计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(5):41−51 [doi: 10.15888/j.cnki.csa.009861] [CSTR: 32024.14.csa.009861]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多尺度特征融合的混合双路径 CNN-MLP 故障诊断模型①

丁传龙[1], 花国祥[1,2], 蒋　亮[2], 郭永信[2]

[1](南京信息工程大学 自动化学院, 南京 210044)
[2](无锡学院 自动化学院, 无锡 214105)
通信作者: 蒋　亮, E-mail: rezgh0214@163.com

摘　要: 滚动轴承在机械系统中至关重要, 低频率故障通常由于其发生概率低而导致数据样本稀缺, 这使得相关数据的采集和处理面临挑战, 若处理不当可导致严重的安全隐患和经济损失. 为应对这一问题, 本研究提出了一种结合传统信号处理方法与深度学习模型的卷积神经网络 (CNN) 与多层感知机 (MLP) 的双路径故障诊断模型. 特征工程提取方面, 本研究采用离散小波变换 (DWT) 和连续小波变换 (CWT) 相结合的方法, 结合平均下采样技术从原始信号中提取多尺度的时频特征和时域特征. 模型包含两条路径: 一条通过将 efficient channel attention (ECA) 注意力机制嵌入残差 CNN 中提取特征工程的时频特征, 另一条利用 MLP 处理下采样的多尺度时域特征, 最后融合两者进行分类. 小样本评估显示, 该特征工程方法在凯斯西储大学 (CWRU) 数据集上平均诊断准确率达到 99.34%, 高于传统方法的 98.97%. 混合 CNN-MLP 双路径模型在 CWRU 数据集上达到了 99.90% 的高准确率, 在江南大学 (JNU) 数据集上取得 98.38% 的准确率. 表明其在小样本滚动轴承故障诊断中的应用潜力.

关键词: 故障诊断; 特征融合; 小样本; 注意力机制; 双路径

引用格式: 丁传龙,花国祥,蒋亮,郭永信.基于多尺度特征融合的混合双路径 CNN-MLP 故障诊断模型.计算机系统应用,2025,34(5):41–51. http://www.c-s-a.org.cn/1003-3254/9861.html

Hybrid Dual-path CNN-MLP Fault Diagnosis Model Based on Multi-scale Feature Fusion

DING Chuan-Long[1], HUA Guo-Xiang[1,2], JIANG Liang[2], GUO Yong-Xin[2]

[1](School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Automation, Wuxi University, Wuxi 214105, China)

Abstract: Rolling bearings are crucial components in mechanical systems. As low-frequency faults are less likely to occur, data samples related to those are rare, bringing difficulties to the collection and processing of related data. If not properly addressed, such faults can lead to severe safety hazards and substantial economic losses. To deal with this problem, this study proposes a dual-path fault diagnosis model that integrates traditional signal processing methods with convolutional neural network (CNN) and multilayer perceptron (MLP). In terms of feature extraction, the study employs a combination of discrete wavelet transform (DWT) and continuous wavelet transform (CWT), along with average downsampling techniques, to extract multi-scale time-frequency and time-domain features from the raw signals. The model contains two paths: one extracts time-frequency features of feature engineering by embedding the efficient channel attention (ECA) mechanism into the residual CNN, and the other uses MLP to process down-sampled multi-scale time-domain features, and finally fuses the two paths for classification. Small sample evaluation shows that the feature engineering method achieves an average diagnostic accuracy of 99.34% on the Case Western Reserve University

① 基金项目: 江苏省高等学校自然科学基金 (22KJB460039)
收稿时间: 2024-10-25; 修改时间: 2024-11-18, 2024-12-09; 采用时间: 2024-12-18; csa 在线出版时间: 2025-03-31
CNKI 网络首发时间: 2025-04-01

(CWRU) dataset, which is higher than the 98.97% achieved by the traditional method. The hybrid CNN-MLP dual-path model achieves a high accuracy of 99.90% on the CWRU dataset and an accuracy of 98.38% on the Jiangnan University (JNU) dataset. It shows its application potential in small sample rolling bearing fault diagnosis.
Key words: fault diagnosis; feature fusion; small sample; attention mechanism; dual-path

滚动轴承作为关键部件, 广泛应用于各行各业, 若故障未能及时检测, 将可能导致严重的经济损失和安全风险[1]. 开发高效的故障诊断模型对于提升设备的可靠性、预防事故的发生、优化维护策略、减少停机时间和提高生产效率至关重要[2].

故障诊断技术最初依赖基于物理和工程原理的数学模型, 但其对精确模型的依赖性使得在实际应用中面临诸多挑战. 相比之下, 基于信号分析的方法无需依赖先验模型, 具备更高的灵活性[3]. 其中, 振动信号分析是旋转机械故障诊断中的关键方法, 主要包括信号分解、特征提取和模式识别. 信号分解旨在将周期性分量从噪声中分离出来, 但在随机噪声存在时, 故障信号可能被掩盖, 从而降低诊断准确性[4]. 传统的时域和频域分析主要适用于平稳信号, 难以揭示时间与频率之间的关系. 时频分析方法能够将非平稳信号分解为多个子信号, 从而获得更多的故障信息. 例如, 小波变换 (WT) 和经验模态分解 (EMD) 常被应用于振动信号的降噪和分解. 然而, EMD 及其变体易受到模式混合的影响, 从而导致分解性能下降[5]. 小波分解能够有效处理非线性、非平滑信号, 并具有多分辨率分析和局部特征表征的能力[6].

