

# 基于卷积白盒 Transformer 的滚动轴承剩余寿命预测<sup>①</sup>



张宇, 孙渝林, 居文军

(山东科技大学 计算机科学与工程学院, 青岛 266590)

通信作者: 居文军, E-mail: [juwj@sdust.edu.cn](mailto:juwj@sdust.edu.cn)

**摘要:** 滚动轴承的振动信号具有非线性和非平稳性. 为增强剩余寿命预测方法对长时间依赖性与局部退化信息的同步捕获能力, 提出了一种结合卷积结构的白盒 Transformer (convolutional white-box Transformer, CWTR) 轴承剩余寿命预测模型. 首先, 设计融合膨胀因果卷积的子空间注意力机制, 以扩展注意力机制的感受野, 增强信号中局部依赖关系的建模能力; 其次, 构建多尺度卷积模块, 增强不同时间尺度下通道特征的交互建模能力, 从而更精细地提取不同退化阶段的局部特征; 此外, 基于 Pearson 相关系数量化评估轴承健康状态; 最后, 采用改进损失函数优化网络训练. 在真实轴承数据集上进行实验, 并与其他预测模型的预测结果进行比较, 均方根误差和平均绝对误差分别改进了 27.88% 与 27.85%, 验证了 CWTR 模型的有效性.

**关键词:** 剩余使用寿命; 滚动轴承; 卷积神经网络; 白盒 Transformer; Pearson 相关系数

引用格式: 张宇, 孙渝林, 居文军. 基于卷积白盒 Transformer 的滚动轴承剩余寿命预测. 计算机系统应用, 2025, 34(11): 242-252. <http://www.c-s-a.org.cn/1003-3254/10010.html>

## Remaining Useful Life Prediction for Rolling Bearings Based on Convolutional White-box Transformer

ZHANG Yu, SUN Yu-Lin, JU Wen-Jun

(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

**Abstract:** Vibration signals from rolling bearings exhibit nonlinear and non-stationary characteristics. To improve the ability of remaining useful life (RUL) prediction methods in simultaneously capturing long-term dependency and local degradation information, this study proposes a convolutional white-box Transformer (CWTR) model for RUL prediction of rolling bearings. Firstly, a subspace attention mechanism integrating dilated causal convolution is designed to expand the receptive field of the attention mechanism and enhance the modeling ability of local dependency relationships in signals. Secondly, a multi-scale convolutional module is constructed to improve the interactive modeling ability of channel features under different time scales, allowing for finer extraction of local features at different degradation stages. Additionally, a Pearson correlation coefficient-based method is introduced to quantitatively assess the health status of bearings. Finally, an improved loss function is applied to optimize network training. Experiments are conducted on the real bearing dataset and the prediction results are compared with those of other prediction models. The root mean square error and mean absolute error are improved by 27.88% and 27.85% respectively, verifying the effectiveness of the CWTR model.

**Key words:** remaining useful life (RUL); rolling bearing; convolutional neural network (CNN); white-box Transformer; Pearson correlation coefficient

① 收稿时间: 2025-04-24; 修改时间: 2025-05-15, 2025-06-03; 采用时间: 2025-06-09; csa 在线出版时间: 2025-09-30  
CNKI 网络首发时间: 2025-10-09

滚动轴承是被广泛应用在工业设备旋转机械<sup>[1]</sup>中的关键机械部件,其健康状态会对设备的运行效率和安全性产生直接的影响.随着长时间运行和高负载工作,轴承的性能不可避免地会随着时间的推移逐渐恶化<sup>[2]</sup>,严重时甚至可能导致设备停机乃至人员伤亡.故障预测与健康管理的(prognostic and health management, PHM)<sup>[3,4]</sup>包括异常状态检测、故障诊断、剩余使用寿命(remaining useful life, RUL)预测等方法,通过对设备或系统进行监测与分析,以预测可能出现的故障并进行设备维护.对滚动轴承进行RUL预测作为其中的关键技术之一,有助于实现有效的预测性维护,从而进一步保证工业设备的可靠性与安全性.

近年来,随着高精度传感器技术与工业物联网技术的快速发展,工业设备运行数据的采集与存储能力得到了显著提升.然而,滚动轴承作为旋转机械的核心部件,其运行过程中产生的监测信号具有典型的非平稳性与突变性<sup>[5]</sup>.这种复杂的信号特征使得传统方法难以有效提取设备退化状态的关键信息,导致大量监测数据中的退化模式信息未被充分挖掘与利用.在此背景下,融合统计学习与人工智能的数据驱动方法,因其能从复杂监测数据中自动学习退化规律<sup>[6-8]</sup>,而受到广泛关注.

数据驱动方法的发展推动了滚动轴承剩余寿命(RUL)预测技术的革新,其中基于Transformer的深度神经网络表现尤为突出<sup>[9]</sup>.自注意力机制使Transformer架构在序列建模中具备两方面优势:一是通过全局依赖建模克服了传统循环结构在长序列处理中的性能瓶颈;二是凭借高度并行的计算结构,在提升建模能力的同时加速了模型训练.为优化对时序数据的特征提取,Liu等人<sup>[10]</sup>设计了金字塔注意力模块,通过多尺度特征融合在保证 $O(L)$ 时空复杂度的前提下显著提升了长时依赖建模能力.Peng等人<sup>[11]</sup>设计了一种卷积自注意力机制并将时间卷积网络(TCN)与挤压激励(SE)注意力机制融合,通过改造Transformer中的注意力机制提升了其建模能力.然而,Transformer本身的“黑盒”特性导致特征提取过程不够透明,存在缺乏可解释性等问题.Yu等人<sup>[12]</sup>通过对稀疏率下降这一目标进行逐步优化,推导出来了一个在数学上完全可解释的白盒Transformer深度网络架构.尽管该方法目前仅在图像领域得到验证,但其模块化、可推理结构为工业时序数据建模的可解释性迁移提供了启发.虽然自注意力机制

在时序数据全局建模方面展现出显著优势,但在轴承剩余寿命预测这一特殊场景中仍存在关键性挑战:一方面需要保持对轴承平稳运行阶段长时依赖关系的建模能力,另一方面又必须精确捕捉衰退期局部退化特征的非线性突变.针对这一双重需求,本文在可解释Transformer架构<sup>[12]</sup>的基础上,提出了一种创新的卷积白盒Transformer模型.

该模型的核心改进体现在3个层面:首先,通过引入膨胀因果卷积重构多头子空间注意力机制,在保持数学模型可解释性的前提下,显著增强了局部退化特征与时间序列特征的建模能力;其次,设计了一种具有全透明特性的多尺度卷积模块并将其嵌入注意力层后,强化时间步间通道特征融合;最后,本文还将非线性轴承退化模型与改进损失的函数结合,使网络训练过程更贴合轴承实际退化规律.在PHM2012标准数据集上的实验验证表明,所提模型不仅保持了全局依赖关系的建模优势,同时也具备准确识别滚动轴承状态退化拐点的能力.

## 1 RUL 预测方法

### 1.1 轴承健康状态划分

滚动轴承在正常工作中大致可以分为稳定运行状态与加速衰退状态.文献<sup>[13]</sup>通过实验观测发现滚动轴承在稳定状态下的时间序列自相关系数要显著高于状态转变阶段的相关系数,这一现象为基于相关性分析的状态监测提供了理论依据.然而,Pearson相关系数在面对非平稳信号时容易产生虚假相关性,影响监测结果的可靠性.为增强模型对非平稳特征的表达能力,本文引入奇异值分解(SVD)方法作为特征预处理步骤,以提取信号中的主要变化趋势并削弱高频噪声干扰.

