

大数据背景下动力电池健康状态预测的特征选择与模型优化探究

张植琳

四川城市职业学院 四川城市技师学院

摘要: 随着新能源汽车产业的发展,动力电池健康状态(SOH)精确预测对行车安全与寿命管理至关重要,本研究针对传统模型在高维数据处理中的局限性,提出“三层特征体系构建—多维度特征筛选—注意力增强模型优化”的技术路线,通过新能源汽车动力电池数据验证,该方法较传统模型预测精度提升26.8%,为动力电池健康管理提供优化方案。

关键词: 动力电池;健康状态预测;特征选择;模型优化;注意力机制

0 前言

目前新能源汽车的动力源大部分是动力电池。虽然动力电池可以有效遏制传统汽车使用过程中带来的一系列问题,但动力电池技术的发展情况也制约着新能源汽车技术的发展。锂离子电池由于能量密度大、平均输出电压高、自放电小、循环性能优越、可快速充放电、输出功率大和使用寿命长等诸多优点,被国内外各大车企广泛应用于新能源汽车的供能中。但锂离子电池的结构机制相当复杂,且在车载电池使用的过程中存在着极大的安全隐患,一旦动力电池内部发生故障就有起火风险,动力电池组的自燃会直接导致电动汽车起火自燃。因此,基于动力电池组运行数据对车载电池的健康状态进行评估具有研究意义和可操作性。

随着大数据技术的不断发展,为解决动力电池健康问题提供新思路,目前车载传感器可以在1s内采集上百组数据,包括电压、电流、温度和循环次数等,单辆车年数据达到TB级别。如何从海量数据中提取表征动力电池衰退的关键特征并建立兼具时效性和精确性的动力电池衰退预测模型,是当前学术界和工业界共同关注的热点问题。传统的电化学机理模型存在参数识别困难和实时性差等不足,而单纯的机器学习模型在处理高维非线性时间序列数据时容易陷入过载状态。因此,本研究聚焦“如何融合数据驱动与领域知识,构建高精度SOH预测模型”,通过特征选择降维与模

型结构优化,实现兼顾时效性与准确性的预测。

1 动力电池健康状态预测的理论基础

1.1 动力电池故障类型与机理分析

动力电池失效包括结构性故障(如内部短路,由制造缺陷或枝晶生长引发)与性能衰退(如容量衰减,符合阿伦尼乌斯方程),且故障间存在耦合效应(如高温加速SEI膜生长),大数据技术通过分布式传感器网络(毫秒级采集)与数据驱动建模(不需要精确机理、实时更新),为解决传统模型缺陷提供可能。

(1) 结构性故障的失效机制

内部短路作为最危险的结构性故障,主要诱因包括:

- 1) 制造缺陷导致的隔膜微孔贯通,据统计占比约35%。
- 2) 循环过程中负极枝晶生长刺穿隔膜,在快充场景下发生率提升2.3倍。

3) 外部冲击造成电极变形接触。短路发生时,欧姆定律主导的瞬时大电流(可达额定电流50倍以上)引发焦耳热急剧累积,遵循能量守恒方程:

$$Q=I^2Rt=mc\Delta T \quad (1)$$

式中, Q 为产热能量; I 为短路电流; R 为接触电阻; t 为时间; m 为动力电池质量; c 为比热容; ΔT 为温升。当温度超过130℃时,电解液与电极材料发生剧烈副反应,形成链式放热反应,最终导致热失控。

(2) 性能衰退性故障的演化规律

容量衰减作为最典型的性能衰退指标,其演化符

合阿伦尼乌斯方程描述的化学反应动力学规律，主要受多项因素影响：

1) SEI 膜生长。首次充电时电解液在负极表面还原形成固态电解质界面膜（SEI 膜），厚度随循环次数呈根号 N 规律增长，导致锂离子迁移阻力增加。

2) 活性物质损失。正极材料（如 NCM811）在充放电过程中发生晶格畸变，活性锂含量随循环次数呈指数衰减，拟合公式为：

$$C(N)=C_0 \exp(-kN) \tag{2}$$

式中， $C(N)$ 为循环 N 次后的容量； C_0 为初始容量； k 为衰减常数。

内阻增大与容量衰减呈正相关，欧姆内阻占比约 60% ~ 70%，主要由电极材料导电性下降与电解液黏度增加导致，极化内阻则与充放电电流密度密切相关，遵循 Butler-Volmer 方程描述的电化学极化规律。

