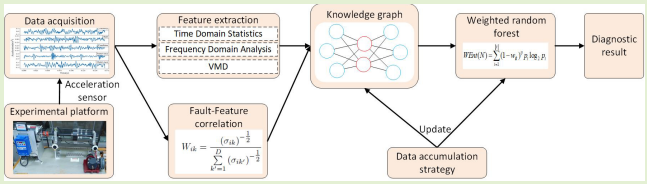
基于知识图谱和数据积累策略的滚动轴承故障诊断

肖向渠、李超顺、 IEEE 会员、黄杰、于田

摘要：滚动轴承的故障诊断对于保证旋转机械的安全稳定运行和维护具有重要作用。传统的轴承故障诊断方法没有考虑故障与特征之间的相关性，也没有充分利用不断增加的监测数据。由此，提出了一种基于知识图谱（KG）和数据积累策略的轴承故障诊断框架。首先，实体



KG 是根据从振动传感器收集的轴承振动数据的时域、频域和时频域提取的多个特征来定义的。然后，设计并计算作为边的特征 - 故障相关性，与实体一起建立 KG 框架。此外，提出了加权随机森林算法作为知识图谱的推理算法，充分利用特征与故障的相关性来提高轴承故障分类的准确性。最后，设计了数据积累策略来不断增加 KG 的训练数据集的大小。在此过程中更新相关参数，使推理算法产生的结果更加准确。通过与相同情况下的多个模型的比较，证明了该方法的优点。测试结果表明，该方法具有良好的前景，对不同工况具有良好的预测精度和鲁棒性。

IndexTerms— 数据积累策略、故障诊断、特征故障关联、知识图谱（KG）、加权 dom 森林（WRF）。

ran

一、简介

ROTATING machinery plays an important role in indus-试验系统。一旦发生故障，往往会造成较大的经济损失，严重时甚至造成重大事故 [1]。滚动轴承是旋转机械的重要零件，也是比较脆弱的零件。据统计，轴承故障约占旋转机械的 30%[2]。因此，开展滚动轴承故障诊断研究，实现故障类别快速准确分类，保障旋转机械稳定高效运行具有重要意义。

在工业发展初期，轴承故障检测主要依靠技术人员的经验。随着计算机技术的出现和发展，频谱分析已从理论变为现实。轴承特性和快速傅里叶变换技术可用于确定轴承故障。随后的共振解调技术也被广泛应用于轴承故障诊断领域。随着计算机技术的快速发展，机器学习和深度学习逐渐成为轴承故障诊断的主流技术。

基于机器学习的滚动轴承故障诊断过程通常可分为两个步骤。首先利用时域统计分析等变换域分析方法提取原始信号的典型特征，然后设计分类器对故障类型进行分类和诊断。其中，贝叶斯准则、线性判别函数和非线性判别函数常用于设计模式分类器。常见的模型包括支持向量机（SVM）[3],[4]，[5]，K 最近邻（KNN）[6]，[7]，朴素贝叶斯（NB）[8]，和人工神经网络（ANN）[9], [10]。这类方法通常实现起来比较简单，但面对复杂的故障情况可能会表现出不足，并具有一定的局限性。深度学习技术

手稿于 2022 年 7 月 29 日收到；2022 年 8 月 14 日接受。出版日期 2022 年 8 月 31 日；当前版本的日期为 2022 年 9 月 30 日。这项工作部分得到了国家自然科学基金委 51879111 的资助，部分得到了武汉科技局应用基础前沿项目的资助 2018010401011269，以及湖北省自然科学基金委 2019CFA068 的部分资助。协调本文评审并批准发表的副主编是王栋博士。（通讯作者：李超顺）

作者单位：华中科技大学土木与水利工程学院，武汉 430074（e-mail：d201981042@hust.edu.cn；csli@hust.edu.cn；hjie@hust.edu.cn；d201880943@hust.edu.cn）。数字对象标识符 10.1109/JSEN.2022.3201839

近年来发展起来的方法展现了将传统两步任务统一到整个流程中的强大能力。深度学习 [11], [12], [13] 可以通过多层网络结构在输入和输出之间建立复杂的非线性关系。通过复杂的非线性变换，可以从训练数据中自动学习底层特征，并通过多层网络逐渐形成。抽象的高层表示克服了人工选择特征的缺点，自适应地获得最优特征组合，实现端到端学习，完成轴承的故障诊断。这些基于深度学习的方法在轴承故障诊断方面取得了良好的效果。然而，他们很少考虑如何处理不断增加的故障轴承监测数据。监测数据的不断积累和使用将丰富故障知识，从而有效提高诊断方法的准确性。因此，考虑到实际工业应用中数据会不断增加且故障知识不是静态的，研究用于轴承故障诊断的知识图谱（ KG ）是很有吸引力的。

KG 是 Google 在 2012 年提出的语义网络，它以网络 [14] 的形式存储知识。 KG 以实体和关系的形式表达事实。实体是图中的节点，关系是连接图中节点的边 [15]。KG 可以有效解决自动问答、个性化推荐、智能信息检索等问题 [16], [17],，广泛应用于反金融欺诈和诊疗决策 [18]。根据知识库的覆盖范围，可分为通用知识库和行业知识库。一般的知识图谱是面向多个领域甚至整个领域的。重点是知识的广度，所涉及的实体和关系的数量通常是巨大的。具有代表性的通用知识图谱有 WordNet [19],DBpedia [20], wikidata [21], 等。行业知识图谱也称为领域知识图谱。本文对轴承故障诊断领域 KG 进行研究。此类知识图谱利用特定领域的行业数据构建，强调知识的专业性和可靠性，为行业人员提供相对准确、详细的信息，例如金融知识图谱 [22] 和医疗知识图谱 [23]。然而，在滚动轴承故障诊断领域，目前还没有比较有代表性的研究。

目前，专业知识图谱在轴承故障诊断领域尚属空白，相关方法和技术还处于起步阶段。参考其他领域的知识图谱系统，得到相应的关键步骤：提取振动信号特征、量化特征与故障的相关性、基于构建的知识图谱推理故障类别、随着数据数量的增加进行知识积累。振动信号特征提取可以获得凸显不同故障类别差异的信息，弱化或去除与故障诊断无关的信息。振动信号特征提取方法大致可分为时域分析、频域分析和时频分析。将会用到这三个方法提取振动信号的特征弥补了单一方法的缺点，从而提高了故障分类的准确性。此外，利用软聚类中的特征权重系数来量化特征与故障之间的相关性，提出改进的随机森林算法进行故障类别推断，利用特征与故障之间的相关性来最大化故障分类精度，并提出数据积累策略。