在预处理阶段,首先通过奇异值分解(SVD)对数据进行处理,得到奇异值矩阵.

$$X = U\Sigma V^T \quad (1)$$

选取前 $k$ 个奇异值作为单一时间步的预处理信号表征,计算时间序列的相关系数.

$$s_t = \text{diag}(\Sigma_t)_{1:k} \in R^k \quad (2)$$

我们将起始时刻的奇异值序列设为稳定状态参考值 $s_{\text{ref}}$ ,接着计算每一时刻与起始状态的相关系数.当相关系数 $\rho$ 下降至预设阈值时,判断找到退化点.

$$\begin{aligned} \rho(t) &= \frac{cov(s_t, s_{ref})}{\sigma_{s_t} \cdot \sigma_{s_{ref}}} \\ &= \frac{\sum_{i=1}^k (s_{t,i} - \bar{s}_t)(s_{ref,i} - \bar{s}_{ref})}{\sqrt{\sum_{i=1}^k (s_{t,i} - \bar{s}_t)^2} \sqrt{\sum_{i=1}^k (s_{ref,i} - \bar{s}_{ref})^2}} \end{aligned} \quad (3)$$

### 1.2 模型总体架构

该模型主要由3个核心部分构成:白盒 Transformer 基础架构、多头子空间卷积自注意力模块和多尺度卷积模块,整体架构如图1所示。

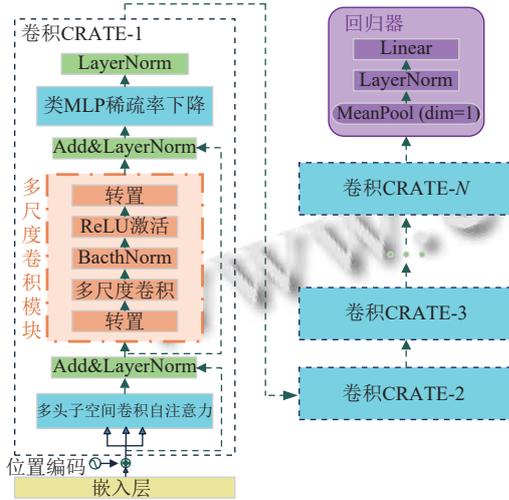


图1 CWTR 网络结构

嵌入层引入位置编码以弥补注意力机制的时序信息缺失。主循环块包含3个核心组件:多头子空间卷积自注意力模块通过扩展注意力机制的感受野,增强对信号中局部依赖关系的建模能力;多尺度卷积模块通过增强不同时间尺度下通道特征的交互能力,更精准地提取不同退化阶段的局部特征;ISTA (iterative shrinkage-thresholding algorithm) 算法作为白盒基础架构的一部分,用来增强稀疏表示。为优化梯度传播效果与提升训练稳定性,模块间采用残差连接与层归一化策略。最终,回归器通过在时间维度上的平均池化操作对全局特征进行聚合,并输出滚动轴承的剩余寿命预测值。

### 1.3 白盒 Transformer 基础架构

CRATE (coding rate reduction Transformer) 作为首个在结构设计中具备数学可解释性的 Transformer 架构,其核心思想是以稀疏率递减为优化目标,引导注意力分布自适应地聚焦于关键特征区域,从而实现特征空间的可控重构与表达优化<sup>[14]</sup>。假设原始数据为  $X = [x^1, x^2, \dots, x^N] \in R^{D \times N}$ , CRATE 的目的就是将其转换为更加线性、更加紧凑的特征  $Z = [z^1, z^2, \dots, z^N] \in R^{d \times N}$ ,

其中,  $d < D$ 。对  $X$  的压缩可用式(4)表示。

$$\begin{aligned} \max E_Z [\Delta R(Z; U_{[k]}) - \lambda \|Z\|_0] = \\ \max E_Z [R(Z) - R^c(Z|U_{[k]}) - \lambda \|Z\|_0] \end{aligned} \quad (4)$$

其中,  $Z = f(X)$ ,  $\lambda$  为稀疏化正则项,编码率  $R(Z)$  是使用高斯码本编码特征  $z$  到精度  $\varepsilon$  (保证失真率在  $\varepsilon$  之内) 所需要的平均比特数(数据量),而  $R^c(Z|U_{[k]})$  则表示将特征  $z$  投影到每个子空间上,达到精度  $\varepsilon$  所需的平均比特数。  $U_{[k]}$  为构成  $Z$  集合的低维子空间。该架构通过两阶段交替优化实现,这两部分分别对应了优化目标(1)中的两项:最小化编码率项  $R^c(Z|U_{[k]})$  和最小化稀疏性惩罚与展开项的差  $\lambda \|Z\|_0 - R(Z)$ 。

#### 1.3.1 压缩:多头子空间自注意力 (multi-head subspace self-attention, MSSA)

采用多头子空间自注意力 (MSSA) 将输入数据投影到  $K$  个低维子空间,压缩对应于优化目标中的  $R^c(Z|U_{[k]})$ 。过程描述为式(5)。其中  $N$  为特征维度,  $P$  为子空间维度。这里的 MSSA 与标准 Transformer 中的多头自注意力算子基本类似,只是这里  $QKV$  的线性算子都被设置为与子空间基相同,即  $Q = K = V = U_{[k]}^*$ 。

$$\begin{aligned} MSSA(Z|U_{[k]}) \doteq \frac{P}{(N+1)\varepsilon^2} [U_1, \dots, U_k] \\ \left[ \begin{array}{c} (U_1^*Z) \text{Softmax}((U_1^*Z)^*(U_1^*Z)) \\ \vdots \\ (U_k^*Z) \text{Softmax}((U_k^*Z)^*(U_k^*Z)) \end{array} \right] \end{aligned} \quad (5)$$

中间产物的表示结果  $Z^{l+1/2} \in R^{d \times N}$  则可以由式(6)表示。其中  $\kappa > 0$ , 为可供学习的超参数。

$$\begin{aligned} Z^{l+1/2} &= Z^l - \kappa \nabla_z R^c(Z^l|U_{[k]}) \\ &\approx \left( 1 - \kappa \cdot \frac{P}{(N+1)\varepsilon^2} \right) Z^l + \frac{P}{(N+1)\varepsilon^2} MSSA(Z^l|U_{[k]}) \end{aligned} \quad (6)$$

#### 1.3.2 稀疏:迭代收缩阈值算法 (ISTA)

稀疏对应了优化目标当中的  $\lambda \|Z\|_0 - R(Z)$ , 在这里我们引入一个可供学习的全局正交字典  $D^l$  来对  $Z^{l+1/2}$  进行稀疏编码<sup>[12]</sup>。

$$\begin{aligned} Z^{l+1} &= ReLU(Z^{l+1/2} + \eta D^l (Z^{l+1/2} - D^l Z^{l+1/2} - \eta \lambda \mathbf{1})) \\ &\doteq ISTA(Z^{l+1/2}|D^l) \end{aligned} \quad (7)$$

其中,  $\lambda$  为稀疏化正则项,  $\eta$  为步长,  $\mathbf{1}$  为全 1 向量。结合 MSSA 块与 ISTA 块,我们就得到了单个 CRATE 定义框架,一个完全数学可解释的白盒 Transformer 架构。

$$\begin{cases} Z^{l+1/2} \doteq Z^l + MSSA(Z^l|U_{[k]}^l) \\ Z^{l+1} \doteq ISTA(Z^{l+1/2}|D^l) \end{cases} \quad (8)$$

值得注意的是, CRATE 的数学可解释性主要强调其架构在设计上的结构透明性与可还原性. 但在具体领域落地当中为了体现实际预测任务中的可解释性, 仍需要结合注意力分布、通道响应等手段进行辅助可视化分析.

