## 特征筛选的原因与背景

轴承故障诊断中的特征筛选是一个关键步骤，其主要原因在于原始数据包含51个特征，存在特征冗余、噪声干扰和维度灾难等问题。通过科学的特征筛选，我们将特征数量从51个减少到26个，减少比例达到49%，既保留了关键信息又提高了模型效率。特征筛选的核心目标是识别对故障诊断最有价值的特征，去除冗余和无关特征，提升分类性能。

## 特征筛选的依据与方法

### 1. 基于随机森林的重要性评估

采用随机森林算法计算每个特征的重要性得分，这是一种基于树模型的特征重要性评估方法，能够有效识别对分类贡献最大的特征。从 特征重要性分析图 可以看出，前10个最重要的特征分别是：snr\_db（0.120）、dynamic\_range（0.099）、zero\_crossing\_rate（0.089）、freq\_rms（0.064）等，这些特征在故障识别中起到关键作用。

### 2. 基于特征类别的平衡选择

将特征按照物理意义分为六大类别：时域统计、时域形状、频域统计、频域能量、调制冲击和其他特征。从 特征类别分析图 可以观察到，频域能量特征类别包含13个特征，总重要性达到0.331，是故障诊断的核心；调制和冲击特征虽然数量较少（3个），但平均重要性最高（0.061），说明这类特征对故障识别具有很强的区分能力。

### 3. 基于相关性分析的冗余去除

通过计算特征间的皮尔逊相关系数，识别高相关性特征对（|r| > 0.7）。从 特征相关性分析图 中发现13对高相关性特征，如energy与freq\_rms的相关系数达到0.938，zero\_crossing\_rate与band\_1\_energy\_ratio的相关系数为-0.932。在保留重要性更高特征的前提下，去除冗余特征以避免信息重复。

## 特征筛选的具体步骤过程

### 第一步：数据预处理与特征提取

从原始轴承振动信号中提取51个特征，涵盖时域、频域、时频域等多个维度的特征。这些特征包括统计特征（均值、方差、偏度、峭度）、能量特征、频谱特征、调制特征等。

### 第二步：特征重要性计算

使用随机森林分类器训练模型，计算每个特征的重要性得分。重要性得分反映了特征在决策树分裂过程中的贡献程度，得分越高表示该特征对分类的贡献越大。

### 第三步：特征类别分析

将特征按照物理意义进行分类，分析每个类别的特征数量、总重要性、平均重要性和最高重要性。这种分析有助于理解不同类型特征在故障诊断中的作用，确保选择的特征具有物理可解释性。

### 第四步：相关性分析与冗余去除

计算特征间的相关性矩阵，识别高相关性特征对。对于相关系数绝对值大于0.7的特征对，保留重要性更高的特征，去除冗余特征。这一步骤有效减少了特征间的多重共线性问题。

### 第五步：综合评估与最终选择

综合考虑特征重要性、类别平衡性和相关性分析结果，最终选择26个特征。选择标准包括：重要性得分较高、物理意义明确、相关性较低、类别分布均衡。

## 基于可视化图表的深入分析

### 1. 特征类别分布分析

从饼图可以看出，频域能量特征占据主导地位（50%），这符合轴承故障诊断的物理机理——轴承故障通常表现为特定频率成分的能量变化。时域形状特征（15.4%）和调制冲击特征（11.5%）也占有重要比例，说明故障信号的非线性特征和冲击特性对诊断具有重要价值。

### 2. 特征重要性对比分析

从条形图可以观察到，不同类别特征的重要性存在显著差异。"其他特征"类别虽然特征数量少，但总重要性和平均重要性都很高，特别是snr\_db特征的重要性达到0.331，成为最重要的单一特征。这表明信号质量相关特征对故障诊断至关重要。

### 3. 特征相关性网络分析

相关性热力图揭示了特征间的复杂关系网络。高相关性特征主要集中在同类特征之间，如能量相关特征（energy、freq\_rms、log\_energy）和频谱特征（spectral\_centroid、spectral\_spread）。这种相关性模式验证了特征分类的合理性，同时指导了冗余特征的去除策略。

### 4. 降维可视化效果分析

PCA和t-SNE降维结果显示，选择的特征能够有效区分不同故障类别。PCA前两个主成分解释了约32%的方差，虽然解释率不高，但不同故障类别在降维空间中呈现出明显的聚类趋势。t-SNE结果更加清晰地展现了故障类别的可分离性，验证了特征选择的有效性。

### 5. 故障类别对比分析

箱线图展示了Top 8重要特征在不同故障类别间的分布差异。可以观察到，不同特征对不同故障类别的区分能力存在差异。例如，snr\_db特征在正常状态（N）和故障状态间存在显著差异，而dynamic\_range特征在内圈故障（IR）和外圈故障（OR）间表现出良好的区分能力。

## 筛选结果的验证与意义

最终筛选出的26个特征在保持高分类性能的同时，显著降低了计算复杂度。这些特征涵盖了轴承故障的多个物理机制：振动信号的统计特性、频谱分布、能量集中度、调制特征和信号质量等。特征选择结果不仅具有良好的数学性质（低冗余、高重要性），更重要的是具有明确的物理意义，为后续的故障诊断模型提供了可靠的输入基础。