尽管故障诊断依赖于振动信号分析等信号处理技术, 并且较为有效, 但在处理复杂信号和噪声数据时, 仍然存在一定局限性[7]. 通过与传统机器学习技术的集成, 可对时频数据进行更复杂的分析, 从而提升对细微故障特征的检测效果. 自 2006 年 Hinton 等[8]提出深度学习以来, 故障诊断已成为一个备受关注的研究方向. 尤其是 CNN, 作为一种强大的自动特征提取和模式识别工具, 显著提升了故障预测的准确性. Chen 等[9]提出了一种结合多尺度 CNN 和 LSTM 的模型, 用于从多个频段提取特征并进行故障分类. 该模型通过减少参数数量, 提高在噪声环境下的诊断准确性, 克服了传统方法在特征提取和分类中的局限性. 尽管卷积神经网络 (CNN) 在自动特征提取和模式识别方面展现出强大的能力, 但在工业噪声环境和小样本条件下依然存在显著局限性. 由于工业数据通常伴随大量背景噪声, 且故障数据的采集成本高导致可用训练样本有限, 这使得 CNN 更难有效学习到代表性的特征[10]. CNN 的卷积操作主要依赖局部特征提取, 在数据稀缺时往往过度拟合于噪声或非代表性模式, 无法有效区分有用信号和干扰信号. 此外, 由于 CNN 在捕捉全局依赖关系方面存在一定局限, 其提取的特征往往缺乏全局视野. 因此, 当数据量不足且信号噪声较大时, CNN 容易陷入泛化能力不足的问题, 从而降低故障诊断的准确性和鲁棒性[11]. 董绍江等[12]提出了一种基于奇异值分解 (SVD) 和经验模态分解 (EMD) 的多层次降噪方法. 该方法首先通过 SVD 去除强噪声, 同时保留微弱的故障特征; 接着, 利用 EMD 提取关键模态分量; 最后, 将降噪后的信号输入带有注意力机制的改进卷积神经网络以进行特征学习和故障诊断. 鉴于轴承的失效可能需要数月甚至数年, 数据的收集通常需要长时间的等待[13]. 这些挑战可能限制 CNN 在故障诊断中的有效应用. 最近的研究表明, 注意力机制能够通过增强关键特征, 提高模型在小样本条件下的故障诊断性能[14]. 此外, MLP 通过处理数值数据 (包括噪声特征数据), 有效弥补了 CNN 在故障诊断中的局限性[15].

深度学习具备卓越的非线性特征挖掘能力, 可以在复杂工况下实现端到端的故障检测, 有效弥补了传统信号处理在复杂场景中的局限性. 结合信号处理的降噪和特征解析, 深度学习能进一步强化特征表达并提升诊断的可靠性和精确性[16]. 本研究旨在结合传统的信号处理技术与先进的深度学习技术, 设计一种适用于小样本数据集的高效且准确的故障诊断方法. 该方法融合了 DWT 和 CWT, 其中 DWT 用于多尺度分解, CWT 用于提供连续的时频特征分析. 同时, 通过对原始信号进行下采样, 获取多尺度的时域特征. 最终, 提出一种结合注意力机制的混合双路径 CNN-MLP 模型, 通过并行结构中集成 ECA 的残差 CNN 和 MLP 分别处理时频特征与时域特征, 并将其融合输出故障诊断结果.

1 理论基础

1.1 离散小波变换 (DWT)

DWT 使用小波缩放形成的一组基函数来将原始信号分解为不同尺度的子信号, 同时保留时间信息, 从而对信号的局部特征进行多尺度分析. 假设X是振动信号$x(t)$的时间序列, 正交小波分解可以由式 (1) 描述:

$$\begin{cases} c_{j,k} = \sum\limits_{n} c_{j-1,n}\, h_{n-2k} \\ d_{j,k} = \sum\limits_{n} c_{j-1,n}\, g_{n-2k} \end{cases} \quad (k = 0, 1, 2, \cdots, N-1) \tag{1}$$

其中, $c_{j,k}$表示尺度系数, $d_{j,k}$表示小波系数, h和g是一对正交镜像滤波器组 (QMF), j是分解层的数目, N是时间序列的数据点数目.

小波重构是小波分解的逆, 且对应的重构方程可以表示为式 (2):

$$c_{j-1,n} = \sum_{n} c_{j,n}\, h_{k-2n} + \sum_{n} d_{j,n}\, g_{k-2n} \tag{2}$$

如图 1 所示显示了对原始信号进行 4 层小波分解的结果, 得到近似信号 A_4 和一系列的细节信号 D_1、D_2、D_3 和 D_4. 因此, 分解后的信号与原始信号 X 的关系可以描述为:

$$X = A_4 + D_4 + D_3 + D_2 + D_1 \tag{3}$$

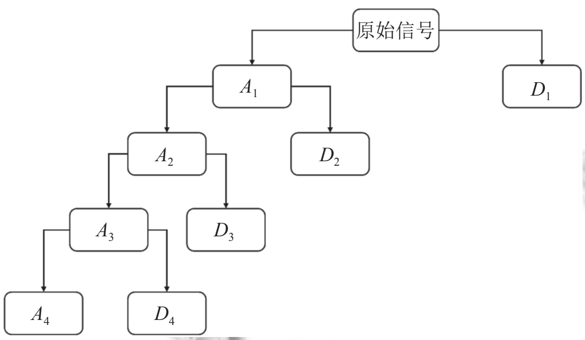


图 1 DWT 示意图

1.2 连续小波变换 (CWT)

CWT 使用随频率变化的时频窗口, 克服了短时傅里叶变换窗口固定的局限性, 能同时提取多尺度信息, 具备优异的时频局部化能力, 能检测振动信号中的故障奇异点, 从而提供更丰富的数据解释.

一个基本函数 $\phi(t)$ 通过平移和伸缩得到式 (4):

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\phi\left(\frac{t-b}{a}\right) \tag{4}$$

对于一个平方可积信号 $f(t)$, 其中 $f(t) \in L^2(\mathbb{R})$, 其 CWT 数学定义如下:

$$WT(\tau, a) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-\tau}{a}\right)\mathrm{d}t \tag{5}$$

其中, $L^2(\mathbb{R})$ 表示能量有限的信号空间. 尺度因子a 表示频率相关的缩放, 平移因子 τ 则决定了小波窗口在频率和时间域中的位置. 当时域信号被转换到尺度域后, 结果信号尺度通过式 (6) 转化为时频表示 TFR 图:

$$f_a = \frac{f_c f_s}{a} \tag{6}$$

其中, f_c 表示小波的中心频率, f_s 是信号的采样频, f_a 表示对应于尺度的实际频率. 为了确保转换后的频率序列成为等距序列, 尺度序列取以下值:

$$\left[totalscal, totalscal - \frac{totalscal-1}{n-1}, \cdots, 1\right] \tag{7}$$

尺度序列的长度 $totalscal$ 设为 256. 此外, 对应于尺度$c/totalscal$ 的实际频率为$f_s/2$, 常数 c 可通过式 (8) 计算得出:

$$c = 2 \times f_c \times totalscal \tag{8}$$

通常, 选择与信号故障特征密切相关的小波基函数是有利的. 复 Morlet 小波 (cmor 小波) 由于与轴承故障特征的冲击波形相似, 非常适合此用途. 因此, 我们选择了 cmor 小波.

1.3 卷积神经网络

CNN 由 LeCun 等[17]提出, 受大脑皮层感知机制启发, 广泛应用于图像和模式识别任务. 典型 CNN 架构如图 2 所示, 包括输入层、卷积层、池化层、全连接层和输出层. 卷积层提取特征, 池化层减少数据维度以防过拟合, 常用优化方法包括 *Dropout*、批量归一化和 Adam 优化器[18].