本研究的主要科学贡献可列举如下。

1 ）定义特征权重系数，可以量化故障与特征之间的相关性，表明各个特征在不同故障类别中的重要性，有效降低无效特征对故障诊断准确性的影响。

2 ）创新性地提出了一种基于知识图谱的轴承故障诊断模型，并提出改进的随机森林算法作为知识图谱的推理算法来实现故障诊断。该算法根据特征重要性对分裂节点进行划分，利用特征与故障之间的相关性来有效提高分类精度。 3 ）提出了 KG 数据的积累策略，可以实现数据的有效积累和模型参数的更新，最大限度地利用数据，实现故障分类精度的提高。

本文的其余部分组织如下。第二节介绍随机森林算法和故障 KG 的结构。第三节详细介绍了 KG 的构建方法、随机森林改进策略以及 KG 更新和数据积累策略。实验过程和结果在第四节中介绍。最后，第五节给出了结论。

ii。相关作品

A. 知识图谱

知识图谱是一个结构化的语义知识库，由一系列节点、边和属性组成，并使用三元组来描述数据。三元组的形式一般为 “ 实体 - 关系 - 实体 ”。节点是实体（或概念）在物理世界中的表现。实体是实际的事物，而概念通常是事物的集合。边指的是节点之间的关系，关系就是实体之间的各种联系。

KG 的构建和应用有一系列的流程和操作。以下部分主要关注实体获取、关系获取、知识推理和累积更新。

1) *Entity Acquisition:* 对于轴承领域的 KG 振动，实体主要包括振动信号的特征和轴承的典型状态。通过时域特征函数、频域特征函数和变分模态分解 (VMD)[24] 可以得到振动信号的特征。

2) *Relationship Acquisition:* 对于实体之间的关系，需要用具体的数据来描述

在轴承故障 KG 中，特征与故障之间的关系往往非常重要，这可以通过统计方法或其他方法获得。

3） *Knowledge Reasoning:* 知识推理是 KG 的重要组成部分，其他未知的事物可以从已知的知识推理中推导出来。知识推理方法可以分为两类：基于逻辑的推理 [25], [26] 和基于图的推理 [27]。在实现上，推理可以通过逻辑规则和逻辑符号来实现，也可以通过机器学习算法和神经网络模型来实现。

4） *Accumulate and Update:* 知识的积累和更新是 KG 的优势之一。随着时间的推移，信息量和知识量会不断增加，因此知识图谱的内容也必须与时俱进。不断增加的知识和数据将使 KG 的功能更加强大。

B. 随机森林

它是由多个决策树模型 [28] 组成的机器学习方法。利用随机化的方法选择样本和特征，生成多棵决策树，然后对所有决策树的结果进行汇总，得到最终的分类结果。

传统的随机森林分类算法主要有四个基本内容。

1）从 N 个 *Randomly Select Samples:* 原始训练样本集中进行放回采样，得到 N 个样本，形成新的训练集。2） *Randomly Select Features:* 在构建决策树的过程中，当节点分裂时，算法需要从 k 特征中无替换地提取 k_0 特征。 3) *Build a Decision Tree:* 根据 k_0 特征的信息增益、增益比或基尼指数选择当前节点的最优分裂特征属性，重复该过程直至构建完成。 4) *Random Forest Voting Mechanism:* 当使用随机森林作为分类模型时，选择决策树中出现次数最多的结果作为最终的分类结果。

随机森林分类算法具有不易产生过拟合、特征要求低、调整参数少等优点，在机器学习领域具有广泛的应用前景。在本文中，我们将尝试改进随机森林中决策树节点分裂的最具特征选择方法，使其能够充分利用不同类别下每个特征的重要性来改善最终的分类结果。

三. 提议的方法

本节将介绍轴承故障知识图谱（ BFKG ）的相关内容。针对不同故障各特征重要性不同的问题，提出故障 - 特征相关性计算方法，建立

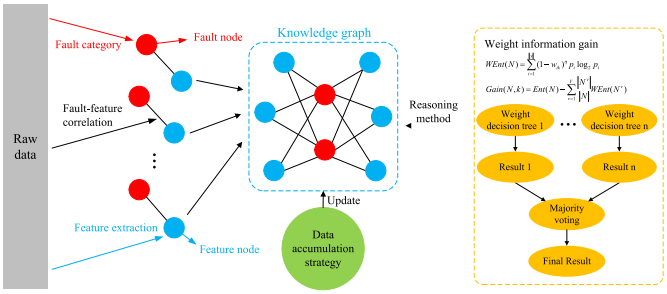


图 1. 用所提出的方法进行机械故障诊断的通用流程。

表 1 时域特征参数

Time-domian feature parameters

$$p_1 = \frac{\sum_{n=1}^{N} x(n)}{N} \qquad p_7 = \frac{\sum_{n=1}^{N} (x(n)-p_1)^4}{(N-1)p_2^4}$$

$$p_2 = \sqrt{\frac{\sum_{n=1}^{N} (x(n)-p_1)^2}{N-1}} \qquad p_8 = \frac{p_5}{p_4}$$

$$p_3 = \left(\frac{\sum_{n=1}^{N} \sqrt{|x(n)|}}{N}\right)^2 \qquad p_9 = \frac{p_5}{p_3}$$

$$p_4 = \sqrt{\frac{\sum_{n=1}^{N} (x(n))^2}{N}} \qquad p_{10} = \frac{p_4}{\frac{1}{N}\sum_{n=1}^{N} |x(n)|}$$

$$p_5 = \max|x(n)| \qquad p_{11} = \frac{p_5}{\frac{1}{N}\sum_{n=1}^{N} |x(n)|}$$

$$p_6 = \frac{\sum_{n=1}^{N} (x(n)-p_1)^3}{(N-1)p_2^3}$$

Where $x(n)$ is a signal series for n = 1,2,...,N, N is the number of data points.