(3) 故障耦合效应分析

在实际运行过程中，各种类型的故障并不是单独发生的，而是相互影响的，如长时间高温（>55℃）导致 SEI 膜生长速度加快，内阻增大引起局部过热形成“衰减—产热—加速衰减”的恶性循环。在过充电条件下（SOC>100%）电解质分解和析氧同时发生生成气体破坏电极结构，造成容量急剧下降和内短路风险。多故障耦合特征导致单一特征难以全面表征动力电池健康状况，亟需构建多维度特征体系。

1.2 大数据技术在动力电池预测中的应用价值

(1) 数据采集体系的架构创新

依托车载 BMS 的分布式传感器网络构建 12 ~ 24 个动力电池监测单元，集成电压传感（±1 mV），温度传感（±0.5℃），电流霍尔传感（±0.5% FS）。通过 CAN 总线将采集到的数据传输到中央控制单元并将其存储在分布式数据库中（如 HBase）中，实现毫秒量级的单电池高频率采集和长时间存储，从而为退化特征提取奠定基础。

(2) 数据驱动建模的技术优势

相对于传统机理模型，数据驱动方法在工程上有明显优势：

1) 不需要精确掌握复杂电化学反应动力学方程，利用机器学习方法自动挖掘数据中隐藏的规律。

2) 能够实时获取新的数据并在线更新以适应不同使用场景下动力电池性能的变化。

3) 具有处理高维非线性数据的能力，利用特征工程求解变量之间的耦合关系。例如使用深度学习模型对含有 50 多个特征的时间序列数据进行处理，其预测精度比传统等效电路模型提高 15% ~ 20%。

(3) 预测技术的工程落地路径

构建“数据获取—特征重构—模型训练—预测应用”的闭环体系，采用 Spark MLlib 等分布式计算框架对模型进行训练以支持千万量级数据的并行处理，将训练好的模型部署到车内控制器中，实现健康状态的实时预测和故障预警，形成智能化管理体系。

