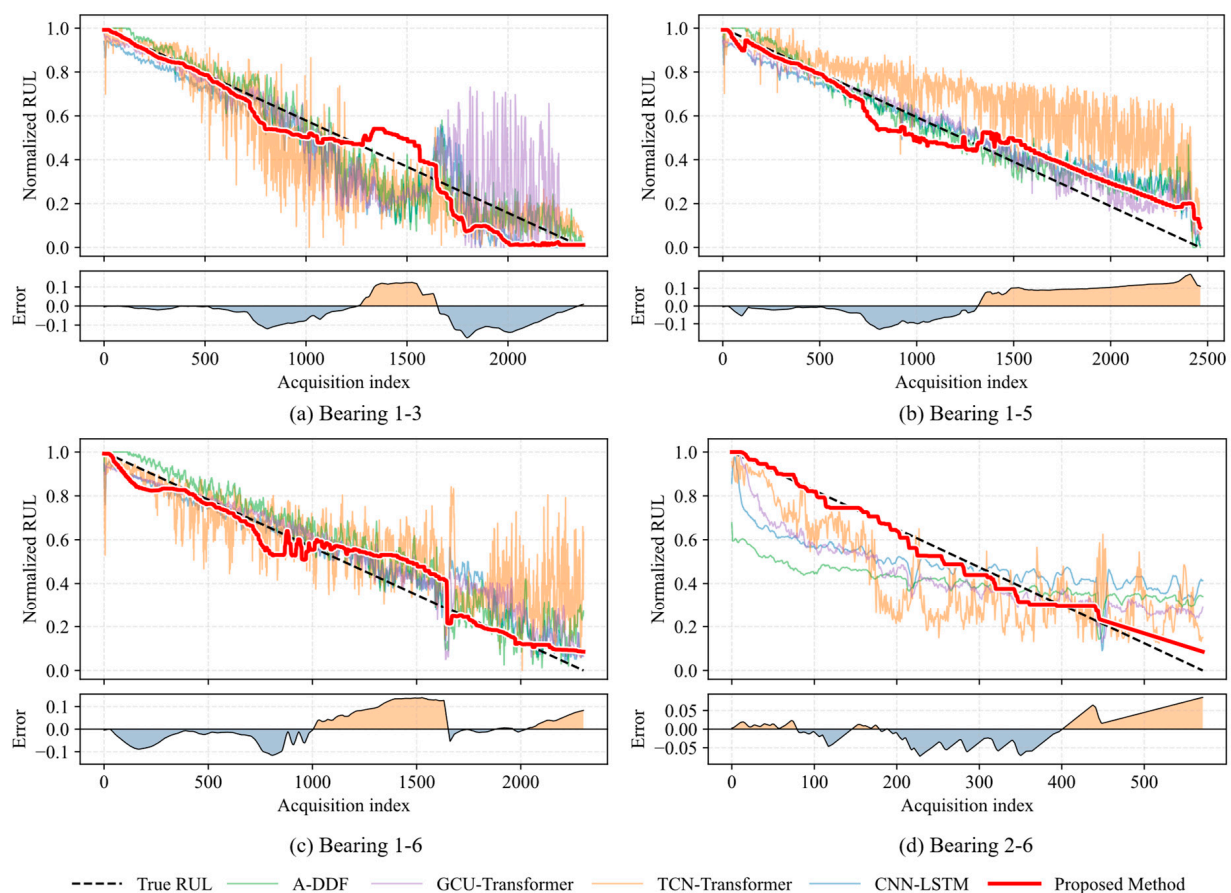


Figure 6. Comparison of RUL prediction curves and residual distributions on the IEEE PHM 2012 dataset.

As shown in Table 5, the four baseline methods are CNN-LSTM, TCN-Transformer [31], A-DDF [32], and GCU-Transformer [33]. The proposed method achieves the best performance on three of the four tested bearings and obtains the best overall performance. Although GCU-Transformer performs better on Bearing 1-5, the proposed method achieves the lowest overall RMSE and the highest overall R^2 across the four test cases.