卷积层通过滤波器核, 对输入数据的局部区域进行卷积运算, 提取特征并通过激活函数进行非线性变换, 生成特征图. 滤波器在不同位置共享相同的权重, 称为权重共享, 减少了参数数量, 提高了训练效率. 卷积运算可以表示为:

$$Y_i^{(l)} = \sum K_{i,j}^{(l)} \cdot X_j^{(l)} + b_i \tag{9}$$

其中, $Y_i^{(l)}$表示第l层第i个滤波器输出特征图, $X_j^{(l)}$表示第l层第j个局部区域的数据输入, b_i为偏置.

近年来, 研究人员广泛探索了 CNN 在捕捉复杂任务特征方面的能力. 然而随着网络加深, 梯度消失问题

阻碍了深层网络的有效训练. 为解决此问题, 引入了残差连接, 如图 3 所示, 通过跳过层来学习残差映射, 增强了梯度流动, 促进了深层网络的训练, 提高了复杂任务的准确性.

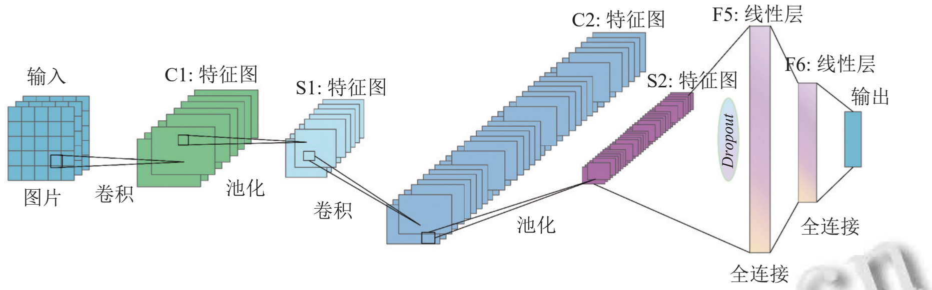


图 2　CNN 网络结构

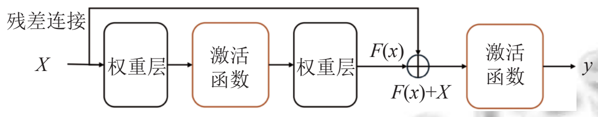


图 3　残差连接结构图

1.4　ECA 注意力机制模块

本文采用 ECA 机制优化模型性能, 如图 4 所示, 以降低 SE 模块计算复杂度, 同时保持性能. 传统 SE 模块通过两层全连接层捕捉通道间依赖, 尽管有效, 但增加了参数和计算开销, 尤其在大规模网络中. 相比之下, ECA 模块利用 1D 卷积实现通道间的局部交互, 避免降维操作, 大幅减少参数, 同时保留捕捉通道依赖的能力.

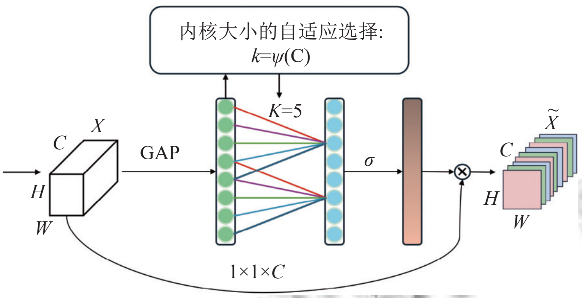


图 4　ECA 结构

与 SE 模块类似, ECA 模块首先需对输入特征图 $X \in \mathbb{R}^{C\times H\times W}$ 进行全局平均池化, 获得通道维度上的平均值向量 $y_c \in \mathbb{R}$:

$$y_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{c,i,j} \tag{10}$$

不同于 SE 模块的全连接层, ECA 通过 1D 卷积在通道维度上进行局部交互, 卷积核大小为 k, 可以动态调节. 这一操作显著减少了计算成本:

$$z = Conv1D(y,k) \tag{11}$$

经过 1D 卷积后, 使用 Sigmoid 函数对注意力权重 z_c 进行归一化处理:

$$a_c = s(z_c) = \frac{1}{1+\exp(-z_c)} \tag{12}$$

最终的输出特征图 $\widetilde{X}$ 通过将注意力权重应用到原始特征图上得到:

$$\widetilde{X}_{c,i,j} = a_c \cdot X_{c,i,j} \tag{13}$$

2　特征工程方法与模型建模

2.1　特征工程方法论

本文对进行归一化的长度为 1 024 的标准化一维信号进行了分解, 得到低频近似分量 (cA) 和高频细节分量 (cD). 近似系数代表信号的整体趋势和主结构, 保留了信号中的低频部分, 适合反映较缓慢变化的成分, 如设备的稳态运行特征. 细节系数则反映不同频带的高频信息, 通常对应信号中的快速变化或异常, 如故障信号、冲击和振动等. 通过 DWT, 可以有效将振动信号的高频与低频信息分离, 使我们能够专注于特定频率范围内的特征进行分析和处理. 本文采用了 Dobesie 小波, 这是一种正交小波, 支持信号的无损分解与重构, 确保各频率分量准确反映原始信号信息. 其多分辨率分析能力能够在不同尺度下捕捉信号的全局趋势和局部细节, 特别适合提取机械故障特征. 选择消失矩为 2 的小波, 旨在去除线性趋势, 保留关键故障特征, 同时保持较高的计算效率.

在故障诊断中, 振动信号往往包含各种频率的成分, 不同的故障类型通常对应于特定频率范围的特征, 轴承内圈或外圈故障则可能集中在某个特定的中频范围, 分解层数与信号的频率带宽和采样率相关. 假设信号的采样频率为 f_s, 每次分解后, 频率范围会减半, 因

此第 j 层分解对应的频率范围为:

$$f_j = \left[\frac{f_s}{2^{j+1}}, \frac{f_s}{2^j}\right] \tag{14}$$

例如, 经过 4 层小波分解后, 采样频率为 12 kHz 的信号可捕捉 0–750 Hz 的频率范围, 涵盖了关键的低频故障信号. 分解后将不同频率分量扩展至相同长度以确保特征平滑性和分布一致性, 增强模型对有限样本的泛化能力并减少过拟合. 随后对每个分量应用 CWT 生成时频图, 最后沿通道方向拼接输入诊断模型进行故障诊断, 流程结构图如图 5 所示.