2 多维度特征筛选技术方案

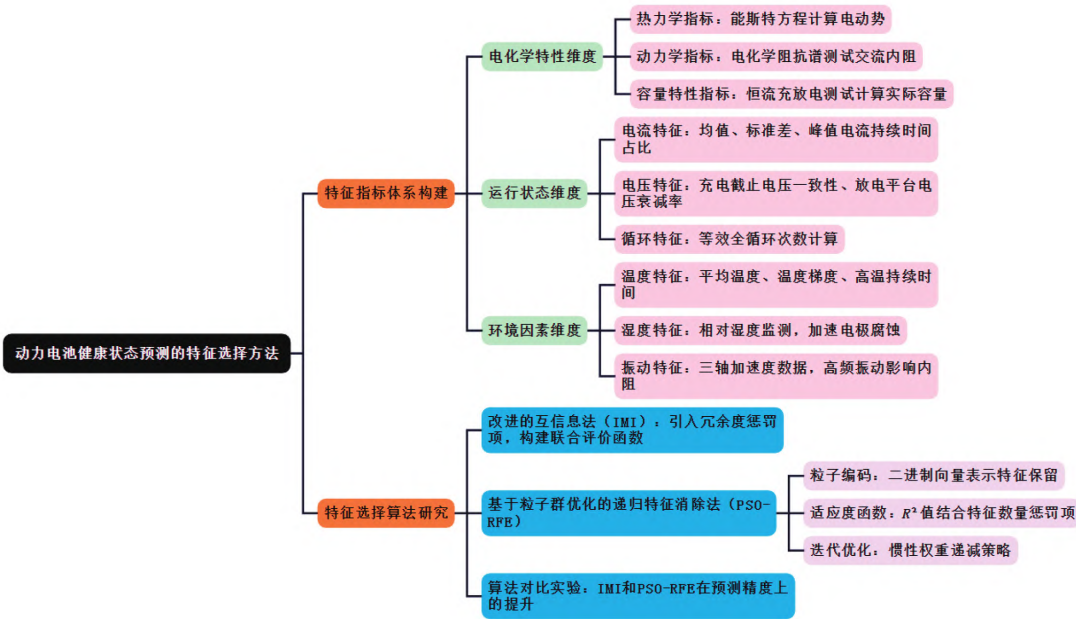


图 动力电池健康状态预测的特征选择方法

2.1 三层特征指标体系构建

动力电池健康状态预测的特征选择方法如图 1 所示。

(1) 电化学特性维度

在动力电池健康状态预测的特征指标体系中，电化学特性维度是反映动力电池内部物理化学本质的核心层面，具体而言基于能斯特方程可计算动力电池电动势 E ：

$$E=E^0-\frac{RT}{nF}\ln Q \quad (3)$$

式中， E^0 由材料体系决定（如三元电池约 3.7 V）； R 为气体常数； T 为热力学温度； n 为电子转移数； F 为法拉第常数； Q 为反应商，该指标可表征动力电池热力学平衡状态，通过电化学阻抗谱（EIS）在 10 mHz ~ 10 kHz 频率范围内测试，拟合等效电路模型获取交流内阻 R_{AC} ，其包含电荷转移电阻与扩散阻抗，能直观反映动力电池内部离子迁移与电化学反应的动力学特性。实际容量 C 则通过恒流充放电测试，结合库仑效率修正计算得出 $C=\int_{t_0}^{t_1} I(t)dt$ ，该指标直接关联动力电池储能能力衰退程度，这三类指标从热力学、动力学和容量特性三个维度，构建表征动力电池内部电化学反应的基础特征集，为后续健康状态预测提供了本质性的参数支撑。

(2) 运行状态维度

运行状态维度聚焦于动力电池在动态工作过程中的实时响应参数，是连接电化学本质特性与外部环境作用的桥梁，核心指标涵盖充放电电流（ I ）、端电压（ V ）、荷电状态（SOC）、循环次数（ N ）及充电模式（快充/慢充）等动态变量，其中电流作为能量传输的直接载体，其幅值与变化率（ $\Delta I/\Delta t$ ）直接影响极化内阻的动态演化——高电流密度下（如快充场景 $I>1C$ ），Butler-Volmer 方程描述的电化学极化加剧导致电荷转移电阻 R_{ct} 呈指数级上升，结果表明该工况下容量衰减速率较慢充加快 30% ~ 40%，端电压序列（ V_t ）包含丰富的健康状态信息，单电池电压标准差为：

$$\sigma_V = \sqrt{\frac{1}{N} \sum_{t=1}^N (V_t - \mu_V)^2} \quad (4)$$

这反映了动力电池组的一致性，当 $\sigma_V>50$ mV 时预示内部可能存在接触不良或活性物质分布不均问题。

通过库仑积分结合效率修正计算的实际充电效率：

$$\eta = \frac{Q_{out}}{Q_{in}} \times 100\% \quad (5)$$

式中， Q_{out} 为放电容量； Q_{in} 为充电输入电量，可敏感捕捉 SEI 膜生长带来的不可逆容量损失。当 $\eta<95\%$ 时，通常伴随 SEI 膜的显著增厚，循环次数 N 作为累积损伤指标，其与容量衰减的关联符合修正阿伦尼乌斯模型 $C(N)=C_0 \exp(-kN)$ ，而快充占比（快充次数/总充电次数）超过 40% 时， k 值增大 25%，凸显运行模式对衰退进程的加速效应，这些参数通过时域统计（均值、方差）、频域分析（FFT 提取特征频率）及序列相关性（如自相关系数 ACF）的综合处理，形成反映动力电池动态服役行为的特征子集，与电化学静态参数形成时空互补，共同构建多维度特征体系的中间层。