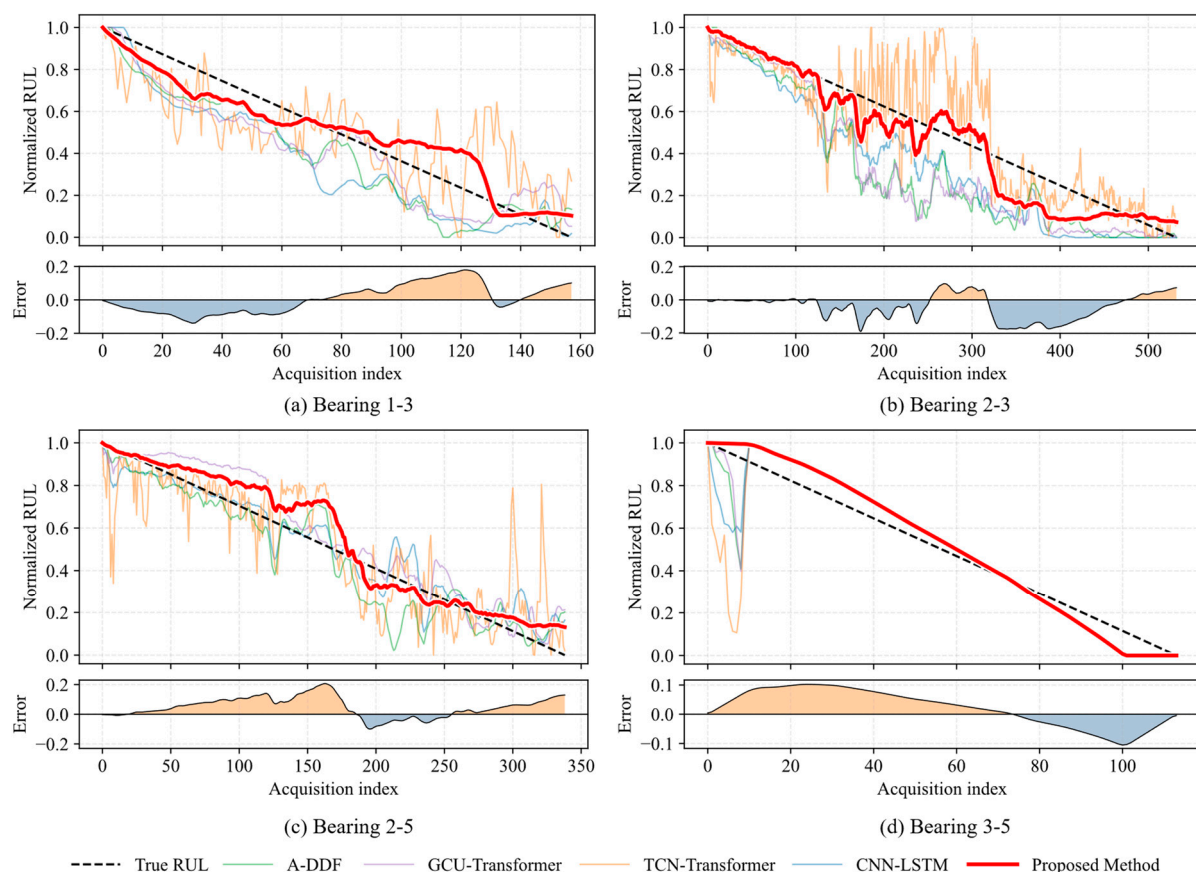
Table 5. Quantitative comparison of RUL prediction performance on the IEEE PHM 2012 dataset.

Method	Metric	Bearing 1-3	Bearing 1-5	Bearing 1-6	Bearing 2-6	Mean
CNN-LSTM	RMSE	0.1199	0.0982	<u>0.0800</u>	0.2004	0.1246 ± 0.0460
	R ²	0.8276	0.8843	<u>0.9233</u>	0.5195	0.7887 ± 0.1591
TCN-Transformer	RMSE	0.1307	0.2348	0.1759	0.1951	0.1841 ± 0.0375
	R ²	0.7953	0.3387	0.6292	0.5450	0.5770 ± 0.1645
A-DDF	RMSE	<u>0.0855</u>	0.0863	0.0937	0.2422	0.1269 ± 0.0666
	R ²	<u>0.9124</u>	0.9108	0.8948	0.2987	0.7542 ± 0.2631
GCU-Transformer	RMSE	0.1435	0.0666	0.0861	<u>0.1667</u>	0.1157 ± 0.0408
	R ²	0.7531	0.9468	0.9112	<u>0.6677</u>	<u>0.8197 ± 0.1141</u>
Proposed method	RMSE	0.0766	0.0858	0.0685	0.0399	0.0677 ± 0.0172
	R ²	0.9296	<u>0.9117</u>	0.9438	0.9809	0.9415 ± 0.0254

Bold and underlined values indicate the best and second-best results, respectively.