KG 节点之间的 “ 边 ”。同时提出一种改进的随机森林算法，旨在在决策树节点分裂时利用故障特征相关性辅助选择最优特征。最后提出了数据积累策略，每次测试后选择可靠性高的结果添加到下一个训练集中，不断增加训练集数据以提高轴承故障分类的准确性。整体流程如图 1 所示。

A. 特征提取

使用长度为 2400 的滑动窗口将测量信号分成更小的信号段。由于采样频率高，2400 的长度已经包含了轴承所必需的健康信息。信号段的特征提取对于故障诊断非常重要。本文从时域、频域、时频域三个方面提取信号特征，以获得更完整的信号特征。

信号的多重特征可以有效提高轴承故障诊断的准确性。本文引用 [29], [30] 来提取相关特征。时域特征包括信号均值、峰度等 11 个指标，如表 1 所示。其中一些特征用于反映信号在时域中的幅度和能量。

表 2 频域特征参数

Frequency-domian feature parameters

$p_{12}=\frac{\sum_{l=1}^{L}s(l)}{L}$	$p_{18}=\sqrt{\frac{\sum_{l=1}^{L}(f_l-p_{16})^2 s(l)}{L}}$
$p_{13}=\frac{\sum_{l=1}^{L}(s(l)-p_{12})^2}{L-1}$	$p_{19}=\sqrt{\frac{\sum_{l=1}^{L}f_l^4 s(l)}{\sum_{l=1}^{L}f_l^2 s(l)}}$
$p_{14}=\frac{\sum_{l=1}^{L}(s(l)-p_{12})^3}{L(\sqrt{p_{13}})^3}$	$p_{20}=\frac{\sum_{l=1}^{L}f_l^2 s(l)}{\sqrt{\sum_{l=1}^{L}s(l)\sum_{l=1}^{L}f_l^4 s(l)}}$
$p_{15}=\frac{\sum_{l=1}^{L}(s(l)-p_{12})^4}{Lp_{13}^2}$	$p_{21}=\frac{p_{18}}{p_{16}}$
$p_{16}=\frac{\sum_{l=1}^{L}f_l s(l)}{\sum_{l=1}^{L}s(l)}$	$p_{22}=\frac{\sum_{l=1}^{L}(f_l-p_{16})^3 s(l)}{Lp_{18}^3}$
$p_{17}=\sqrt{\frac{\sum_{l=1}^{L}f_l^2 s(l)}{\sum_{l=1}^{L}s(l)}}$	$p_{23}=\frac{\sum_{l=1}^{L}(f_l-p_{16})^4 s(l)}{Lp_{18}^4}$

Where $s(l)$ is a spectrum for l = 1,2,…,L, L is the number of spectrum lines; f_l is the frequency value of the l-th spectrum line.

该特征的另一部分反映了信号在时间序列中的分布。频域特征主要是信号经过傅里叶变换后在频域提取的一系列统计指标。本文共提取了 12 个频域特征，如表 2 所示。其中一部分特征用于描述信号频谱的收敛情况，另一部分特征则反映了信号频率分量中主要频率的分布情况。

本文涉及的信号时频域特性主要包括能量特性和奇异值特性。本文利用 VMD 方法对信号进行变换，得到一系列模态，进而得到这些模态的能量特征和奇异值特征。对原始信号进行 VMD 分解，得到一系列模态 $C=\{C_1, C_2, 。。。, C_M\}$，其中 M 是分解后的模式数 —— 在本文中，M 设置为 4。E_1， E_2，。。。，E_m 为各模式对应的能量值。具体计算公式如下：

$$E_m=\int_{\infty}^{+\infty}|C_m|^2 dt. \tag{1}$$

信号的分解模式形成矩阵 $A=[C_1, C_2, 。。。, C_M]^T$。矩阵的奇异值分解可以提取矩阵的奇异值特征。

B. 故障特征相关性

对于不同的轴承故障（即不同的工作条件），相同的特征通常具有不同的重要性。此外，从振动信号中提取的特征可能存在冗余或噪声。直接使用这些特征可能会降低分类的准确性。同时，它不能反映每个特征对不同故障类型的贡献。本文介绍了一种计算方法给出了故障与特征之间的相关性。具体方程如下：

$$w_{ik}=\frac{(\sigma_{ik})^{-\frac{1}{2}}}{\sum_{k'=1}^{D}(\sigma_{ik'})^{-\frac{1}{2}}};\quad i=1,2,\ldots,C;\ k=1,2,\ldots,D \tag{2}$$

$$\sigma_{ik}=\sum_{j=1}^{N}u_{ij}(x_{jk}-v_{ik})^2 \tag{3}$$

$$v_{ik}=\frac{\sum_{j=1}^{N}u_{ij}x_{jk}}{\sum_{j=1}^{N}u_{ij}}. \tag{4}$$

式 (2) 中，w_{ik} 表示第 k 个特征与第 i 个故障之间的相关性，D 为特征类别总数，C 为故障类别个数，σ_{ik} 表示属于该特征的所有信号特征距第 k 维数据的距离之和。i 类别到对应的数据中心，(3) 为 σ_{ik} 的计算方程。(3) 中，u 为隶属度矩阵，其中 u_{ij} 表示第 j 数据属于第 i 种故障的概率，x_{jk} 表示第 j 样本的第 k 特征维度数据，v_{ik} 表示第 i 断层数据的第 k 维数据中心，式（4）为 v_{ik} 的计算方程。(4) 中的符号前面已介绍过。

从 (2) 可以看出，对于属于第 i 种断层的数据的第 k 维特征，越靠近该维的数据中心，w_{ik} 越大，表明第 i 种断层下的数据的第 k 维特征更加集中；则该特征与故障类型的相关性越大。同时可以发现，所有特征与同一故障的相关性之和为 1。利用该方法，可以得到每个特征与每种故障类型的相关性，为知识图谱 “ 边 ” 的建立提供了理论依据。

C. 加权随机森林算法

对于轴承故障 KG 来说，知识推理过程非常重要。知识图谱建立后，需要通过推理过程来实现轴承故障的分类。采用随机森林算法作为 KG 的推理方法。信息熵是衡量样本集纯度的常用指标，通常用于决策树划分节点时进行最优特征选择。传统的信息熵没有利用各个特征与故障的相关性，即没有利用知识图谱中的 “ 边缘 ” 信息，因此传统的随机森林不适合在这里用作故障类型分类的推理算法。本文针对这一点提出了加权随机森林（WRF）算法。对传统的信息熵方程进行调整，旨在利用故障特征相关性辅助决策树节点分裂。公式由下式给出

$$W\mathrm{Ent}(N)=\sum_{i=1}^{|y|}(1-w_{ik})^{\eta}p_i\log_2 p_i \tag{5}$$

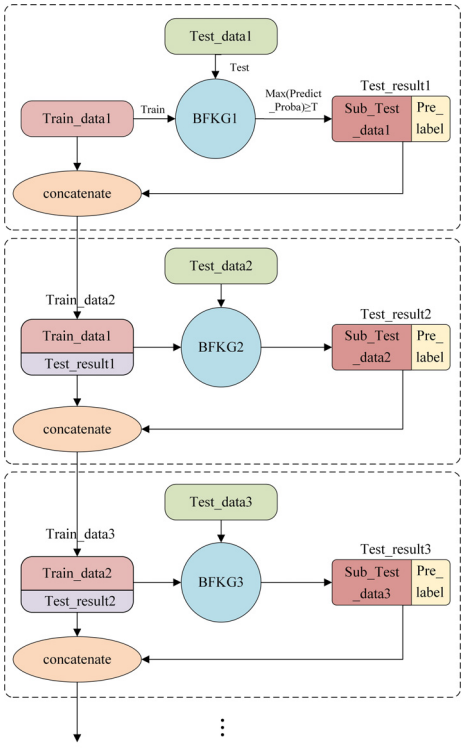


图 2. 数据积累流程图。 BFKG 就是本文提出的诊断方法。如果 Test_data 中数据的最大预测概率大于给定的阈值 T，则将数据记录在 Sub_Test 数据中，Pre_label 是数据的预测标签，Sub_Test 数据和 Pre_label 形成 Test_result。前一个模型的 Train_data 和 Test_result 连接起来形成下一个模型的 Train_data。