### 1.4 多头子空间卷积自注意力模块

膨胀因果卷积 (dilated causal convolution, DCC) 结合了因果卷积和膨胀卷积的特性, 既保持了时间序列的因果性, 又通过引入扩张因子扩展了感受野. 与标准卷积不同, 该结构在时序因果约束下引入了扩张因子  $d$ , 使得卷积核以指定间隔采样特征点, 从而显著扩大感受野. 如图 2 所示, 4 层 DCC 结构中, 每层卷积核大小均为 2, 扩张因子依次为 1、2、4. 图中呈现的因果特性意味着, 下一层在某一时刻的特征只能接收当前时刻及其之前时间步的特征. 为保持时序一致性, 在 Padding 操作中仅对序列左侧执行填充.

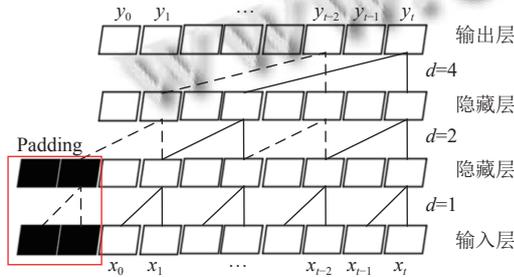


图 2 膨胀因果卷积

多头子空间自注意力机制中, 传统方法通过逐点线性变换生成查询向量  $Q$ 、键向量  $K$  和值向量  $V$ , 但该方法未能显式融合邻近采样点的局部结构信息, 导致局部时序特征建模能力受限. 为此, 我们将逐点自注意力机制替换为卷积自注意力机制. 如图 3 所示, 卷积注意力在生成  $Q$ 、 $K$ 、 $V$  这 3 个向量时采用了膨胀因果卷积. 该结构兼顾因果性与扩展感受野能力, 并能显著抑制未来时刻/突变点的干扰, 从而提升模型对非平稳时序数据的鲁棒性. 为保持白盒模型的可解释性, 卷积注意力中  $Q$ 、 $K$ 、 $V$  的变换矩阵仍与对应子空间保持一致, 具体表现为:

$$Conv1D_d^{(d)}(X) = Q_i = K_i = V_i = U_i^* \quad (9)$$

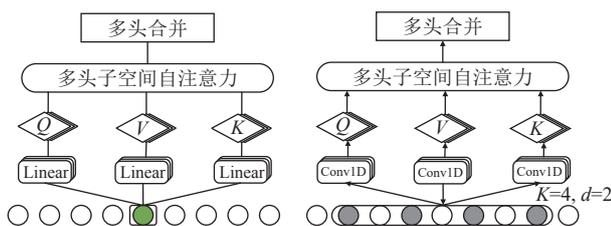


图 3 常规注意力机制与卷积注意力机制

结合式 (5) 后, 便可表示为:

$$MSSA(X|U_{[k]}) \doteq \frac{P}{(N+1)\varepsilon^2} [U_1, \dots, U_k] \begin{bmatrix} (V_1 X) Softmax((Q_1 X)^*(K_1 X)) \\ \vdots \\ (V_i X) Softmax((Q_i X)^*(K_i X)) \end{bmatrix} \quad (10)$$

### 1.5 多尺度卷积模块

为增强模型对轴承信号多尺度退化模式的建模能力, 本文设计了一种并行异构的一维卷积模块, 用于提取多尺度时间依赖特征. 具体地, 该模块结构中并行部署卷积核尺寸为 1、3、5 的卷积层, 以构建差异化感受野的特征采样路径. 其中, 较小卷积核专注于增强单一时间步的特征提取, 较大尺寸卷积核有助于建模时间步间的通道特征融合. 各卷积路径提取的特征随后在通道维度进行融合, 实现不同时间尺度下通道特征的交互建模, 从而提升网络对复杂信号结构的表达能力.

$$Output = torch.cat(Output_1, Output_2, Output_3, Output_4) \quad (11)$$

所设计的特征融合机制通过沿通道维度拼接各个卷积分支的输出, 形成覆盖多尺度时间依赖特征的联合表示. 为保证模块的结构可解释性和计算透明性, 所有卷积核尺寸与分支配置均为预设且固定. 此外, 为确保不同尺度卷积路径的输出在拼接前保持维度一致, 引入了一种尺寸对齐模块 (branch pool), 其在图中表示为  $D_1 + D_2 + D_3 + D_4 = D$ . 数据流动与维度变化如图 4 所示 (图示为批次为 1 的样例).

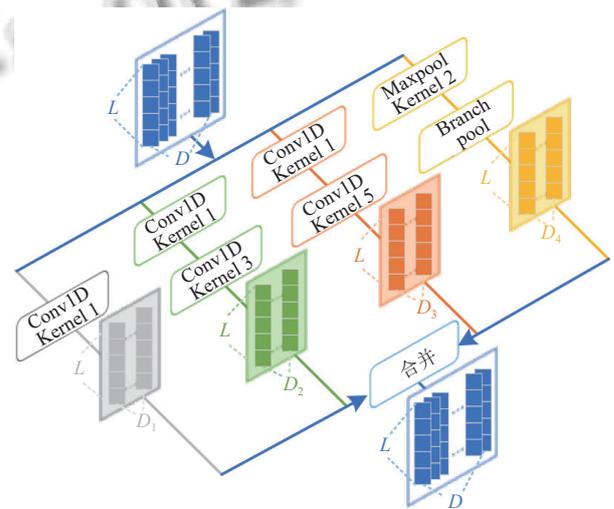


图 4 多尺度卷积

### 1.6 模型训练过程

如图 5 所示, 模型训练流程如下: 首先, 采用时频

域处理技术对原始信号进行预处理, 随后进行健康状态评估. 在此基础上, 划分训练集和测试集, 分别用于模型参数优化与性能评估. 其中, Weibull 损失函数是一种改进后的损失函数, 可将轴承故障演化信息有效地融合于模型训练过程中<sup>[15,16]</sup>, 表达式如式 (12) 所示.

$$L_{\text{weibull}}(t, \hat{t}, T, \hat{T}) = \frac{1}{n} \sum_{i=1}^n (t_i - \hat{t}_i)^2 + \lambda \frac{1}{n} \sum_{i=1}^n (F(T_i) - F(\hat{T}_i))^2 \quad (12)$$

其中,  $t_i$  是真实标签,  $\hat{t}_i$  是预测值,  $T_i$  是实际轴承已使用时间,  $\hat{T}_i$  为预测轴承已使用时间,  $\lambda$  为权重占比超参数.

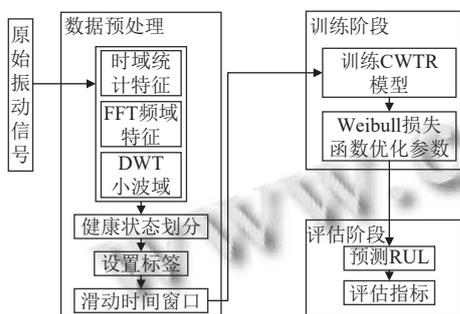


图5 寿命预测流程图

为更准确建模滚动轴承退化过程的非线性特征, 将 Weibull 损失函数与轴承健康状态划分方法相结合. 该损失函数能够在反向传播过程中动态调整网络对不同退化阶段的敏感性, 从而提升模型对平稳期与快速衰退期的区分能力, 使训练过程更贴合轴承的真实退化规律.