3.2. Performance Evaluation on the XJTU-SY Dataset

To evaluate the life-prediction performance of the proposed method, test bearings from all three operating conditions were selected. Figure 7 presents the RUL estimation results of four baseline methods and the proposed method for Bearings 1-3, 2-3, 2-5, and 3-5. As shown in Figure 7, the proposed method effectively captures bearing degradation information and accurately predicts the remaining useful life.

**Figure 7.** Comparison of RUL prediction curves and residual distributions on the XJTU-SY dataset.

As shown in Table 6, the four baseline methods are CNN-LSTM, TCN-Transformer [31], A-DDF [32], and GCU-Transformer [33]. On the XJTU-SY dataset, the proposed method achieves the lowest mean RMSE and the highest mean R^2 across the four test cases. Although CNN-LSTM performs slightly better on Bearing 2-5, the proposed method yields the best overall performance.

Table 6. Quantitative comparison of RUL prediction performance on the XJTU-SY dataset.

Method	Metric	Bearing 1-3	Bearing 2-3	Bearing 2-5	Bearing 3-5	Mean
CNN-LSTM	RMSE	0.1603	0.1781	0.0795	0.1030	<u>0.1302</u> $\pm$ 0.0403
	R^2	0.6954	0.6208	0.9246	0.8748	<u>0.7789</u> $\pm$ 0.1249
TCN-Transformer	RMSE	0.1868	<u>0.1686</u>	0.1639	0.1794	0.1747 $\pm$ 0.0090
	R^2	0.5866	<u>0.6603</u>	0.6796	0.6203	0.6367 $\pm$ 0.0360
A-DDF	RMSE	<u>0.1389</u>	0.2139	0.1065	0.0946	0.1385 $\pm$ 0.0465
	R^2	<u>0.7713</u>	0.4528	0.8648	0.8944	0.7458 $\pm$ 0.1752
GCU-Transformer	RMSE	0.1397	0.2114	0.0953	<u>0.0906</u>	0.1342 $\pm$ 0.0485
	R^2	0.7688	0.4658	0.8916	<u>0.9032</u>	0.7573 $\pm$ 0.1764
Proposed method	RMSE	0.0872	0.0888	0.0838	0.0661	0.0815 $\pm$ 0.0091
	R^2	0.9098	0.9058	0.9162	0.9485	0.9201 $\pm$ 0.0168

Bold and underlined values indicate the best and second-best results, respectively.

3.3. Computational-Efficiency Analysis

To evaluate the computational efficiency of the proposed method, the model complexity was measured using a history length of eight, a two-channel waveform input, and a batch size of one. As shown in Table 7, the compact inference path contains 98,940 parameters and requires 5.703 M FLOPs per sample, representing reductions of 59.53% and 54.16%, respectively, compared with the reference representation path. Moreover, the model size is reduced from 0.933 MB to 0.377 MB. Because the reference representation path is removed after collaborative training, only the compact path is retained during deployment. These results demonstrate that the proposed method substantially reduces inference and storage costs while maintaining competitive RUL prediction performance.

Table 7. Computational complexity of the reference and compact paths.

Path	Parameters	Model Size	FLOPs/Sample
Reference representation path	244,501	0.933 MB	12.441 M
Compact inference path	98,940	0.377 MB	5.703 M

3.4. Ablation Studies on Model Components

In these experiments, the developed approach achieves favorable performance in life prediction. To verify the effectiveness of the proposed innovations, we conducted ablation studies to analyze the role of each component in the method. The ablation experiments were conducted on eight bearing test cases selected from the XJTU-SY and IEEE PHM 2012 datasets. Five ablation variants were constructed by independently removing knowledge distillation (KD), multi-task learning (MTL), FPT-based calibration, waveform features, and time-fusion processing. The corresponding configurations are presented in Table 8.