其中 w_{ik} 表示第 k 个特征与第 i 个故障之间的相关性，$|y|$ 表示当前分支类别的数量，p_i 表示第 i 个数据在集合 N 中的比例，η 是调整参数 —— 在本文中，η 设置为 1/3。

WRF 采用信息增益比分割准则，其中利用特征 k 划分样本 N 得到的信息增益比方程为

$$\mathrm{Gain}(N,k)=\mathrm{Ent}(N)-\sum_{v=1}^{V}\frac{|N^v|}{|N|}W\mathrm{Ent}(N^v) \tag{6}$$

其中 Ent(·) 是传统的信息熵方程；V 是分裂产生的分支数，本文中 V 设置为 2；$|N^v|$ 是第 v 个分支上的样本数。

D. KG 数据积累策略

对于数据驱动模型来说，数据规模非常重要。用于训练的数据集越大，测试结果的效果越好。因此，增加数据规模、提高数据利用程度是有意义的。

本文提出了一种针对 KG 设计的数据积累策略，如图 2 所示。该策略在每次数据测试后选择合适的测试数据和相应的预测结果作为下一个模型训练的新数据。该策略的主要工作是引入一个阈值 T。这个阈值的作用是筛选合适的数据测试结果，即选择结果中预测概率最大不小于 T 的数据，对应的标签值就是模型预测的标签值。

这个阈值通常与模型在数据集上的性能有关。当模型对数据集有很好的分类效果时，阈值 T 往往会比较小，以便可以筛选出更多的测试结果用于下一步的训练。当模型对数据集的分类效果较差时，阈值往往需要较大，以便过滤后的结果中存在尽可能少的错误样本。

具体实施步骤如下。*Step 1:* 将原始数据集分成多个数据集，其中一个作为训练集，其余作为测试集。*Step 2:* 根据训练样本数据确定隶属度矩阵 u 。当第 j 样本数据属于第 i 类故障时，$u_{ij}=1$；否则，$u_{ij}=0$。根据式 (4) 计算各类数据的样本中心矩阵 v。根据（2）和（3）计算特征与故障之间的相关矩阵 W。*Step 3:* 训练 BFKG 模型，得到测试集样本的预测标签和概率矩阵。*Step 4:* 取测试样本进行分析。如果该测试样本的概率矩阵最大值不小于 T，则将此测试样本与预测的标签组合添加到下一个训练样本中，并将概率矩阵添加到先前的隶属度上，作为矩阵中新样本的隶属度矩阵，为下一次训练做好准备。

四. 实验结果与比较

将使用三个轴承数据集，其中两个由机械故障预防技术协会和凯斯西储大学 (CWRU) 提供。另一种是由我们实验室的机械故障综合模拟实验系统采集的，包括轴承的各种状态。在这些数据集上设计了案例。案例一是一个常规实验，将 BFKG 与其他方法的分类准确率进行比较，以验证 BFKG 在诊断方面的优势。在案例二中，用原始训练集训练模型后，每个测试集中值得保留的数据将添加到训练集中以供下一次模型训练。通过与训练集中不添加数据的情况进行对比，验证了数据增长模式下数据积累策略的有效性。通过与其他使用数据积累策略的分类方法进行比较，验证了 KG 中使用数据积累的优势。在案例 III 中，通过在 CWRU 数据集上构建不平衡数据集进行实验，验证了所提方法的强大鲁棒性。

A. 数据集描述

1） MFPT 数据集（MT0 和 MT1）：MFPT 数据集由机械故障预防技术协会 [31] 提供。该数据集包含轴承测试的数据

表 III CWRU 数据集的总共 28 个工作条件

Health condition	0W	735.5W	1471W	2206.5W
Normal	√	√	√	√
OR 0.1778mm	√	√	√	√
OR 0.5334mm	√	√	√	√
IR 0.1778mm	√	√	√	√
IR 0.5334mm	√	√	√	√
B 0.1778mm	√	√	√	√
B 0.5334mm	√	√	√	√

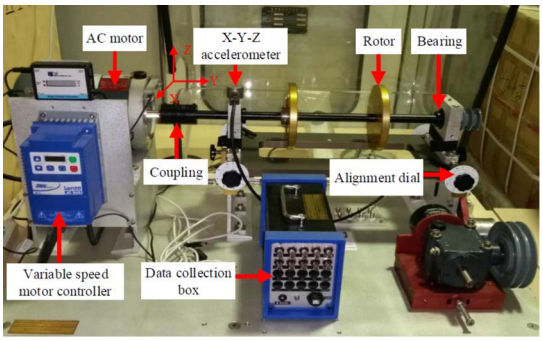


图 3 实验室轴承实验平台。

钻机和三个真实世界的故障数据。轴承试验台的数据包括 3 种基线数据集、 10 种外圈故障数据集和 7 种内圈故障数据集。选取 15 个状态的信号数据进行实验。具体组成部分包括基线条件数据；七种不同载荷条件下的外圈失效数据，载荷分别为 11.34 、 22.68 、 45.36 、 68.04 、 90.72 、 113.40 和 136.08 kg ；以及七种不同负载条件下的内圈故障数据，负载分别为 0 、 22.68 、 45.36 、 68.04 、 90.72 、 113.40 和 136.08 kg。轴承振动信号的采样频率为 48828 Hz，采样时间为 3 s。每个州的样本数为 50，数据集中的样本总数为 750。

2） CWRU 数据集（ CU0 和 CU1 ）： CWRU 电气工程实验室提供了轴承振动数据集，广泛应用于轴承故障诊断领域 [32]。测试台由 2 马力电机、扭矩传感器 / 编码器和测功机组成。该数据集包含四种轴承状态，包括正常（ N ）、外圈故障（ OR ）、内圈故障（ IR ）和滚珠故障（ B ）。数据集包含四种不同的电机负载条件：0 、 735.5 、 1471 和 2206.5 W。选择 28 种状态的信号数据进行实验。具体组件包括四种不同载荷下的正常状态数据：八种不同载荷和故障直径下的外环故障、内环故障和滚动体故障。具体信息见表三。每个状态的样本数为 50，数据集中的样本总数为 1400。信号采样频率为 12 kHz，采样时间为 10 s。