## 2 实验验证

### 2.1 数据集介绍

本文采用的 PHM2012 挑战赛数据集来验证所提出方法. 数据集由加速轴承退化平台 PRONOSTIA 对轴承加速降解并采样得来的<sup>[17]</sup>. 如图 6 所示, 该平台由旋转部分、负载部分与测量部分组成. 测量部分中, 采用加速度传感器按照 25.6 kHz 的采样频率对振动信号进行采样. 单次采样时间为 0.1 s 且采样间隔为 10 s, 这意味着间隔 10 s 采集一个有 2560 个点的样本.

实验平台设定的停止条件为轴承振动信号幅值超过 20g (即轴承完全失效, g 表示重力加速度). 因此, 数据集包括在 3 种恒定工况条件下轴承的全寿命振动信号, 具体包括:

工况 1: 转速 1 800 rpm, 横向负载 4 000 N, 对应的 7 个轴承 ( Bearing1\_1–Bearing1\_7);

工况 2: 转速 1 650 rpm, 横向负载 4 200 N, 对应 7 个轴承 ( Bearing2\_1–Bearing2\_7).

工况 3: 转速 1 500 rpm, 横向负载 5 000 N 的 3 个轴承 Bearing3\_1–Bearing3\_3.

加速退化实验会不可避免地引入一定的信号非平稳性, 如频谱的漂移和幅值波动. 这与实际工程场景中, 轴承由于设备载荷波动、润滑状态变化导致的非平稳类似. 为了降低这种变化对后续 RUL 预测的影响, 本文在数据预处理阶段采用奇异值分解来抑制高频噪声, 并通过模型设计增强对非平稳数据的建模能力, 从而缓解加速退化平台对模型泛化性能可能带来的负面影响.

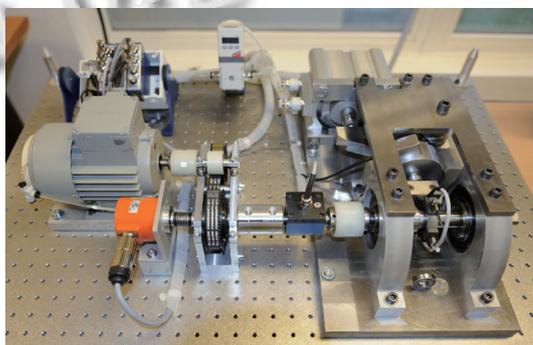


图6 PRONOSTIA 平台

与采用人工方式预设初始缺陷的轴承数据集相比, PRONOSTIA 数据集源于轴承在长期运行过程中发生的自然退化过程<sup>[17]</sup>. 例如, CWRU 数据集<sup>[18]</sup>通过电火花蚀刻的方式人工预设故障. 相比之下, PRONOSTIA 未在实验前人为植入缺陷, 因此采集到的振动信号更真实地反映了轴承在实际服役过程中的退化演化规律, 在统计特征与退化模式上更贴近工程场景下的自然损伤过程. 如图 7 所示为 Bearing1\_1 的振动信号曲线, 图中 g 表示重力加速度. 从图 7 中可观察其早期阶段信号幅值相对稳定, 随后逐渐出现幅值增大和波动增强的趋势, 呈现明显的平稳期与衰退期划分特征. 这种退化演化规律符合真实工程环境下轴承的运行特性, 有助于提升模型在实际场景中的应用效果.

为了验证模型的有效性与泛化性, 本文在不同工况条件下采用差异化的数据划分策略. 工况 1 为了充分评估模型性能, 依次选择 7 个轴承作为测试集, 其余作为训练集; 工况 2、工况 3 用来分析模型变工况条件下的适应性, 依据文献<sup>[19]</sup>数据集划分策略, 工况 2 选择 Bearing2\_1、Bearing2\_2、Bearing2\_4 作为测试集, 其余作为训练集; 工况 3 选择 Bearing3\_1 作为测试集,

其余作为训练集. 本文全部实验均通过 PyTorch 1.9 实现, 并运行于硬件条件为 Intel(R) Core(TM) i5-9300H CPU、8 GB RAM 和 NVIDIA GeForce GTX 1650 的计算机上.

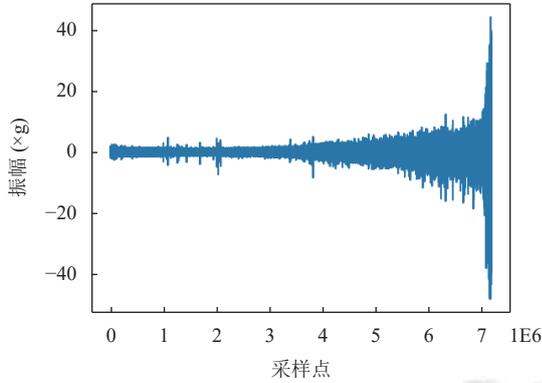


图7 Bearing1\_1 全寿命振动数据

### 2.2 分段寿命标签

由于数据集中缺乏详细物理失效信息, 同时指定振动信号幅值超过 20g 时轴承失效, 因此, 本文采用振动信号幅值的均方根 (root mean square, RMS) 作为衡量轴承健康状态的指标. 图 8 中两条曲线分别代表了 Bearing1\_1 平滑处理之后的 RMS 与 Pearson 相关系数. 从图中可以看出, 稳定状态下 Pearson 相关系数保持稳定, 但当 RMS 急剧上升时, 相关系数会迅速下降. 这表明了 Pearson 相关系数能够通过信号内部相关性变化来反映系统性能劣化. 值得注意的是, 该变化往往早于裂纹等可观测故障, 具备一定的早期预测能力, 有助于提高模型对早期退化的敏感性.

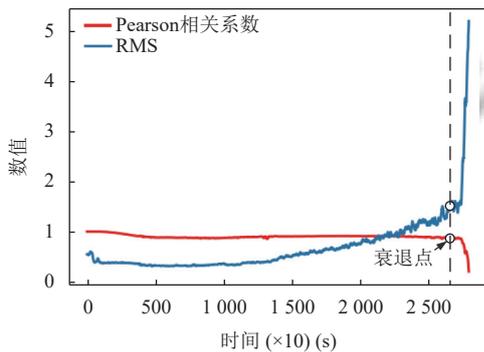


图8 Bearing1\_1 健康状态划分

为此, 本文将 Bearing1\_1 衰退点的触发阈值相关系数  $\rho=0.9$  设为 PHM2012 数据集中各轴承退化状态划分的统一标准. 如表 1 所示, 采用该阈值确定的各轴承退化起始点与当前先进的滚动轴承状态划分方法所得结论具有高度一致性<sup>[11]</sup>.

表 1 轴承衰退点

Bearing	衰退点	Bearing	衰退点
Bearing1_1	27400	Bearing2_3	4500
Bearing1_2	8270	Bearing2_4	7400
Bearing1_3	22430	Bearing2_5	4430
Bearing1_4	10870	Bearing2_6	6850
Bearing1_5	24100	Bearing2_7	2220
Bearing1_6	24140	Bearing3_1	4890
Bearing1_7	22110	Bearing3_2	1430
Bearing2_1	8700	Bearing3_3	3100
Bearing2_2	7210	—	—

因此, 我们可以将轴承的 RUL 标签分为两个部分, 其中  $T$  为轴承总寿命,  $T_p$  为衰退点时间,  $t$  为当前时间:

$$\text{稳定状态: } RUL_{\text{normal}}(t) = T - T_p.$$

$$\text{快速衰退状态: } RUL_{\text{rapid}}(t) = T - t.$$

为了保证模型具有良好的收敛速度与泛化能力, 我们将 RUL 归一到 [0, 1] 之间. 图 9 展示了 Bearing1\_4 的原始振动信号、Pearson 相关系数曲线以及归一化后的 RUL 寿命曲线, 以直观的展示状态划分的有效性. 如图所示, 在选取衰退点前后, 振动信号的幅值和波动程度明显上升, 呈现出典型的失效特征. 此外, 我们在多个轴承样本中观察到, 基于 Pearson 相关系数划分出来的不同健康状态与频域能量分布变化高度一致. 这种现象表明衰退点划分方法在信号层面与实际失效过程具备较强一致性.