Specifically, the w/o KD variant removes all reference-guidance losses and trains the compact path using ground-truth supervision only. The w/o MTL variant removes the degradation-stage and remaining-life auxiliary objectives together with their associated task-level guidance, while retaining direct RUL-related learning. The w/o FPT variant removes the FPT-related input features and disables the causal FPT-based output calibration. The w/o Waveform variant removes the raw vibration-waveform input while retaining the

auxiliary degradation descriptors. The w/o Time Fusion variant disables the remaining-life-based time branch and directly uses the stage-aware signal-RUL output as the prediction result. The full model retains all the above components.

Table 8. Configuration of different ablation variants.

Variant	KD	MTL	FPT	Waveform Feature	Time Fusion
w/o KD	×	✓	✓	✓	✓
w/o MTL	✓	×	✓	✓	✓
w/o FPT	✓	✓	×	✓	✓
w/o Waveform	✓	✓	✓	×	✓
w/o Time Fusion	✓	✓	✓	✓	×
Full model	✓	✓	✓	✓	✓

✓ indicates that the corresponding component is included, whereas × indicates that it is removed.

Table 9 presents the quantitative ablation results on the eight bearing test cases, while Figure 8 illustrates the prediction curves and residual distributions of representative bearings. Overall, the complete model achieves the best aggregate prediction performance and follows the actual RUL degradation trend more consistently, with relatively concentrated residual distributions. Removing KD or MTL causes a moderate reduction in overall accuracy, confirming the effectiveness of knowledge transfer and auxiliary degradation-state supervision. Removing waveform features or time-fusion processing generally leads to larger prediction deviations across the evaluated test cases, although the magnitude of degradation varies among individual bearings. In particular, the w/o FPT variant exhibits severe performance deterioration on XJTU Bearing 3-5, indicating that FPT-based calibration plays an important role in degradation-stage alignment. Although individual ablation variants occasionally outperform the complete model on specific bearings, none achieves consistently better performance across both datasets.

Table 9. Quantitative results of the ablation study on eight bearing test cases.

Method	Metric	XJTU				PHM				Mean
		Bearing 1-3	Bearing 2-3	Bearing 2-5	Bearing 3-5	Bearing 1-3	Bearing 1-5	Bearing 1-6	Bearing 2-6	
Full model	RMSE	0.0872	0.0888	0.0838	0.0661	0.0766	0.0858	0.0685	0.0399	0.0746 ± 0.0154
	R ²	0.9098	0.9058	0.9162	0.9485	0.9296	0.9117	0.9438	0.9809	0.9308 ± 0.0241
w/o KD	RMSE	0.1099	0.1170	0.1139	0.0661	0.0807	0.0876	<u>0.0682</u>	<u>0.0374</u>	<u>0.0851 ± 0.0260</u>
	R ²	0.8569	0.8363	0.8453	0.9485	0.9218	0.9079	<u>0.9442</u>	<u>0.9833</u>	<u>0.9055 ± 0.0506</u>
w/o MTL	RMSE	0.0978	<u>0.1106</u>	0.0554	<u>0.0661</u>	<u>0.0751</u>	0.0852	0.0689	0.1301	0.0862 ± 0.0235
	R ²	0.8866	<u>0.8537</u>	0.9634	<u>0.9485</u>	<u>0.9324</u>	0.9129	0.9431	0.7976	0.9048 ± 0.0525
w/o FPT	RMSE	<u>0.0890</u>	0.1319	<u>0.0783</u>	0.5344	0.0753	0.1074	0.1056	0.0336	0.1444 ± 0.1498
	R ²	<u>0.9060</u>	0.7919	<u>0.9269</u>	−2.3675	0.9321	0.8618	0.8664	0.9865	0.4880 ± 1.0806
w/o Waveform	RMSE	0.1349	0.1635	0.1725	0.0661	0.0686	<u>0.0811</u>	0.0685	0.1513	0.1133 ± 0.0436
	R ²	0.7842	0.6803	0.6452	0.9485	0.9436	<u>0.9212</u>	0.9437	0.7263	0.8241 ± 0.1211
w/o Time Fusion	RMSE	0.3681	0.1468	0.1571	0.0651	0.0753	0.0806	0.0681	0.0725	0.1292 ± 0.0965
	R ²	−0.6057	0.7424	0.7056	0.9501	0.9320	0.9221	0.9444	0.9372	0.6910 ± 0.4985

Bold and underlined values indicate the best and second-best results, respectively.