3） MFSDatasets （ MS0 和 MS1 ）：此外，更多的轴承数据是由我们自己的实验室产生的。图 3 显示了我们实验室的实验装置。该试验台为机械故障综合模拟实验系统（ MFS ），

(a) (b) (c) (d)

图 4. 具有不同故障的轴承。 (a) 外圈故障。 (b) 内圈故障。 (c) 球故障。 (d) 组合故障。

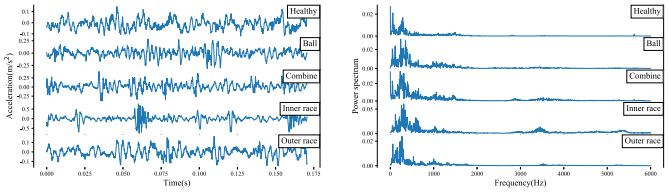


图 5。 MFS 数据集中样本的时域和频域输出。

表四 案例一数据说明

Dataset	Data description	Train	Test
MT0	15 classes of vibration data consists of baseline, Inner Race Fault and Out Race Fault with different loads.	15×40	15×10
CU0	28 classes of fans end vibration data. They consist of N, IRF, ORF, and BF with loads ranging from 0 to 3 loads and fault diameter are 0.1778 and 0.5334 mm.	28×40	28×10
MS0	15 classes of vibration data. They consist of N, IRF, ORF, BF, and CF, the speed ranging from 10rev/s to 30rev/s.	15×48	15×12

美国 Spectra Quest 公司生产的模拟旋转机械轴承故障的实验装置。该系统由机械故障综合仿真实验平台、实验平台套件、传感器、数据采集系统、 PC 终端、数据采集分析软件组成。本实验的数据集包括正常数据（ N ）、外圈故障数据（ OR ）、内圈故障数据（ IR ）、滚珠故障数据（ B ）和组合故障数据（ C ）。每种故障情况的详细描述如图 4 所示。实验装置的轴转速为 10 、 20 和 30 转 / 秒。信号采样帧为 12 kHz，采样时间为 12 s。选取 15 个状态的信号数据进行实验。具体成分包括外环故障、内环故障、滚动体故障、组合故障以及三种不同转速下的正常故障。具体信息如表四所示。每个状态的样本数量为 60，数据集中的样本总数为 900。图 5 给出了部分健康状态样本的时频图。

根据不同的实验对数据集进行不同的处理。在案例一中，每个数据集被简单地分为训练集和测试集。训练集数据占原始数据的 80%，其余 20% 的数据集为测试集；将三个原始数据集处理后得到的新数据集命名为 MT0 、 CU0 和 MS0，如表 5 所示。在案例 II 中，每个数据集均分为 5 个

表五 案例二数据说明

Dataset	Data description	Train	Test
MT1	15 classes of vibration data consists of baseline, Inner Race Fault and Out Race Fault with different loads.	15×10	15×10×4
CU1	28 classes of fans end vibration data. They consist of N, IRF, ORF and BF with loads ranging from 0 to 3 loads and fault diameter are 0.1778 and 0.5334 mm.	28×10	28×10×4
MS1	15 classes of vibration data. They consist of N, IRF, ORF, BF, and CF, the speed ranging from 10rev/s to 30rev/s.	15×12	15×12×4

表 VI MFS 数据集总共 15 个工作条件

Health condition	10rev/s	20rev/s	30rev/s
N	√	√	√
OR	√	√	√
IR	√	√	√
B	√	√	√
C	√	√	√

部分。其中一个数据集作为初始训练集，其余四个数据集作为测试集，测试数据积累策略的效果。共有三个原始数据集。处理后得到的新数据集分别命名为 MT1 、 CU1 、 MS1，如表六所示。

B. 案例一：BFKG 分类效果验证

在案例一中，BFKG 与其他分类模型进行了比较，包括 RF 、 SVM 、 ANN 、一维卷积神经网络（ 1-D CNN ）、CNN-PCA-FCM[33]，改进的随机森林 [34]，深度残差收缩网络 [35]，图卷积网络（ GCN ） [36]，和其他 WRF [37]。比较的模型包括传统的机器学习算法和神经网络模型。一维 CNN 和 CNN-PCA-FCM 的输入是长度为 2400 的信号段，因为它们可以自动提取深层特征。验证了 BFKG 在故障分类方面较其他分类模型具有一定的优势。这些模型在数据集 CU0 、 MT0 和 MS0 上进行比较。SVM 、 RF 等模型的超参数设置如表七所示。

案例一中的十种方法在三个数据集（ CU0 、 MT0 和 MS0 ）上的结果如表 VIII 所示，每个数据集上的最佳结果以粗体突出显示。从表中可以看出，BFKG 在多个数据集上表现出了分类的优越性。为了更直观地展示 BFKG 的优越性，将结果以条形图的形式呈现在图 6 中。实验表明，利用特征 - 故障相关性进行决策树节点分裂属性选择可以促进随机森林的分类，即基于 KG 的 WRF 推理算法与其他机器学习算法和简单的神经网络模型相比具有一定的优势。 GCN 利用

表 7 其他分类方法设置

Method	Structure description
SVM	C=1, Kernel: rbf, gamma=$1/f_n$, where f_n represents the dimension of the input feature vector, using rbf kernel.
RF	The number of decision trees in a random forest is 40, which is consistent with the number of decision trees in BFKG.
ANN	Number of neurons per layer:(number of feature)-256-512-256-(number of classes), Epochs=500, batch size=128.
1D-CNN	Structure of each layer:Input-Conv1D-MaxPooling-Conv1D-MaxPooling-Flatten-Dense-Dropout-Dense, Epochs=50, batch size=32.
CNN-PCA-FCM	The number of principal components extracted by PCA is 10, and the remaining parameters refer to the original paper.
Improved RF	Detailed parameters are set at [34].
DRSN-CS	Detailed parameters are set at [35].
GCN	Contains two graph convolutional layers and a fully connected layer.Number of neurons per layer:(number of feature)-32-32-(number of classes), Epochs=500.
WRF-OOB	The parameters are consistent with BFKG.