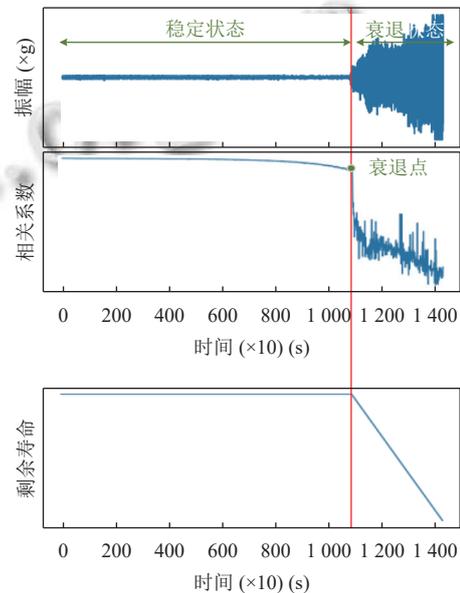


图9 Bearing1\_4 的健康状态划分与 RUL 标签

为进一步验证分界点在高维特征空间中的合理性, 本文引入 t-SNE (t-distributed stochastic neighbor embedding) 非线性降维方法<sup>[20]</sup>, 将高维特征映射至二维空间.

如图 10 所示, t-SNE 可视化结果划分点前后的样本在嵌入空间中呈现出明显的聚类差异. 这一聚类变化支持了健康状态划分在高维表示空间中的有效性, 进一步从几何角度印证了相关系数划分的合理性.

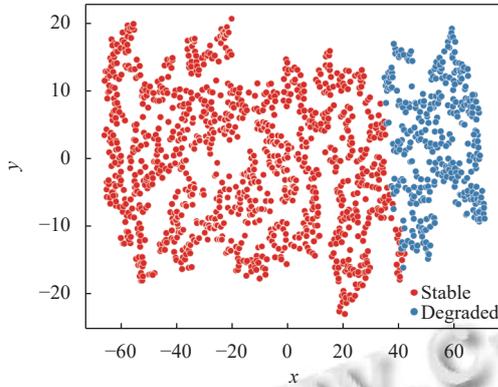


图 10 t-SNE 聚类降维

### 2.3 评价指标

为了更好地评估模型性能, 采用均方根误差 (root mean square error, *RMSE*)、平均绝对值误差 (mean absolute error, *MAE*) 以及评分函数用来衡量预测精度.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (13)$$

$$MAE = \frac{1}{N} |\hat{y}_i - y_i| \quad (14)$$

$$Score = \frac{1}{N} \sum_{i=1}^N A_i \quad (15)$$

其中,  $A_i = \begin{cases} e^{-\ln(0.5) \cdot (Er_i/5)}, & \text{if } Er_i \leq 0 \\ e^{\ln(0.5) \cdot (Er_i/20)}, & \text{if } Er_i > 0 \end{cases}$ ,  $Er_i = \frac{y_i - \hat{y}_i}{y_i} \times 100$ .

### 2.4 超参数设置

在网络搭建过程中, 我们采用网格搜索法来筛选合适的超参数, 并使用交叉验证进行公平地评估. 本文重点说明两个关键超参数的设计依据与优化过程: DCA 卷积核大小和主循环块层数.

1) 膨胀因果卷积核尺寸: 卷积核尺寸作为膨胀因果卷积注意力模块中的关键超参数, 对模型性能有显著影响. 本研究设置卷积核大小为 3、5、7 和 10, 并进行多次重复实验. 结果如图 11 所示, 最终选取性能最优的卷积核大小为 7. 实验表明, 适当的卷积核大小有助于提升注意力机制对局部退化信息的提取能力, 而过大或过小的设置则可能引起局部信息丢失, 降低预测精度.

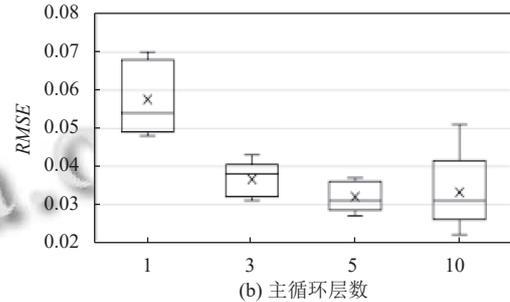
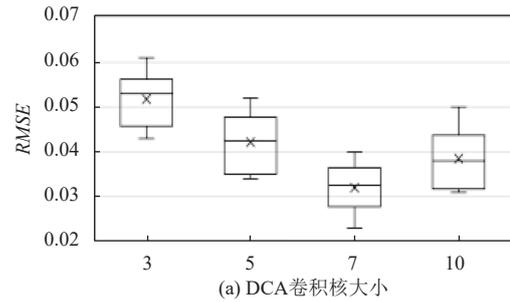


图 11 不同参数模型预测性能

2) 主循环块深度: 主循环块构成了网络的核心结构, 其层数设计对于提高 RUL 预测精度与降低计算成本具有重要意义. 本研究将循环块层数设置为 1、3、5 和 10, 并进行了多组重复实验, 最终确定层数为 5. 实验结果如图 11 所示. 结果表明, 增加模型深度虽可提升一定性能, 但盲目堆叠网络层数不仅带来额外的计算开销, 而且可能无法显著提升预测效果.

3) 网络的其他超参数我们通过表 2、表 3 来展示.

表 2 超参数列表

超参数	参数值	超参数	参数值
模型维度	192	循环周期	150
模型输入	[30, 177]	批大小	256
学习率	8E-4	丢弃率	0.1
优化器	Adam	注意力头数	4

表 3 网络模型参数

网络结构	网络层	参数	输出参数
Embedding	Input	177×192	Batch_size×30×192
	Embedding	1×30×192	Batch_size×30×192
DCA	QKV_conv	192×inner_size	Batch_size×30×inner_size
	Attention_out	inner_size×192	Batch_size×30×192
MSC	Conv_1	192×48	Batch_size×30×48
	Conv_5	(192×64), (64×48)	Batch_size×30×48
	Conv_3	(192×64), (64×64)	Batch_size×30×64
	Branch_pool	(192×96), (96×32)	Batch_size×30×32
ISTA	Grad	192×192	Batch_size×30×192
Regress	Mean_pool	30×1	Batch_size×1×192
	Output	192×1	Batch_size×1×1

### 3 实验结果与分析

#### 3.1 预测结果

我们选取 Bearing1\_1、Bearing1\_2、Bearing1\_3、Bearing1\_7、Bearing2\_1、Bearing3\_1 预测结果进行展示,如图 12 所示,具体数据在表 4.

在图 12 中可以看出,我们提出的方法在多数轴

承样本上均达到了预期的效果,成功区分了轴承的两种工作状态并精准的提取到轴承衰退点附近的局部信息.对于个别轴承预测误差较大的情况,主要原因是在于单一工况轴承之间数据差异性较大,且数据量较小,导致模型无法从训练集中学到充分的信息来实现预测.