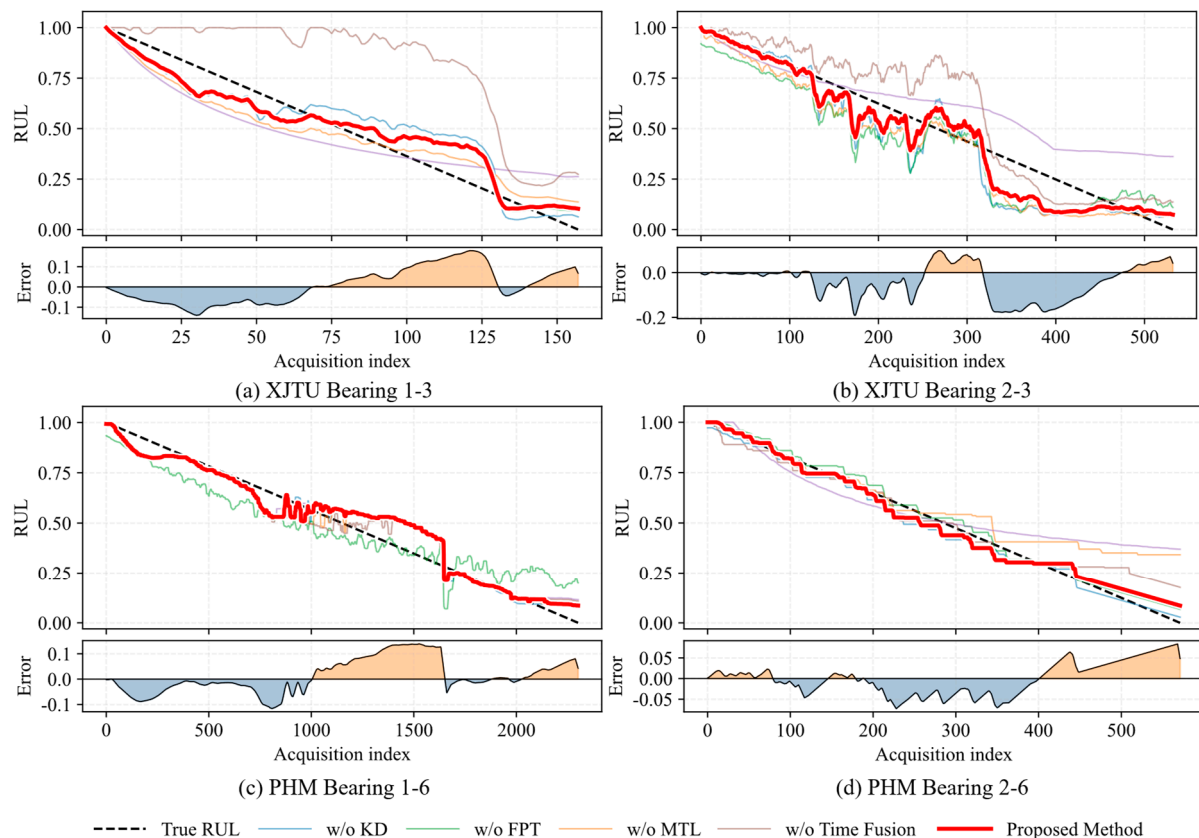


Figure 8. Comparison of RUL prediction curves and residual distributions for different ablation variants.

4. Discussion

The experimental results demonstrate that the proposed method provides accurate RUL prediction across both datasets. On the IEEE PHM 2012 dataset, it achieved an average RMSE of 0.0677 and an R^2 of 0.9415; on the XJTU-SY dataset, it achieved an average RMSE of 0.0815 and an R^2 of 0.9201. These results indicate that the method maintains favorable predictive performance under different bearing degradation trajectories and operating conditions.

The performance gain can be attributed to the combination of stage-aware multi-task modeling and causal degradation-prior fusion. The stage-classification and stage-conditioned regression branches enable differentiated RUL mappings for different degradation stages, which helps reduce prediction discontinuities near stage transitions. In addition, the healthy-state anchor, causally identified first prediction time, and cumulative-damage information enable the model to adapt its prediction trajectory to bearing-specific degradation onset and lifetime variation.