表 VIII BFKG 与其他方法的比较

Dataset	MT0	CU0	MS0
BFKG	**0.8500**	**0.9857**	**0.9556**
RF	0.8350	0.9750	0.9444
SVM	0.8400	0.9678	0.7944
ANN	0.7200	0.9607	0.9389
1D-CNN	0.7300	0.9607	0.8167
CNN-PCA-FCM	0.8350	0.9500	0.8500
Improve RF	0.8450	0.9821	0.9500
DRSN-CS	0.7350	0.9750	0.9333
GCN	0.8133	0.9321	0.9000
WRF-OOB	0.8450	0.9786	0.9500

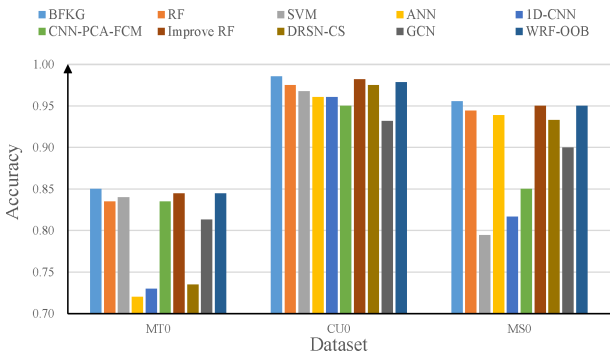


图 6。数据 MT0 、 CU0 和 MS0 的实验结果。结果表明，与其他方法相比，BFKG 更加稳定并保持最佳精度。

样本之间的关系较好，得到的故障分类精度较好。改进的 RF 和 WRF-OOB 在传统 RF 的基础上进行了改进，都取得了不错的效果。

表九 是否采用策略结果比较

Dataset	T	Strategy	Test1	Test2	Test3	Test4
MT1	0.75	NO strategy	0.6950	0.7150	0.7400	0.7550
		Data accumulate	0.6950	0.7200	0.7600	0.7700
CU1	0.50	NO strategy	0.9143	0.9286	0.9071	0.9357
		Data accumulate	0.9143	0.9321	0.9393	0.9429
MS1	0.55	NO strategy	0.8833	0.9111	0.8500	0.8667
		Data accumulate	0.8833	0.9167	0.8611	0.8778

C. 案例二：数据积累效果验证

在案例二中，需要为不同的数据集设置不同的阈值 T。本案例的数据集分为五部分，其中一部分作为原始训练集，另外四部分作为测试集。实验组将遵循第三节中的步骤。首先使用训练集训练模型，然后将这个测试集中值得保留的数据添加到模型的下一个训练集中，记录测试集的准确率，重复上述过程，直到四个测试集都测试完毕。对照组实验：使用训练集训练模型，然后测试四个测试集，得到每个测试集的准确率。

模型是否实现数据积累策略的实验结果如表 9 所示，其中包括三个数据集（ CU1 、 MT1 和 MS1 ）上的结果。实验结果表明，实施数据积累策略后的模型的分类准确率高于未实施数据积累策略的模型。由于模型首次训练时的训练集和其他条件相同，因此 Test1 的准确率对于每个数据集的结果都是相同的，这也反映在表中。在图 7 中，使用折线图更清楚地展示了数据积累策略的有效性。

将数据积累策略引入到其他方法中。这些方法的参数设置与案例一相同，各方法的实验均按照上述流程进行。实现数据积累策略的每种方法的实验结果如表 X 所示，其中包括三个数据集（ MT1 、 CU1 和 MS1 ）和各种方法的结果。

实验表明，该方法在数据积累后的分类精度总体优于其他方法。其中，一维 CNN 和 ANN 的分类精度较差。原因可能是初始训练所用的数据样本太少，而传统神经网络模型的效果取决于训练集的样本量；否则，可能会出现训练过度拟合，导致测试结果不佳。 GCN 充分利用了样本之间的关系，因而具有良好的效果。该方法充分利用了样本之间的关系

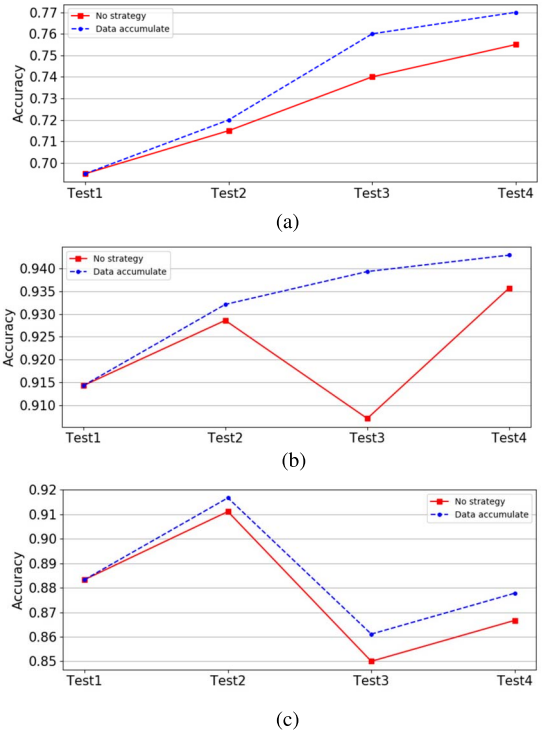


图 7. 数据集 MT1 、 CU1 和 MS1 上的实验结果。 (a) 案例 II 在数据集 MT1 上执行的结果。 (b) 在数据集 CU1 上执行案例 II 的结果。 (c) 在数据集 MS1 上执行案例 II 的结果。

表 X 使用数据积累策略的其他方法的比较

Dataset	Methods	Test1	Test2	Test3	Test4
MT1	BFKG	0.6950	**0.7200**	**0.7600**	**0.7700**
	SVM	**0.7350**	0.6850	0.6800	0.7500
	ANN	0.4100	0.5400	0.5300	0.6100
	1D-CNN	0.3200	0.3300	0.3250	0.4300
	Improve RF	0.6950	0.6750	0.6800	0.7550
	DRSN-CS	0.6700	0.6450	0.7000	0.7150
	GCN	0.5667	0.4733	0.6000	0.3333
	WRF-OOB	0.6750	0.6850	0.6600	0.7600
CU1	BFKG	0.9143	**0.9321**	**0.9393**	**0.9429**
	SVM	**0.9179**	0.8857	0.8821	0.8924
	ANN	0.7750	0.8036	0.7714	0.7500
	1D-CNN	0.5286	0.6500	0.5964	0.5929
	Improve RF	0.9036	0.9143	0.9214	0.9357
	DRSN-CS	0.9107	0.9071	0.9250	0.9179
	GCN	0.8629	0.8679	0.8750	0.8929
	WRF-OOB	**0.9179**	0.9071	0.9357	0.9357
MS1	BFKG	**0.8833**	**0.9167**	**0.8611**	**0.8778**
	SVM	0.8222	0.8000	0.8222	0.8500
	ANN	0.8056	0.8056	0.8444	0.7722
	1D-CNN	0.6000	0.4722	0.4722	0.5556
	Improve RF	0.8278	0.8667	0.8344	0.8611
	DRSN-CS	0.8389	0.8556	0.8222	0.8167
	GCN	0.7278	0.8056	0.7444	0.7500
	WRF-OOB	0.8222	0.8778	0.8389	0.8722