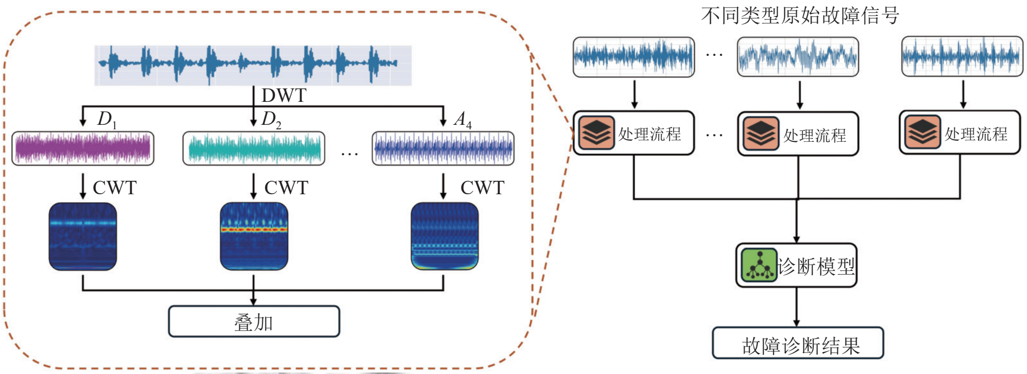


图 5　特征方法流程图

2.2　混合 CNN-MLP 双路径故障诊断模型

本文提出了一种混合 CNN-MLP 双路径模型, 如图 6 所示, 有效结合了 DWT-CWT 特征工程与多尺度原始信号分析. 在该架构中, 第 1 路径通过改进的 CNN, 捕捉了振动信号的多尺度时频特性. 同时, 第 2 路径采用 3 种不同的步长对原始信号进行平均下采样, 得到原始信号的多尺度时域特征. 接着利用 MLP 处理每个下采样的多尺度信号. 最后将两个路径得到特征进行融合输出. 通过 MLP 增强了模型对输入数据不同部分关键特征的关注能力, 从而在小样本条件下提升了性能.

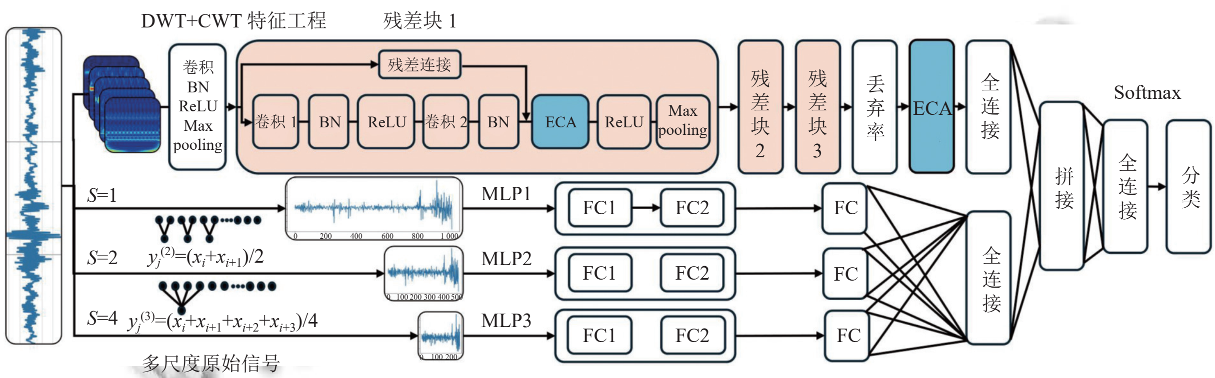


图 6　混合 CNN-MLP 双路径模型的结构

CNN 卷积层通过局部滤波器有效提取时频图像的局部特征, 但处理一维原始信号时, 故障特征稀疏分布于整个序列, CNN 易忽略跨时间段的关键信息. 同时, CNN 依赖大量训练数据来优化卷积核, 在小样本条件下易过拟合, 难以稳定提取普遍特征. 相比之下, MLP 全连接层一次性处理整个信号序列, 更适合全局特征建模, 能在数据稀缺情况下减少过拟合风险, 准确识别关键模式. 因此, 本文设计了双路径模型, 结合 CNN 提取时频特征与 MLP 建模时域特征, 有效捕获潜在故障信息. CNN 模型结合残差连接和 ECA 注意力, 用于处理特征工程信号, 作为混合 CNN-MLP 双路径模型的一个分支. 该模型有输入层、卷积层、残差层、归一化层、ReLU 层、池化层、全连接层、Softmax 层和输出层, 具体参数如表 1. 残差连接和 ECA 注意力的结合增强了梯度流, 实现了更深的网络, 并优先考虑关键特征, 从而显著提高了小样本环境下的性能和准确性.

表 1　CNN 网络参数

主模块名称	子模块	输入通道	输出通道	卷积核大小	步长	填充	补充参数
卷积2D	—	15	16	6×6	1	1	批量归一化、ReLU
最大池化	—	16	16	2×2	2	—	—
残差块1	卷积2D	16	32	3×3	1	1	批量归一化、ReLU
	卷积2D	32	32	3×3	1	1	批量归一化、ReLU
	残差连接	—	—	—	—	—	ReLU
	ECA block	—	—	—	—	—	ReLU
	最大池化	32	32	2×2	2	—	—
残差块2	卷积2D	32	64	3×3	1	1	批量归一化、ReLU
	卷积2D	64	64	3×3	1	1	批量归一化、ReLU
	残差连接	—	—	—	—	—	ReLU
	ECA block	—	—	—	—	—	ReLU
	最大池化	64	64	2×2	2	—	—
残差块3	卷积2D	64	128	3×3	1	1	批量归一化、ReLU
	卷积2D	128	128	3×3	1	1	批量归一化、ReLU
	残差连接	—	—	—	—	—	ReLU
	ECA block	—	—	—	—	—	ReLU
	最大池化	128	128	2×2	2	—	—
ECA block	—	—	—	—	—	—	—
全连接	—	128×14×14	128	—	—	—	*Dropout*=0.3

在模型的第 2 条路径中, 具体参数如表 2 所示. 首先对原始信号采用不同步长 (步幅 1、2 和 4) 进行下采样, 生成多个尺度的信号数据, 从而有效降低了数据冗余, 保留了信号中的关键信息, 并使得不同时间尺度上的特征更加明显. 该多尺度下采样策略不仅增强了对故障特征的多分辨率感知, 还通过压缩数据长度提升了后续特征提取的效率. 处理后的下采样信号被输入到 MLP 中, 以捕捉各尺度下的全局特征并进一步整合, 全面表征低频故障的潜在模式, 从而确保了对短期和长期依赖关系的准确建模, 提升了故障诊断的精度. 然而, MLP 部分的输入数据量和网络结构的复杂性可能引发“参数爆炸”问题, 即参数量急剧增加, 导致计算负担加重并可能影响训练稳定性. 正是模型通过多尺度下采样有效压缩了输入数据的维度, 减少了高维数据输入带来的计算复杂度. 同时, MLP 部分采用逐层降维设计, 逐步减少特征维度, 避免了参数量的过度增长. 此外, 模型还引入了 *Dropout* (丢弃率为 0.3) 进行正则化, 有效控制了模型的复杂性, 增强了训练过程的稳定性. 通过这些策略, 模型避免了“参数爆炸”现象, 确保了计算效率和训练的稳定性.