The compact inference path also provides an important practical advantage. Compared with the reference representation path, it reduces the number of parameters and FLOPs by 59.53% and 54.16%, respectively, while retaining competitive prediction accuracy. This result suggests that collaborative distillation can transfer useful degradation knowledge to a lightweight model and thereby improve its suitability for resource-constrained condition-monitoring scenarios.

The ablation results further confirm that the individual components contribute differently to the overall performance. In particular, removing the waveform features or time-fusion processing leads to more pronounced performance degradation, whereas the deterioration of the w/o FPT variant on specific bearings highlights the importance of

degradation-stage alignment. Nevertheless, the present study is limited to accelerated bearing-degradation datasets. Future work should validate the method using long-term field data from wind turbines and investigate its robustness under time-varying operating conditions.

5. Conclusions

This study proposed a stage-aware multi-task RUL prediction method with causal degradation-prior fusion and collaborative distillation for bearing prognostics. The proposed framework jointly performs degradation-stage classification, stage-conditioned RUL regression, and continuous remaining-life estimation, and then fuses stage-specific predictions according to their probabilities. By incorporating a healthy-state anchor, a causally identified first prediction time, and cumulative damage, the causal degradation-prior fusion module further adapts the predicted RUL trajectory to bearing-specific degradation processes. Experiments on the IEEE PHM 2012 and XJTU SY datasets achieved average RMSE/ R^2 values of 0.0677/0.9415 and 0.0815/0.9201, respectively. Ablation studies verified the contributions of collaborative distillation, multi-task learning, FPT-based calibration, waveform features, and time fusion. In addition, compared with the reference representation path, the compact inference path reduced the number of parameters and FLOPs by 59.53% and 54.16%, respectively, indicating that the proposed method achieves a favorable balance between prediction accuracy and computational efficiency.

Nevertheless, the present study provides a data-driven methodological validation using controlled run-to-failure datasets rather than a field-deployment study. The results therefore do not establish performance under nonstationary wind-turbine operating conditions. Moreover, given the limited number of bearings available under each condition, they should not be interpreted as a statistical assessment of bearing fatigue life. Future work will validate the method using long-term wind-turbine field data. In parallel, we will investigate the integration of operating-condition information and conventional fatigue-life knowledge to support more reliable practical prognostics.

Author Contributions: Conceptualization, L.S., J.Z. and J.L.; methodology, J.L. and C.Z.; software, C.Z. and J.L.; validation, C.Z. and S.Z.; formal analysis, J.Z. and J.L.; investigation, S.Z. and F.W.; resources, L.S.; data curation, S.Z. and C.Z.; writing—original draft preparation, J.L. and C.Z.; writing—review and editing, L.S., C.Z., Q.B., J.Z. and J.L.; visualization, C.Z. and J.L.; supervision, Z.D. and J.Z.; project administration, L.S. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. The XJTU-SY and IEEE PHM 2012 (PRONOSTIA) bearing datasets are available from the sources cited in References [29] and [30], respectively.

Conflicts of Interest: Author Lei Song, Qin Bie, Zhixiang Dai and Feng Wang were employed by the company PetroChina Southwest Oil & Gasfield Surface Engineering Design Center. The company had no role in the study design, data collection, data analysis, interpretation of the results, manuscript preparation, or the decision to submit the manuscript for publication. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
RUL	Remaining useful life

FPT	First prediction time
KD	Knowledge distillation
MTL	Multi-task learning
CNN	Convolutional neural network
TCN	Temporal convolutional network
GRU	Gated recurrent unit
BiLSTM	Bidirectional long short-term memory
RMSE	Root mean square error
FLOPs	Floating-point operations
R ²	Coefficient of determination