特征和故障类别并优化随机森林算法以获得最佳结果。

D. 案例三：BFKG 的稳健性

为了验证所提方法的鲁棒性，添加了不平衡数据集实验。基于

表 XI 不平衡 CWRU 数据集描述

Health condition	Label	The ratio of training samples (%)	The ratio of testing samples (%)
N	1-4	80	20
IR	5-12	50	20
OR	13-20	30	20
B	21-28	20	20

表 XII 不平衡数据集的实验结果

Dataset	Imbalance
BFKG	**0.9643**
RF	0.9571
SVM	0.9071
ANN	0.9357
1D-CNN	0.6750
CNN-PCA-FCM	0.8643
Improve RF	0.9571
DRSN-CS	0.9321
GCN	0.9107
WRF-OOB	0.9537

CWRU 数据集，通过调整不同状态下的样本数量来构建不平衡的数据集。表 XI 中显示了不平衡数据集的详细信息，而不是每个州有 50 个样本。四种健康状态各自的训练样本数为 40。八种内环故障状态各自的训练样本数为 25。八种外环故障状态各自的训练样本数为 15。八种钢球故障状态各自的训练样本数为 10。

使用十种不同的方法对该数据集进行训练，并记录每种方法在训练集上的准确性，如表十二所示。从表 XII 可以看出，与平衡数据集相比，所提出的方法仍然具有最高的精度，并且精度下降较小。表明该方法能够有效处理训练数据样本的不平衡问题，具有较强的鲁棒性。

V. 结论

本文提出了一种基于 KG 的轴承诊断方法。该方法提供了一种构建知识图谱的方法，以信号特征和故障类别作为知识图谱的节点，以特征与故障的相关性作为知识图谱的边。该方法构建的知识图谱可以反映不同故障类别下各个特征的重要性，为后续的故障推理和诊断奠定基础。提出了一种改进的随机森林算法，该算法充分利用了特征与故障的相关性，相对于传统的分类算法具有一定的优势。此外，提出了知识图谱的知识积累和更新策略，以充分利用数据来实现训练数据集的逐步扩展，提高后续数据分类的准确性。

使用两个经典数据集和一个实验室数据集来实验所提出的方法。实验的结果验证了 WRF 算法的优越性，表明使用该算法可以充分利用 KG 提供的信息。因此，用它作为 KG 的推理算法是非常合适和有效的。实验结果表明，数据积累后的分类准确率明显高于没有数据积累情况下的分类准确率，说明数据积累策略的有效性，可以有效增加 KG 中的数据量，从而提高推理精度。同时，该策略也可以为其他类型 KG 的数据积累提供研究思路。但本文使用的数据集均为恒速轴承振动数据，而在实际工况下，机械系统的振动信号大多为时变信号。因此，在后续工作中，我们将尝试使用时变信号数据集，并采用时变信号的特征提取方法。另外，本文数据积累策略部分使用的阈值 T 是根据模型对数据集的分类效果粗略制定的，需要在后续工作中进行详细分析并准确赋值。

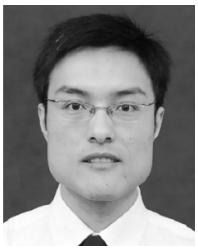
参考

[1] Z。杨正宇，谢春，黄宇，“Hilbert-Huang 变换在声发射信号中的应用，用于表面磨削过程烧伤特征提取”，*Measurement*，第 1 期。 47，第 14-21 页，2014 年 1 月。[2] 问。胡志强，张世良，杨世良，“基于时域和人工智能的变工况轴承故障诊断”，载 *Applied Mechanics and Materials*，第 1 期。 203. 沃勒劳，瑞士：Trans Tech Publications，2012 年，第 329–333 页。[3] Z. 霍，张，舒，M. Gallimore，“一种基于细到粗的多尺度排列熵、拉普拉斯评分和支持向量机的轴承故障诊断新方法”，*IEEE Access*，第 1 卷。 7，第 17050–17066 页，2019 年。[4] M. Kang，J. Kim，J. M. Kim，“基于 FPGA 的多核系统，利用超采样率 AE 信号进行实时轴承故障诊断”，*IEEE Trans. Ind. Electron.*，卷。 62，没有。 4，第 2319–2329 页，2015 年 4 月。[5] M. 崔永旺，林晓，钟明，“基于改进的堆栈自编码器和支持向量机的滚动轴承故障诊断”，*IEEE Sensors J.*，第 1 期。 21、没有。 4，第 4927–4937 页，2021 年 2 月。[6] J. 熊强，张庆，孙刚，朱旭，刘明，李子，“一种基于静态贴现因子和 KNN 无量纲指标的信息融合故障诊断方法”，*IEEE Sensors J.*，第 1 期。 16、没有。 7，第 2060–2069 页，2016 年 4 月。[7] Y。 Benmahamed、 M. Teguar 和 A. Boubakeur，“SVM 和 KNN 在 duval pentagon 1 变压器油诊断中的应用”，*IEEE Trans. Dielectr. Electr. Insul.*，卷。 24、没有。 6，第 3443–3451 页，2017 年 12 月。[8] S. E. Pandarakone、 S. Gunasekaran、 Y. Mizuno 和 H. Nakamura，“朴素贝叶斯分类器定理在检测感应电机轴承故障中的应用”，*Proc. XIII Int. Conf. Electr. Mach. (ICEM)*，2018 年 9 月，第 1761–1767 页。[9] J. 史晓武、周杰、王树，“基于差分进化和 EEMD 降噪的 BP 神经网络轴承故障诊断”，载 *Proc. 9th Int. Conf. Modelling、Identificat. Control (ICMIC)*，2017 年 7 月，第 1038-1043 页。[10] L. 江强，李建，崔建，席建，“基于高阶累积量和 BP 神经网络的滚动轴承故障诊断”，载 *Proc. 27th Chin.Control Decis. Conf. (CCDC)*，2015 年 5 月，2664-2667 页。[11] 沉书生，贾芳，左浩，马建，“机器智能故障诊断的深度多标签学习框架”，*IEEE Access*，第 1 卷。 8，第 113557–113566 页，2020 年。[12] D. T. Hoang 和 H. J. Kang，“基于深度学习和信息融合的基于电机电流信号的轴承故障诊断”，*IEEE Trans.Instrum. Meas.*，卷。 69，没有。 6，第 3325–3333 页，2020 年 6 月。[13] S. 高，裴正，张勇，李涛，“基于 Nesterov 动量自适应卷积神经网络的轴承故障诊断”，*IEEESensors J.*，第 1 期。 21、没有。 7，第 9268–9276 页，2021 年 4 月。