在本研究提出的混合 CNN-MLP 双路径模型中, 采用了端到端的训练策略, 即 CNN 和 MLP 分支在整个训练过程中共享同一损失函数, 进行联合优化. 通过同时训练这两个分支, 模型能够在整体上更有效地提取多尺度时域和时频特征, 从而在小样本条件下提升故障诊断性能.

表 2　MLP 网络参数

主模块名称	子模块	输入维度	输出维度	其他参数
MLP1	FC1	1 024	512	ReLU
	FC2	512	256	ReLU
MLP2	FC1	512	256	ReLU
	FC2	256	128	ReLU
MLP3	FC1	256	128	ReLU
	FC2	128	64	ReLU
FC	—	256+128+64	128	*Dropout*=0.3

3　实验与分析

为了评估本文提出的特征工程处理方法和双路径故障诊断模型的有效性和准确性, 利用凯斯西储大学 (CWRU) 和江南大学 (JNU) 数据集的轴承数据进行实验验证, 此处, 使用 PyCharm 在 Windows 11 系统上使用 PyTorch 2.2.2 框架进行训练和测试. 计算任务在配备 Intel(R) Core (TM) i5-9300H CPU@2.40 GHz、NVIDIA RTX 1660 Ti GPU 和 16 GB RAM 的系统上执行. 本实验中所有模型均使用相同的超参数进行训练, 最大轮次为 20, 批次大小为 32. 使用指数衰减策略调整学习率, 以衰减系数 0.96 每个轮次动态调整学习率. 初始学习率设置为 3E–4, 选择损失函数收敛稳定的模型作为最终的分类模型. 具体的参数如表 3.

表 3　用于模型训练的参数设置

参数类型	参数
轮次	20
批次大小	32
损失函数	交叉熵损失函数
优化器	Adam
学习率	3E–4
学习率衰减系数	0.96
权重衰减	1E–4

3.1　CWRU 数据集

实验采用了 CWRU 提供的滚动轴承故障数据集, 该数据集在基于振动信号的轴承故障诊断研究中被广泛应用. 如图 7 所示, CWRU 试验台由 1.5 kW 电机、扭矩传感器和测功机组成. 实验中电机负载功率分别设置为 0 HP、1 HP、2 HP 和 3 HP (1 HP =735.498 75 W), 对应转速分别为 1 797 r/min、1 772 r/min、1 750 r/min 和 1 730 r/min. 使用的滚动轴承型号为 SKF6205-2RS JEM SKF, 采样频率为 12 kHz. 在轴承内圈、外圈和滚动体上, 分别通过电火花加工引入直径为 0.007 ms、0.014 ms 和 0.021 ms 的单点故障. 外圈故障的位置为 3 点钟、6 点钟和 12 点钟方向, 实验中使用了 6 点钟位置的数据. 从滚动轴承的 9 种故障状态和 1 种正常状态中选择数据, 每种状态包含 233 组样本. 每组样本由 1 024 个采样点组成. 重叠率为 0.5, 实验样本数据集的详细信息在表 4 中.

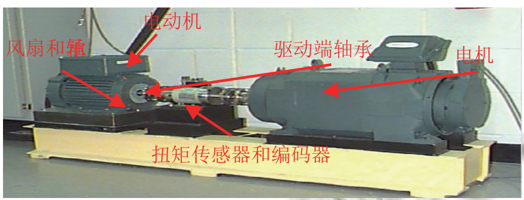


图 7　CWRU 实验装置实物图

表 4　CWRU 实验样本数据集

标签	故障类型	故障直径	数据长度	负载	转速 (r/min)
No (0)	正常	—	1 024	1	1 772
In (1)	轴承内圈	0.007	1 024	1	1 772
Ball (2)	轴承滚子	0.007	1 024	1	1 772
Out (3)	轴承外圈	0.007	1 024	1	1 772
In (4)	轴承内圈	0.014	1 024	1	1 772
Ball (5)	轴承滚子	0.014	1 024	1	1 772
Out (6)	轴承外圈	0.014	1 024	1	1 772
In (7)	轴承内圈	0.021	1 024	1	1 772
Ball (8)	轴承滚子	0.021	1 024	1	1 772
Out (9)	轴承外圈	0.021	1 024	1	1 772

3.1.1　数据集的特征工程处理

对原始数据样本进行 DWT 和 CWT, 可以提取不同尺度的子信号, 并生成它们的时频表示 TFR 图. 图 8 展示了一维详细子信号和近似子信号及其对应的 TFR 图, 图 8 显示出这些子信号在不同频带中的振动特性差异. 通过 TFR 图可以从多个维度清晰呈现各子信号的变化, 能够有效捕捉故障的细微特征. 这进一步证明, 不同尺度的子信号包含了与故障相关的丰富信息.

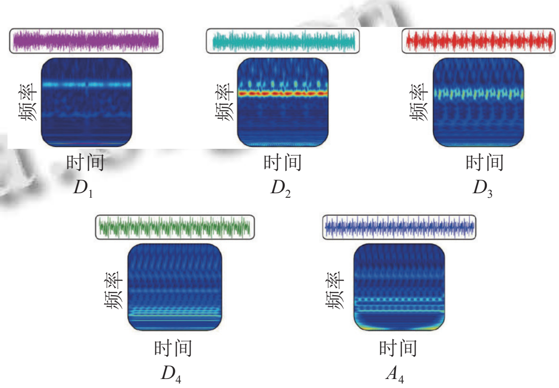


图 8　分量对应的时频图

3.1.2　特征工程性能评估

为了严格评估所提方法在有限样本条件下的特征提取和泛化能力, 实验使用了 CWRU 数据集的子集, 遵循机器学习中常用的小样本训练集设置, 以模拟标记数据稀缺的现实情况. 具体而言, 准备 3 种数据集进行测试, 数据集 A 每种故障类型选取 20 个标记样本用于训练, 80 个样本用于测试; 数据集 B 选取 20 个标记样本用于训练, 100 个样本用于测试, 数据集 C 选取 20 个标记样本用于训练, 160 个样本用于测试. 此调整旨在全面评估方法的性能和鲁棒性, 确保模型不仅在训练中表现优异, 还能有效推广到新数据.