References

- Peng, Y.; Shao, H.; Xiao, Y.; Yan, S.; Wang, J.; Liu, B. A systematic review on interpretability research of intelligent fault diagnosis models. *Meas. Sci. Technol.* **2025**, *36*, 012009. [\[CrossRef\]](#)
- Wu, J.; Zhao, P.; Shu, W.; Yang, J.; Wang, L.; Han, S. Self-adaptive physics-informed neural networks (SA-PINN) coupled with thermal network model (TNM) for cross-scale temperature field modeling of wet friction components. *Tribol. Int.* **2026**, *214*, 111208. [\[CrossRef\]](#)
- Peng, Y.; Shao, H.; Xiao, Y.; Wang, J.; Liu, B. Dual-stage interpretable domain generalization fault diagnosis: Integrating prior knowledge and gradient-weighted class activation mapping. *Eng. Appl. Artif. Intell.* **2026**, *166*, 113655. [\[CrossRef\]](#)
- Hart, E.; Turnbull, A.; Feuchtwang, J.; McMillan, D.; Golysheva, E.; Elliott, R. Wind turbine main-bearing loading and wind field characteristics. *Wind Energy* **2019**, *22*, 1534–1547. [\[CrossRef\]](#)
- Hart, E.; Clarke, B.; Nicholas, G.; Amiri, A.K.; Stirling, J.; Carroll, J.; Dwyer-Joyce, R.; McDonald, A.; Long, H. A review of wind turbine main bearings: Design, operation, modelling, damage mechanisms and fault detection. *Wind Energy Sci.* **2020**, *5*, 105–124. [\[CrossRef\]](#)
- Zhang, H.B.; Cheng, D.J.; Zhou, K.L.; Zhang, S.W. Deep transfer learning-based hierarchical adaptive remaining useful life prediction of bearings considering the correlation of multistage degradation. *Knowl.-Based Syst.* **2023**, *266*, 110391. [\[CrossRef\]](#)
- Luo, H.; Bo, L.; Liu, X.; Zhang, H. A novel method for remaining useful life prediction of roller bearings involving the discrepancy and similarity of degradation trajectories. *Comput. Intell. Neurosci.* **2021**, *2021*, 2500997. [\[CrossRef\]](#) [\[PubMed\]](#)
- SKF. *Rolling Bearings*; Publication No. 17000/1 EN; SKF Group: Gothenburg, Sweden, 2018. Available online: https://www.skf.com/binaries/pub12/Images/0901d196802809de-Rolling-bearings---17000_1-EN_tcm_12-121486.pdf (accessed on 2 September 2026).
- Lugt, P.M.; Berens, F. The grease life factor concept for ball bearings. *Tribol. Int.* **2022**, *169*, 107460. [\[CrossRef\]](#)
- Song, L.; Jin, Y.; Lin, T.; Zhao, S.; Wei, Z.; Wang, H. Remaining useful life prediction method based on the spatiotemporal graph and GCN nested parallel route model. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 3511912. [\[CrossRef\]](#)
- Wang, H.; Lin, T.; Cui, L.; Ma, B.; Dong, Z.; Song, L. Multitask learning-based self-attention encoding atrous convolutional neural network for remaining useful life prediction. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 3516108. [\[CrossRef\]](#)
- Zhu, J.; Chen, N.; Peng, W. Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Trans. Ind. Electron.* **2018**, *66*, 3208–3216. [\[CrossRef\]](#)
- Wang, H.; Zhang, X.; Guo, X.; Lin, T.; Song, L. Remaining useful life prediction of bearings based on multiple-feature fusion health indicator and weighted temporal convolution network. *Meas. Sci. Technol.* **2022**, *33*, 104003. [\[CrossRef\]](#)
- Lin, T.; Song, L.; Cui, L.; Wang, H. Advancing RUL prediction in mechanical systems: A hybrid deep learning approach utilizing non-full lifecycle data. *Adv. Eng. Inform.* **2024**, *61*, 102524. [\[CrossRef\]](#)
- Han, T.; Tian, J.; Wei, Y.M. Challenges and opportunities for battery health estimation: Bridging laboratory research and real-world applications. *J. Energy Chem.* **2024**, *89*, 434–436. [\[CrossRef\]](#)
- Fan, W.; Gao, X.; Han, L.; Xu, Z.; Chen, C.; Chen, L. Adaptive multistage sparse Gaussian process regression for bearing remaining useful life prediction. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3566010. [\[CrossRef\]](#)
- Spirito, M.; Melluso, F.; Nicoletta, A.; Malfi, P.; Cosenza, C.; Savino, S.; Niola, V. Rolling bearing fault detection using symmetrized dot pattern indices and feedforward neural network. *Struct. Health Monit.* **2026**. [\[CrossRef\]](#)
- Chen, J.; Li, T.; He, J.; Liu, T. An interpretable wavelet Kolmogorov–Arnold convolutional LSTM for spatial-temporal feature extraction and intelligent fault diagnosis. *J. Dyn. Monit. Diagn.* **2025**, *4*, 183–193. [\[CrossRef\]](#)
- Xu, Q.; Chen, Z.; Wu, K.; Wang, C.; Wu, M.; Li, X. KDnet-RUL: A knowledge distillation framework to compress deep neural networks for machine remaining useful life prediction. *IEEE Trans. Ind. Electron.* **2022**, *69*, 2022–2032. [\[CrossRef\]](#)
- Ma, Y.; Wang, Z.; Gao, J.; Chen, H. A novel method for remaining useful life of solid-state lithium-ion battery based on improved CNN and health indicators derivation. *Mech. Syst. Signal Process.* **2024**, *220*, 111646. [\[CrossRef\]](#)