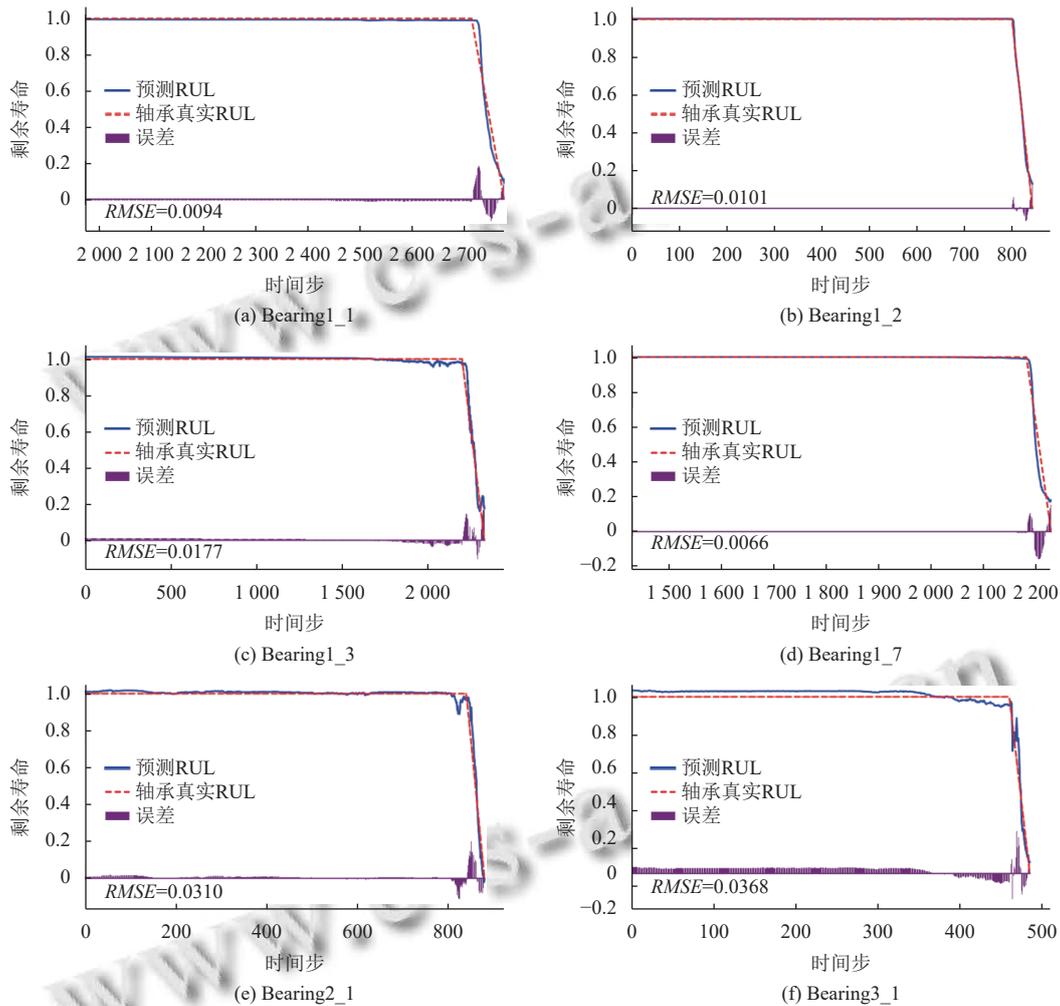


图 12 轴承预测结果

#### 3.2 预测过程可视化

为验证所提出模型对关键退化特征的提取能力及其在预测任务当中的可解释性,本文提取了卷积自注意力模块中每个子空间  $U_{[k]}$  与其伴随矩阵  $U_{[k]}^*$  的乘积(即为传统注意力中每一注意力头的输出得分),并对预测阶段的注意力权重进行可视化分析.如图 13 所示,我们选取了 Bearing1\_3 样本在退化阶段的第 1 个时间步和衰退点时间步附近的样本.在第 1 个样本中,轴承

处于早期运行阶段,故障特征不明显,因此注意力更倾向于掌握整体信息.而在衰退点附近的样本中,注意力为后半段时间步(即出现衰退点之后的退化阶段)赋予了更高的注意力权重,说明模型成功捕获了轴承状态变化的退化信息.

为了验证 CWTR 符合 CRATE 最初的设计目标,我们将优化目标 (1) 的过程结果进行可视化,其中包括压缩之后的  $R^c(z^{l+1/2})$  和稀疏之后的  $\|z^{l+1/2}\|$ .我们对划

分出的测试集和验证集的测试过程进行了可视化展示 (此处采用 10 层网络架构以增强可视化效果), 并对这

两个数据集在不同层数下的变化进行了对比 (展示的数据均来自单一批次).

表 4 实验结果对比

工况	测试轴承	COT <sup>[21]</sup>		MTCT <sup>[11]</sup>		CDCT <sup>[19]</sup>		本文方法	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
工况1	Bearing1_1	0.0134	0.0125	0.0229	0.0222	0.0159	0.0129	<b>0.0094</b>	<b>0.0075</b>
	Bearing1_2	0.0167	0.0123	0.0363	0.0323	0.0648	0.0398	<b>0.0101</b>	<b>0.0064</b>
	Bearing1_3	0.0199	0.0152	0.0506	0.0418	0.0329	0.0236	<b>0.0174</b>	<b>0.0130</b>
	Bearing1_4	0.0423	0.0356	0.1237	0.1179	0.0659	0.0535	<b>0.0416</b>	<b>0.0335</b>
	Bearing1_5	0.0197	0.0140	0.0385	0.0245	0.0377	0.0211	<b>0.0143</b>	<b>0.0105</b>
	Bearing1_6	0.0112	0.0069	0.0365	0.0260	0.0252	0.0097	<b>0.0105</b>	<b>0.0075</b>
	Bearing1_7	0.0170	0.0125	0.0391	0.0300	0.0618	0.0572	<b>0.0066</b>	<b>0.0036</b>
工况2	Bearing2_1	0.0529	0.0269	0.0566	0.0515	0.0459	0.0225	<b>0.0310</b>	<b>0.0210</b>
	Bearing2_2	0.0499	0.0294	0.0507	0.0470	0.0556	0.0454	<b>0.0382</b>	<b>0.0215</b>
	Bearing2_4	0.0338	0.0242	0.0564	0.0525	0.0755	0.0553	<b>0.0233</b>	<b>0.0169</b>
工况3	Bearing3_1	0.0550	0.0499	0.0702	0.0680	0.0647	0.0532	<b>0.0368</b>	<b>0.0313</b>

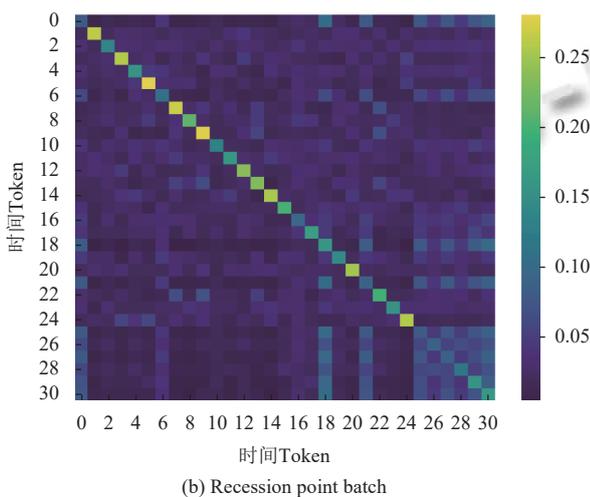
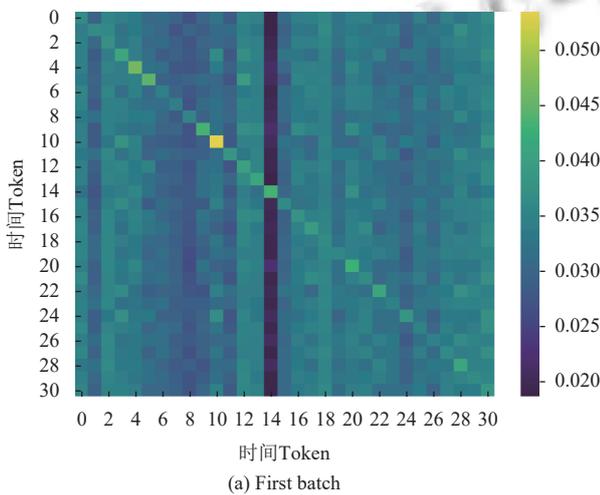