实验中, CWT 用于提供详细的信号 TFR, 捕捉瞬态和振荡特征. 实验将直接对原始信号应用 CWT 的方法, 并与本研究设计的 DWT+CWT 方法进行对比, 通过与普通 CNN 进行结合, 旨在验证 DWT+CWT 特征工程方法在增强特征提取和提高泛化能力方面的有效性.

图 9 和表 5 的结果显示了在有限样本条件下 CWT 方法与提出的 DWT+CWT 方法的对比. 在所有测试数据集上, DWT+CWT 始终实现了更高的准确率. 例如, 在 800 个测试样本中, DWT+CWT 的平均准确率为 99.42%, 而 CWT 为 98.97%. 随着测试集规模的增加,

DWT+CWT 的优势依旧明显, 在 1 000 和 1 600 个测试样本中, DWT+CWT 分别保持了 99.46% 和 99.49% 的准确率, 而 CWT 则分别下降至 98.75% 和 98.43%. 箱型图进一步突显了 DWT+CWT 方法的稳定性. 随着测试集规模增大, CWT 表现出更大的方差和更低的中位数准确率, 反映出其较弱的泛化能力. 相较之下, DWT+CWT 即使在较大的测试集下仍以最小方差提供高精度, 展现了其稳健性和可靠性.

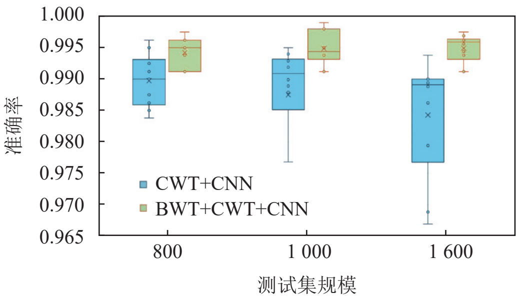


图 9　测试集 10 次实验的准确性分布

表 5　特征工程比较结果

数据集	方法	故障类型	训练集	测试集	轮次	实验次数	最低准确率 (%)	最高准确率 (%)	平均准确率 (%)
A	CWT+CNN	10	200	800	20	10	98.38	99.62	98.97
	DWT+CWT+CNN	10	200	800	20	10	99.12	99.75	**99.42**
B	CWT+CNN	10	200	1 000	20	10	96.47	99.50	98.75
	DWT+CWT+CNN	10	200	1 000	20	10	99.12	99.90	**99.46**
C	CWT+CNN	10	200	1 600	20	10	96.69	99	98.43
	DWT+CWT+CNN	10	200	1 600	20	10	99.12	99.75	**99.49**

3.1.3　不同模型的对比实验分析

在这一部分, 通过对比不同模型的实验结果, 验证了所提出方法的可行性和通用性. 对比的模型包括 CWT+CNN、MCNN+LSTM、CWT-DCNN-LSTM、IFE-CBAM-Inception、MRSVD+CWT+CNN+LSTM 等. 根据表 6 的结果, 所提出的混合双路径模型在仅需 20 个训练样本和 100 个测试样本的极端小样本条件下, 取得了 99.90% 的最高平均准确率, 显著优于依赖大规模训练数据的对比模型. 其他故障诊断模型的平均准确率分别为 98.75%、98.46%、99.00%、99.50% 和 99.83%. 图 10 中的直方图直观展示了各模型的准确率差异, 其中混合双路径模型表现最佳.

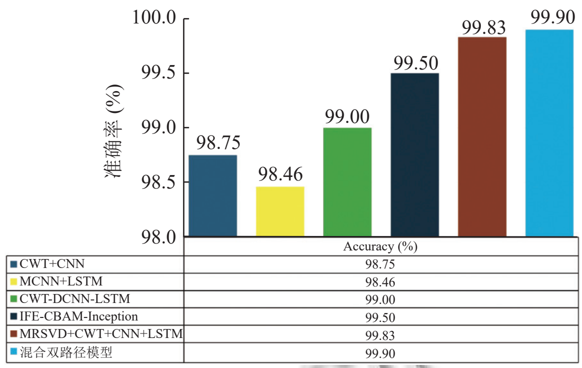


图 10　不同模型准确率直方图

表 6　不同模型的准确率比较

方法	训练	测试	年份	平均准确率 (%)
CWT+CNN	20	100	2018	98.75
MCNN+LSTM[9]	168	72	2021	98.46
CWT-DCNN-LSTM[19]	1 600	480	2023	99.00
IFE-CBAM-Inception[20]	60	20	2023	99.50
MRSVD+CWT+CNN+LSTM[21]	180	60	2024	99.83
混合CNN-MLP双路径模型	20	100	2024	**99.90**

3.1.4　模型消融实验评估

使用数据集 B 进行消融实验, 数据处理方法仍采用 DWT 与 CWT 融合的方式. 为探讨模型中不同模块对最终性能的具体影响, 逐步去除模型中的残差连接层、ECA 注意力机制以及并行的 MLP 分支, 具体实验结果如表 7 所示. 除模块改动外, 其他实验超参数保持不变.

表 7　消融实验结果

模型	残差连接	ECA注意力机制	并行MLP分支	准确率 (%)
1	—	—	—	99.46
2	√	—	—	99.56
3	√	√	—	99.70
4	√	√	√	99.90

模型 1 去除了残差连接、ECA 注意力机制及并行 MLP 分支, 仅保留一条普通 CNN 路径结合 DWT 和 CWT 提取的特征进行分类. 实验结果显示, 该模型达到了 99.46% 的分类准确率.

模型 2 保留了残差连接, 同时去除了 ECA 注意力机制和并行 MLP 分支. 相较于普通 CNN, 准确率稳定提升了 0.10%, 达到 99.56%.

模型 3 在模型 2 的基础上加入了 ECA 注意力机制, 进一步提升了分类能力, 准确率达到 99.70%.

模型 4 为本文完整提出的混合双路径 CNN-MLP

故障诊断模型, 结合了残差连接、ECA 注意力机制及并行的 MLP 分支, 实验中实现了最高的 99.90% 分类准确率.

图 11 展示了不同模型在消融实验中的准确率分布情况. 从图 11 中可以直观地看出, 随着模型组件的逐步加入 (如残差连接、ECA 注意力机制、并行 MLP 分支), 分类准确率逐渐提高, 最终在完整模型中达到最高值. 相比于最基础的普通 CNN 模型, 完整模型的准确率提升显著, 验证了各组件对模型性能的贡献.