21. Ye, G.; Feng, L.; Guo, J.; Chen, Y.; Li, S. A simple and effective KAN-based architecture for accurate battery RUL prediction. *IEEE Access* **2025**, *13*, 128700–128715. [[CrossRef](#)]
22. Ren, L.; Wang, T.; Jia, Z.; Li, F.; Han, H. A lightweight and adaptive knowledge distillation framework for remaining useful life prediction. *IEEE Trans. Ind. Inform.* **2022**, *19*, 9060–9070. [[CrossRef](#)]
23. Yan, J.; He, Z.; He, S. Multitask learning of health state assessment and remaining useful life prediction for sensor-equipped machines. *Reliab. Eng. Syst. Saf.* **2023**, *234*, 109141. [[CrossRef](#)]
24. Wu, J.Y.; Wu, M.; Chen, Z.; Li, X.; Yan, R. A joint classification-regression method for multi-stage remaining useful life prediction. *J. Manuf. Syst.* **2021**, *58*, 109–119. [[CrossRef](#)]
25. Wang, G.; Xiang, J. Remaining useful life prediction of rolling bearings based on exponential model optimized by gradient method. *Measurement* **2021**, *176*, 109161. [[CrossRef](#)]
26. Li, N.; Lei, Y.; Lin, J.; Ding, S.X. An improved exponential model for predicting remaining useful life of rolling element bearings. *IEEE Trans. Ind. Electron.* **2015**, *62*, 7762–7773. [[CrossRef](#)]
27. Peng, B.; Jin, X.; Li, D.; Zhou, S.; Wu, Y.; Liu, J.; Zhang, Z.; Liu, Y. Correlation congruence for knowledge distillation. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 5006–5015. [[CrossRef](#)]
28. Liu, Y.; Cao, J.; Li, B.; Hu, W.; Ding, J.; Li, L.; Maybank, S. Cross-architecture knowledge distillation. *Int. J. Comput. Vis.* **2024**, *132*, 2798–2824. [[CrossRef](#)]
29. Wang, B.; Lei, Y.; Li, N.; Li, N. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Trans. Reliab.* **2018**, *69*, 401–412. [[CrossRef](#)]
30. Nectoux, P.; Gouriveau, R.; Medjaher, K.; Ramasso, E.; Chebel-Morello, B.; Zerhouni, N.; Varnier, C. PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *Proceedings of the IEEE International Conference on Prognostics and Health Management*, Denver, CO, USA, 18–21 June 2012; pp. 1–8.
31. Cao, W.; Meng, Z.; Li, J.; Wu, J.; Fan, F. A remaining useful life prediction method for rolling bearing based on TCN-transformer. *IEEE Trans. Instrum. Meas.* **2024**, *74*, 3501309. [[CrossRef](#)]
32. Cheng, X.; Chaw, J.K.; Sahrani, S.; Ang, M.C.; Gunasekaran, S.S.; Mahmoud, M.A.; Zaman, H.B.; Zhao, Y.; Ren, F. An adaptive dual distillation framework for efficient remaining useful life prediction. *Complex Intell. Syst.* **2025**, *11*, 253. [[CrossRef](#)]
33. Mo, Y.; Wu, Q.; Li, X.; Huang, B. Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit. *J. Intell. Manuf.* **2021**, *32*, 1997–2006. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.