图 13 注意力权重可视化

图 14 显示了模型上述两个指标在不同层数之间的变化过程, 我们发现压缩项 (a) 与稀疏化项 (b) 随着

层数的增加而下降, 与优化目标相吻合. 这意味着网络通过一系列压缩、卷积与稀疏操作学习到了原始数据中蕴含的信息, 并以此建立了层数与优化目标的线性关系. 同时可以看到, 在 5 层之后压缩项的下降趋势开始平缓, 网络的学习能力提升较慢, 这进一步解释了为何当主循环层数超过 5 时, 优化效果并不明显.

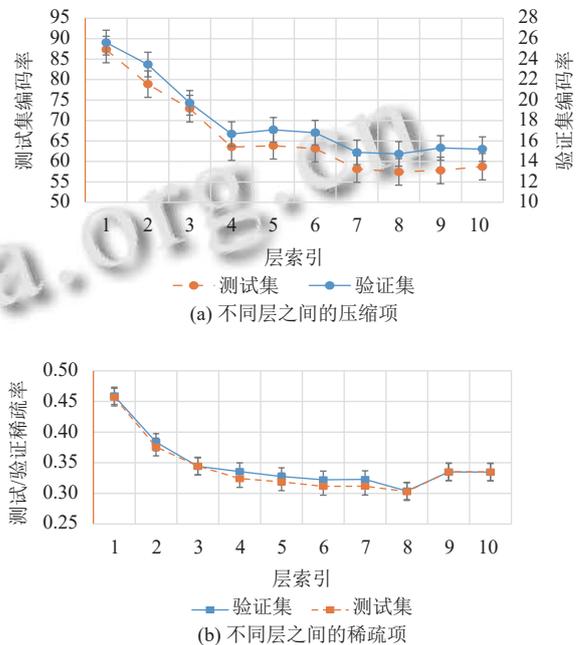


图 14 网络优化目标可视化

同时, 为了更好地理解网络中数据的分布形式, 我们以 Bearing1\_3 为例, 将 1、5、10 层的最终输出  $z^{l+1}$  的热力图 (选取前 50 列特征) 进行展示.

如图 15 所示, 数据特征变得越来越稀疏且具结构

化. 这同样表明了数据正被卷积与注意力机制逐渐压缩至更加结构化的子空间  $U_{[k]}$  上, 符合网络的设计初衷.

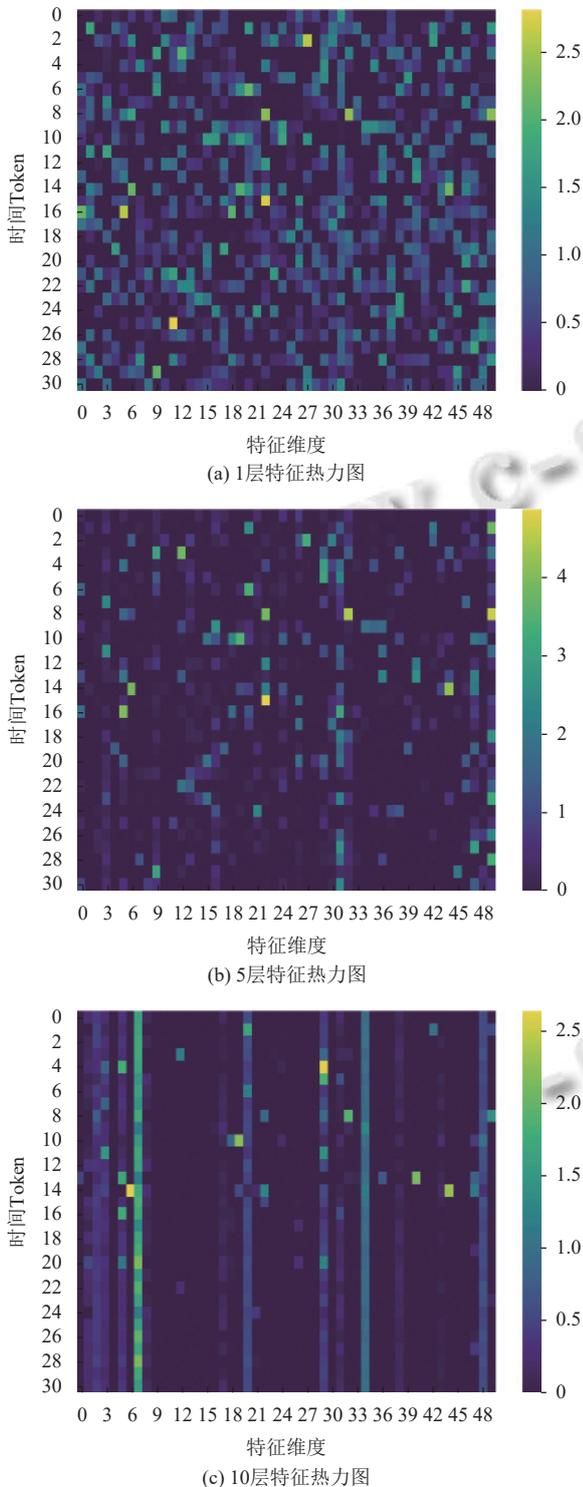


图 15  $z^{l+1}$  数据热力图

### 3.3 结果对比与消融实验

为了进一步表示提出网络的性能, 在相同训练环

境下将 MTCT<sup>[11]</sup>、COT<sup>[21]</sup>、CDCT<sup>[19]</sup>等先进预测方法与本文提出方法进行比较. 同时为了评估模型在不同工况场景下的适应能力, 我们分别在工况 1 (高转速、低负载)、工况 2 (中转速、中负载) 和工况 3 (低转速、高负载) 上进行对比实验, 结果详见表 4.

与 COT 的比较结果表明, 膨胀因果卷积拓宽了注意力机制的感受野并削减了局部噪声的干扰, 增强了对 RUL 的预测精度. 与 MTCT 与 CDCT 的比较结果表明, 多尺度卷积模块在跨通道特征的空间提取上表现优异, 优于过度依赖时间卷积网络 (TCN) 结构的时间注意力机制以及并联卷积注意力模型. 最后, 在各种工况下, 所提出的方法在 RMSE 与 MAE 上均取得了最优效果, 充分验证了方法的有效性与泛化能力.

此外, 我们还探究了模型中各个子模块对模型 RUL 预测的影响. 相较于基础的 CRATE 模型, 我们做出了膨胀因果卷积注意力 (DCA)、多尺度卷积 (MSC) 等改进, 此外首次将 Weibull 损失函数与非线性轴承退化曲线结合. 消融实验包括: 1) DCA 对模型的影响; 2) MSC 对模型的影响; 3) Weibull 损失函数对模型的影响.