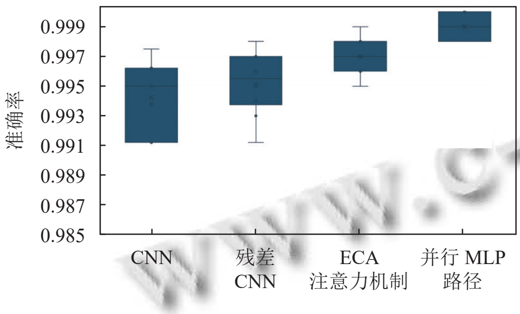


图 11 数据集 B 上的 10 次消融实验结果分布

3.2 JNU 数据集

本节实验使用的数据集来自江南大学的轴承数据集, 具体参数如表 8 所示. 该数据集的采样频率为 50 kHz, 采样时长为 20s. 数据集包含 4 类故障类型, 分别为: 内圈故障、外圈故障、滚动体故障以及正常状态. 每种故障类型均对应 3 种不同的工况, 具体为 3 种转速条件, 分别为 600 r/min、800 r/min 和 1 000 r/min.

3.2.1 混合 CNN-MLP 双路径诊断模型的性能验证

为了保证实验结果的可靠性, 每个数据集进行了 10 次独立实验, 并记录各实验的平均性能指标. 数据集中每种故障类型选取 150 个样本用于训练, 50 个样本用于测试, 50 个样本用于验证. 模型超参数设置和上述一致. 图 12 中的 3 个子图分别展示了在 600 r/min、800 r/min 和 1 000 r/min 工况下的训练和验证准确率曲线、损失曲线以及测试集的混淆矩阵. 在每个工况下, 模型的训练准确率和验证准确率以及损失值都在前 5 轮之后迅速收敛, 并趋于稳定. 此外, 图 12 中的混淆矩阵展示了模型在测试集上的分类表现. 不同工况下, 分类准确率有所差异. 在 600 r/min 时, 模型表现最佳, 准确率达到 98.83%; 而在 800 r/min 和 1 000 r/min 时, 分类准确率略有下降, 但仍维持在较高水平. 这表明不同工况对诊断模型有一定影响, 但整体而言, 模型在各工况下的诊断效果良好. 表 9 实验结果进一步验证了所提方法在不同工况下对轴承故障诊断的有效性和鲁棒性.

表 8 JNU 数据集故障类型

标签	故障类型	数据长度	转速 (r/min)
In (0)	内圈故障	1 024	600
Normal (1)	正常状态	1 024	600
Out (2)	轴承外圈	1 024	600
Ball (3)	轴承滚子	1 024	600
In (0)	内圈故障	1 024	800
Normal (1)	正常状态	1 024	800
Out (2)	轴承外圈	1 024	800
Ball (3)	轴承滚子	1 024	800
In (0)	内圈故障	1 024	1 000
Normal (1)	正常状态	1 024	1 000
Out (2)	轴承外圈	1 024	1 000
Ball (3)	轴承滚子	1 024	1 000

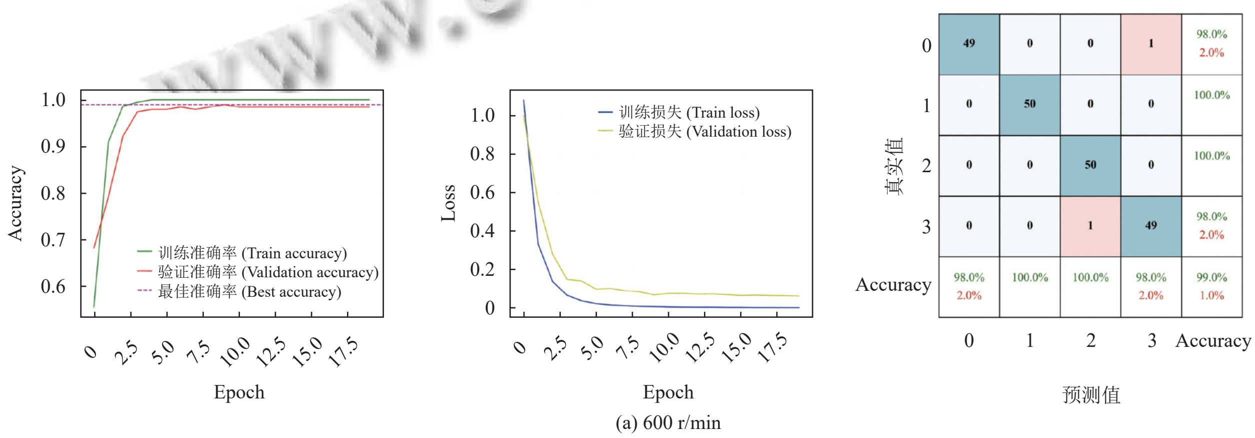
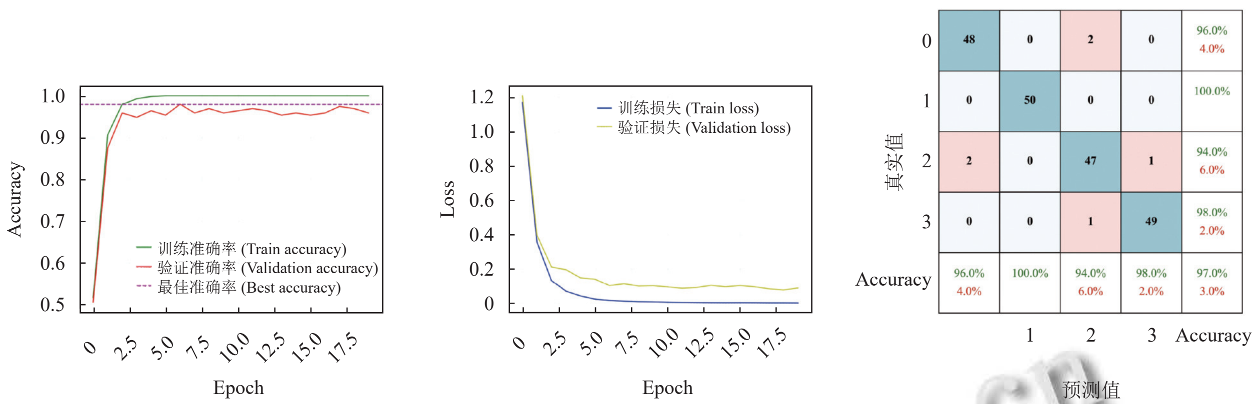
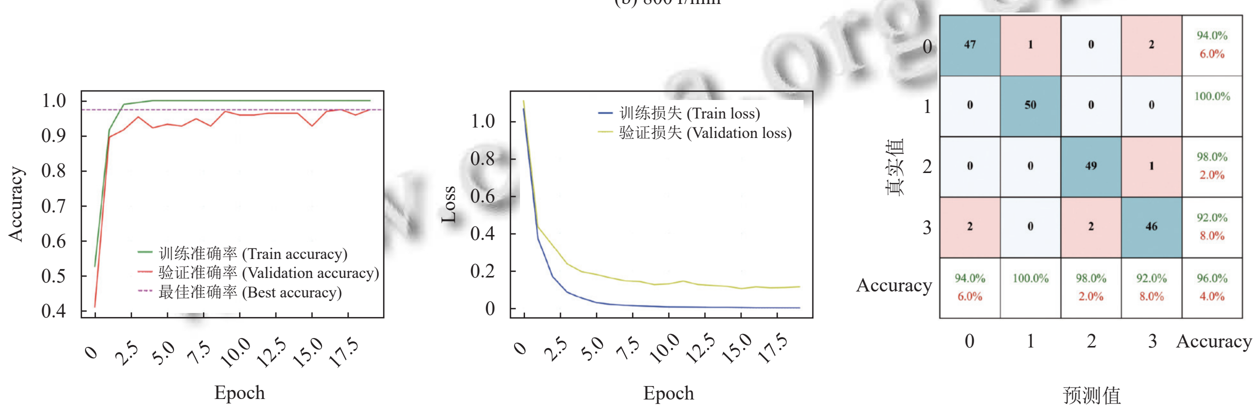