[14] X。王春马、刘鹏、潘斌、康子，“智能能源管理的潜在解决方案 —— 知识图谱”，载 *Proc. IEEE Int.Conf. Energy Internet (ICEI)*，2018 年 5 月，第 281-286 页。[15] Alshahrani 、 M. A. Khan 、 O. Maddouri 、 A. R. Kinjo 、 N. Queralt-Rosinach 和 R. Hoehndorf，“生物知识图谱上的神经符号表示学习”，*Bioinformatics*，卷。 33 、没有。 17，第 2723–2730 页，2017 年 9 月。[16] X。 Yao 和 B. Van Durme，“结构化数据的信息提取：使用 freebase 回答问题”，*Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*，2014 年，第 956–966 页。[17] A. Fader 、 L. Zettlemoyer 和 O. Etzioni，“针对策划和提取的知识库进行开放式问答”，*Proc. 20th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining*，2014 年 8 月，第 1156–1165 页。[18] M. 盛，胡，Y. 张，C.Xing，和 T.Zhang，“基于知识图谱的数据密集型 CDSS 平台”，载于 *Proc. Int. Conf. Health Inf. Sci.* 瑞士 Cham：Springer，2018 年，第 146-155 页。[19] G. A. Miller，“WordNet：英语词汇数据库”，*Commun. ACM*，卷。 38，没有。 11，第 39–41 页，1995 年。[20] C. Bizer *et al.*，“DBpedia - 数据网络的结晶点”，*J. Web Semantics*，卷。 7 、没有。 3，第 154–165 页，2009 年 9 月。[21] D. Vrande˘ci´c 和 M. Krötzsch，“维基数据：免费协作知识库”*Commun. ACM*，卷。 57 、没有。 10，第 78-85 页，2014 年。[22] R. Miao，X.Zhang，H.Yan，C.Chen，“基于强化学习和迁移学习的动态金融知识图”，*Proc. IEEE Int. Conf. Big Data (Big Data)*，2019 年 12 月，第 5370-5378 页。[23] Z. Lin，D. Yang，X. Yin，“通过医学知识图和医学实体描述的联合嵌入实现患者相似性”，*IEEE Access*，卷。 8，第 156663–156676 页，2020 年。[24] K. Dragomiretskiy 和 D. Zosso，“变分模态分解”，*IEEE Trans. Signal Process.*，卷。 62，没有。 3，第 531–544 页，2014 年 2 月。[25] S. Y. Lu，K.-H。许和 L.-J。 Kuo，“基于 WordNet 和 SWRL 规则的语义服务匹配方法”，*Proc. IEEE 10th Int. Conf. e-Bus. Eng.*，2013 年 9 月，第 419–422 页。[26] T.-W. Lee 、 M. S. Lewicki 、 M. Girolami 和 T. J. Sejnowski，“使用过完备表示法对更多来源进行盲源分离，”*IEEE Signal Process. Lett.*，卷。 6 、不。 4，第 87–90 页，1999 年 4 月。[27] R. Socher 、 D. Chen 、 C. D. Manning 和 A. Ng，“利用神经张量网络进行推理以完成知识库”，*Proc. Adv. NeuralInf. Process. Syst.*，2013 年，第 926–934 页。[28] L. Breiman，“随机森林”，*Mach. Learn.*，卷。 45，没有。 1，第 5–32 页，2001 年。[29] M。罗春莉，张晓，李瑞，安小安，“滚动轴承故障诊断的极限学习机复合特征选择与参数优化”，*ISA Trans.*，第 1 期。 65，第 556–566 页，2016 年 11 月。[30] Y。雷志和，訾勇，胡强，“基于多 ANFIS 与气体结合的旋转机械故障诊断”，*Mech. Syst. SignalProcess.*，2014 年第 1 期。 21 、没有。 5，第 2280–2294 页，2007 年 7 月。[31] MFPT。*Home-Society for Machinery Failure Prevention Technology*。访问日期：2021 年 7 月 13 日。[在线]。可用：https://www.mfpt.org/[32] CWRU。*Bearing Data Center*。访问日期：2021 年 7 月 13 日。[在线]。可用：https://csegroups.case.edu/bearingdatacenter/home[33] D. 张，陈勇，郭凤，H.R.Karimi，董浩，Q. 轩，“滚动轴承故障诊断的一种新的可解释学习方法”，*IEEE Trans. Instrum. Meas.*，第 1 期。 70，第 1-10 页，2021 年。[34] X。董国光，贾勇，徐坤，“谱图小波变换结合改进随机森林的图视角多尺度特征提取滚刀故障诊断”，*Measurement*，2015 年第 1 期。176，2021 年 5 月，艺术。不。 109178。 [在线]。可用：https://www.sciencedirect.com/science/article/pii/S0263224121001986[35] M。赵世钟，付旭，唐波，M. Pecht，“用于故障诊断的深度残差收缩网络”，*IEEE Trans. Ind. Informat.*，第 1 卷。 16 、没有。 7，第 4681–4690 页，2020 年 7 月。[36] Y。高明，陈明，余德，“半监督图卷积网络及其在旋转机械智能故障诊断中的应用”，*Measurement*，第 1 期。 186，2021 年 12 月，艺术。不。 110084。[在线]。可用：https://www.sciencedirect.com/science/article/pii/S026322412101006X[37] M。 Shahhosseini 和 G. Hu，“针对分类问题的改进加权随机森林”，*Proc. Int. Online Conf. Intell. Decis. Sci.*Cham，瑞士：Springer，2020 年，第 42-56 页。

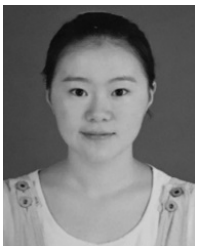
肖向渠获得学士学位他于 2019 年在中国武汉华中科技大学 (HUST) 获得水利水电工程学士学位，目前正在攻读博士学位。土木与水利工程学院博士学位。

研究方向为非平稳信号处理和旋转机械故障诊断。



李朝顺 (IEEE 会员) 获得理学学士学位 2005 年毕业于武汉大学热能与动力工程专业，获工学博士学位； 2010 年毕业于华中科技大学水利水电工程专业，获工学博士学位。

现任华中科技大学水电与信息工程学院教授。研究兴趣包括机器学习、智能优化、控制理论旋转机械及其应用和故障诊断。



黄杰获得理学学士学位 2019 年毕业于昆明理工大学能源与动力工程专业，获工学博士学位。目前正在攻读博士学位。毕业于华中科技大学土木水利工程学院，获学士学位。

研究方向为非平稳信号处理和旋转机械故障诊断。

田宇获得学士学位 2018 年毕业于华北水利水电大学（郑州），获博士学位。目前正在攻读博士学位。毕业于华中科技大学土木水利工程学院，获学士学位。

她目前的研究兴趣包括机械故障诊断和强化学习。