表 5 展示了不同消融模块带来的结果. 从表中可知, 在相同的训练条件下, 3 种改进方案都增强了网络的预测性能. 三者的组合方法取得了最优的预测结果, 表明各个子模块在轴承 RUL 预测中均发挥了重要且积极的作用.

表 5 消融实验

模块	RMSE	MAE	Score
本文方法	<b>0.0157</b>	<b>0.0117</b>	<b>0.9363</b>
-DCA	0.0181	0.0275	0.9034
-MSC	0.0298	0.0234	0.8848
-Weibull	0.0349	0.0268	0.8963

## 4 结束语

本文提出了一种卷积白盒 Transformer 的滚动轴承剩余寿命预测模型, 融合膨胀因果卷积注意力和多尺度卷积 (DCA-MSC) 技术. 膨胀因果卷积注意力机制拓宽了注意力机制的感受野并削减了噪声, 而多尺度卷积则增强了不同时间尺度下通道特征的交互建模能力, 这有助于更好地捕捉轴承的长时依赖性与局部退化特征. 此外, 本文将新型轴承状态评估方法与 Weibull 损失函数相结合, 优化了模型训练. 在 PHM2012 多工况数据集场景下的实验结果表明, 该方法相比 COT<sup>[21]</sup>、MTCT<sup>[11]</sup>和 CDCT<sup>[19]</sup>的 RMSE 分别降低了 27.88%、

58.85% 和 56.18%，充分证明了本文方法在轴承剩余寿命预测领域的优越性。

尽管所提出模型取得了良好预测效果，但由于不同轴承之间在结构参数、运行条件存在差异，导致其振动信号退化模型变化较大，这可能影响模型的泛化能力。未来的研究可通过引入迁移学习等技术，进一步提高模型在处理不同类型轴承数据时的鲁棒性。此外，所使用的 PHM2012 轴承数据集样本数量有限，且依托加速退化平台采集，存在一定的工程应用局限性。因此，下一步工作将致力于将所提方法应用于大规模轴承数据集与实际工业场景中，以全面评估模型的潜在性能。

### 参考文献

- 1 Zhao HM, Liu HD, Jin Y, *et al.* Feature extraction for data-driven remaining useful life prediction of rolling bearings. *IEEE Transactions on Instrumentation and Measurement*, 2021, 70: 3511910. [doi: [10.1109/TIM.2021.3059500](https://doi.org/10.1109/TIM.2021.3059500)]
- 2 陈志强, 陈旭东, de Oliveira JV, 等. 深度学习在设备故障预测与健康中的应用. *仪器仪表学报*, 2019, 40(9): 206–226.
- 3 Zhao MH, Kang M, Tang BP, *et al.* Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes. *IEEE Transactions on Industrial Electronics*, 2018, 65(5): 4290–4300. [doi: [10.1109/TIE.2017.2762639](https://doi.org/10.1109/TIE.2017.2762639)]
- 4 张晓东, 秦子轩, 李敏, 等. 基于多特征融合的航空发动机剩余寿命预测. *计算机系统应用*, 2023, 32(3): 95–103. [doi: [10.15888/j.cnki.csa.008958](https://doi.org/10.15888/j.cnki.csa.008958)]
- 5 Xiang S, Zhou JH, Luo J, *et al.* Cocktail LSTM and its application into machine remaining useful life prediction. *IEEE/ASME Transactions on Mechatronics*, 2023, 28(5): 2425–2436. [doi: [10.1109/TMECH.2023.3244282](https://doi.org/10.1109/TMECH.2023.3244282)]
- 6 周壮, 周凤. 基于 E2E Deep VAE-LSTM 的轴承退化预测应用研究. *计算机应用研究*, 2022, 39(7): 2091–2097. [doi: [10.19734/j.issn.1001-3695.2021.11.0676](https://doi.org/10.19734/j.issn.1001-3695.2021.11.0676)]
- 7 Lei YG, Li NP, Gontarz S, *et al.* A model-based method for remaining useful life prediction of machinery. *IEEE Transactions on Reliability*, 2016, 65(3): 1314–1326. [doi: [10.1109/TR.2016.2570568](https://doi.org/10.1109/TR.2016.2570568)]
- 8 黄晓玲, 周磊, 张德平. 基于特征融合和混类增强的深度学习滚动轴承故障诊断. *计算机系统应用*, 2022, 31(8): 345–353. [doi: [10.15888/j.cnki.csa.008630](https://doi.org/10.15888/j.cnki.csa.008630)]
- 9 Zollanvari A, Kunanbayev K, Akhavan Bitaghsir S, *et al.* Transformer fault prognosis using deep recurrent neural network over vibration signals. *IEEE Transactions on Instrumentation and Measurement*, 2021, 70: 2502011.
- 10 Liu SZ, Yu H, Liao C, *et al.* Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. *Proceedings of the 10th International Conference on Learning Representations*. OpenReview.net, 2022.
- 11 Peng HC, Jiang B, Mao ZH, *et al.* Local enhancing Transformer with temporal convolutional attention mechanism for bearings remaining useful life prediction. *IEEE Transactions on Instrumentation and Measurement*, 2023, 72: 3522312. [doi: [10.1109/TIM.2023.3291787](https://doi.org/10.1109/TIM.2023.3291787)]
- 12 Yu YD, Buchanan S, Pai D, *et al.* White-box Transformers via sparse rate reduction. *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2023. 413.
- 13 Mao WT, He JL, Zuo MJ. Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 2020, 69(4): 1594–1608. [doi: [10.1109/TIM.2019.2917735](https://doi.org/10.1109/TIM.2019.2917735)]
- 14 Yu YD, Chu TZ, Tong SB, *et al.* Emergence of segmentation with minimalistic white-box Transformers. *Proceedings of the 2024 Conference on Parsimony and Learning*. Cambridge: PMLR, 2024. 72–93.
- 15 Weibull W. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 1951, 18(3): 293–297. [doi: [10.1115/1.4010337](https://doi.org/10.1115/1.4010337)]
- 16 Von Hahn T, Mechefske C. Knowledge informed machine learning using a Weibull-based loss function. *Journal of Prognostics and Health Management*, 2022, 2(1): 9–44. [doi: [10.22215/jphm.v2i1.3162](https://doi.org/10.22215/jphm.v2i1.3162)]
- 17 Nectoux P, Gouriveau R, Medjaher K, *et al.* PRONOSTIA: An experimental platform for bearings accelerated degradation tests. *Proceedings of the 2012 IEEE International Conference on Prognostics and Health Management*. Denver: IEEE, 2012. 1–8.
- 18 Smith WA, Randall RB. Rolling element bearing diagnostics using the case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, 2015, 64–65: 100–131.
- 19 Jiang L, Zhang TA, Lei W, *et al.* A new convolutional dual-channel Transformer network with time window concatenation for remaining useful life prediction of rolling bearings. *Advanced Engineering Informatics*, 2023, 56: 101966. [doi: [10.1016/j.aei.2023.101966](https://doi.org/10.1016/j.aei.2023.101966)]
- 20 Chatzimparmpas A, Martins RM, Kerren A. t-visNE: Interactive assessment and interpretation of t-SNE projections. *IEEE Transactions on Visualization and Computer Graphics*, 2020, 26(8): 2696–2714. [doi: [10.1109/TVCG.2020.2986996](https://doi.org/10.1109/TVCG.2020.2986996)]
- 21 Ding YF, Jia MP. Convolutional Transformer: An enhanced attention mechanism architecture for remaining useful life estimation of bearings. *IEEE Transactions on Instrumentation and Measurement*, 2022, 71: 3515010. [doi: [10.1109/TIM.2022.3181933](https://doi.org/10.1109/TIM.2022.3181933)]

(校对责编: 张重毅)