图 12 3 种工况训练验证准确率和测试混淆矩阵可视化



(b) 800 r/min



(c) 1 000 r/min

图 12 3 种工况训练验证准确率和测试混淆矩阵可视化 (续)

表 9 双路模型不同工况下的准确率

工况 (r/min)	实验次数	最低准确率 (%)	最高准确率 (%)	平均准确率 (%)
600	10	97.40	99.00	98.38
800	10	96.35	97.92	97.20
1 000	10	95.31	96.88	96.03

4 结论

(1) 本研究提出了一种混合 CNN-MLP 双路径故障诊断方法. 通过特征工程结合 DWT 和 CWT, 将振动信号转换为时频特征图并进行通道堆叠. 得到多尺度时频特征, 同时对原始信号下采样得到多尺度时域特征.

(2) 随后利用结合 ECA 注意力机制的残差 CNN 路径对这些时频特征进行处理同时, 将下采样后的多尺度时域特征输入 MLP 路径进行分析. 通过混合双路径模型, 时频特征与时域特征得以有效融合, 充分结合了特征工程与深度学习技术.

(3) 在 CWRU 数据集的小样本数据集上, 混合双路模型达到了 99.90% 的准确率, 在江南大学数据集上, 达到了 98.38% 的准确率. 所提出的模型均取得了卓越的准确率, 表明该方法能够显著提升故障诊断的准确性和鲁棒性.

参考文献

1 王婧, 许志伟, 刘文静, 等. 滚动轴承健康智能监测和故障诊断机制研究综述. 计算机科学与探索, 2024, 18(4): 878–898.

2 李政, 张炜, 明安波, 等. 基于 IEWT 和 MCKD 的滚动轴承故障诊断方法. 机械工程学报, 2019, 55(23): 136–146.

3 Zhang PB, Chen RX, Xu XY, *et al*. Recent progress and prospective evaluation of fault diagnosis strategies for electrified drive powertrains: A comprehensive review. Measurement, 2023, 222: 113711. [doi: 10.1016/j.measurement.2023.113711]

4 陈是扦, 彭志科, 周鹏. 信号分解及其在机械故障诊断中的应用研究综述. 机械工程学报, 2020, 56(17): 91–107.

5 孟事业, 罗倩. 基于 CEEMDAN 优化的轴承故障变分推断诊断算法. 电子测量技术, 2023, 46(22): 94–101.

6 彭成, 王松松, 贺婧, 等. 基于离散小波变换和随机森林的轴承故障诊断研究. 计算机应用研究, 2021, 38(1): 101–105.

7 王永鼎, 金子琦. 基于融合 CNN 与 PSO-SVM 的滚动轴承故障诊断. 机械强度, 2021, 43(4): 793–797.
8 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]
9 Chen XH, Zhang BK, Gao D. Bearing fault diagnosis base on multi-scale CNN and LSTM model. Journal of Intelligent Manufacturing, 2021, 32(4): 971–987. [doi: 10.1007/s10845-020-01600-2]
10 Jiao JY, Zhao M, Lin J, *et al*. A comprehensive review on convolutional neural network in machine fault diagnosis. Neurocomputing, 2020, 417: 36–63. [doi: 10.1016/j.neucom.2020.07.088]
11 Li J, Liu YB, Li QJ. Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method. Measurement, 2022, 189: 110500. [doi: 10.1016/j.measurement.2021.110500]
12 董绍江, 裴雪武, 吴文亮, 等. 基于多层降噪技术及改进卷积神经网络的滚动轴承故障诊断方法. 机械工程学报, 2021, 57(1): 148–156.
13 董绍江, 朱朋, 朱孙科, 等. 基于仿真数据驱动和领域自适应的滚动轴承故障诊断方法. 中国机械工程, 2023, 34(6): 694–702.
14 司伟伟, 岑健, 伍银波, 等. 小样本轴承故障诊断研究综述. 计算机工程与应用, 2023, 59(6): 45–56.
15 Sinitsin V, Ibryaeva O, Sakovskaya V, *et al*. Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. Mechanical Systems and Signal Processing, 2022, 180: 109454. [doi: 10.1016/j.ymssp.2022.109454]
16 陈鹏. 基于振动信号的滚动轴承故障诊断方法综述. 轴承, 2022(6): 1–6.
17 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444.
18 Jalayer M, Orsenigo C, Vercellis C. Fault detection and diagnosis for rotating machinery: A model based on convolutional LSTM, fast Fourier and continuous wavelet transforms. Computers in Industry, 2021, 125: 103378. [doi: 10.1016/j.compind.2020.103378]
19 Wang Y, Zhu CF, Wang QR, *et al*. Research on fault detection of rolling bearing based on CWT-DCCNN-LSTM. Engineering Letters, 2023, 31(3): 987–1000.
20 Xu SJ, Yuan R, Lv Y, *et al*. A novel fault diagnosis approach of rolling bearing using intrinsic feature extraction and CBAM-enhanced InceptionNet. Measurement Science and Technology, 2023, 34(10): 105111. [doi: 10.1088/1361-6501/ace19c]
21 Guo Y, Zhou J, Dong ZB, *et al*. Research on bearing fault diagnosis based on novel MRSVD-CWT and improved CNN-LSTM. Measurement Science and Technology, 2024, 35(9): 095003. [doi: 10.1088/1361-6501/ad4fb3]

(校对责编: 王欣欣)