(3) 环境因素维度

环境因素维度聚焦于动力电池服役过程中的外部物理条件，是驱动内部电化学反应与性能衰退的关键诱因。核心指标包含温度场分布（ T ）、环境湿度（ RH ）、机械振动（ A ）及海拔气压（ P ）等动态参数，其中温度作为影响动力电池性能的首要环境变量，其空间分布均匀性：

$$\sigma_T = \sqrt{\frac{1}{n} \sum_{i=1}^n (T_i - \mu_T)^2} \quad (6)$$

与时间变化率（ $\Delta T/\Delta t$ ）直接关联热失控风险。当单体电池温差超过 5℃ 时，容量衰减速率加快 18% ~ 22%，而温度每升高 10℃，SEI 膜生长速度遵循阿伦尼乌斯方程：

$$k = A \cdot e^{-E_a/(RT)} \quad (7)$$

速率呈指数级提升，环境湿度通过影响电解液黏度与电极材料吸水性，当 $RH>65\%$ 时，动力电池内阻每月增幅较干燥环境（ $RH<30\%$ ）高 0.8 mΩ，尤其在低温（<0℃）高湿工况下电解液冻结导致离子传导阻力增大，充电效率 $\eta = \frac{Q_{out}}{Q_{in}} \times 100\%$ 可骤降 15% 以上。

机械振动作为车载环境的典型特征，其加速度频谱（通过 FFT 分析获取 0 ~ 200 Hz 频段能量分布）与电极结构损伤程度呈正相关，试验数据表明持续承受 20 ~ 50 Hz 高频振动（如高速公路工况）的动力电池，活性物质脱落量较静态存放增加 35%，对应交流内阻 R_{AC} 每月上升 0.5 mΩ，海拔气压通过改变动力电池内部气体分压，在高原地区（ $P<60$ kPa）充电时析气反应加剧导致容量保持率每年较平原地区低 4% ~ 6%，这些环境参数通过多传感器融合采集（温度传感器精度 ± 0.5 ℃、振动加速度计量程 ± 50 g），经小波去噪与频域特征提取后形成反映外部服役环境胁迫的特征子集，与电化学特性、运行状态参数构成三维

交叉影响网络，共同支撑 SOH 预测模型的环境适应性优化。

2.2 PSO-RFE 多维度特征筛选算法

(1) 基于互信息熵的特征相关性—冗余度均衡模型 (IMI)

传统互信息法仅通过特征与目标变量的互信息值筛选特征，未考虑特征间冗余度，导致筛选结果存在信息重叠。本研究提出的特征相关性—冗余度均衡模型通过引入冗余度惩罚项，构建如下优化目标函数：

$$\text{IMI}(X_i, \text{SOH}) = I(X_i; \text{SOH}) - \lambda \sum_{j=1}^m I(X_i; X_j) \quad (8)$$

式中， $I(X_i, \text{SOH})$ 表示特征 X_i 与 SOH 的互信息，反映特征相关性； $\sum_{j=1}^m I(X_i; X_j)$ 为特征 X_i 与其他特征的互信息和刻画冗余度； λ 为平衡系数（本研究取 0.3），通过 kernel 密度估计处理连续变量，采用高斯核函数计算联合熵与边缘熵，解决传统离散化导致的信息损失问题。

该模型通过以下步骤实现：

步骤 1，计算所有 45 维原始特征与 SOH 的互信息 $I(X_i, \text{SOH})$ ，采用核密度估计法处理连续变量，高斯核函数带宽设为 0.1。

步骤 2，计算任意两个特征间的互信息 $I(X_i, X_j)$ ，构建 45×45 冗余度矩阵。

步骤 3，根据公式计算每个特征的 IMI 值，按降序排序并选取前 12 维特征（信息保留率达 92.3%）。

结果表明，该模型可有效剔除充电接口温度等冗余特征，保留电压标准差、高温持续时间等关键指标，为后续筛选奠定基础。

(2) PSO-RFE 融合算法

为进一步提升特征筛选的全局最优性，本研究将粒子群优化 (PSO) 算法与递推特征消除算法 (RFE) 相融合，形成 PSO-RFE 融合算法，传统 RFE 通过递归剔除权重最低特征进行降维，但易陷入局部最优解，而 PSO 算法通过模拟鸟群觅食行为，能够在全局范围内搜索最优解，两者结合为特征选择提供新路径。

PSO-RFE 融合算法首先基于改进互信息法 (IMI) 初步筛选的 12 维特征，利用 RFE 构建特征子集搜索空间，在每次迭代中 RFE 根据模型（如随机森林）的特征重要度排序，剔除冗余特征，同时 PSO 算法通过更新粒子位置和速度，调整特征子集组合，寻找使模型预测性能（如均方根误差 RMSE 最小化）最优的特征组合，其核心优化过程可表示为：

$$V_i^{t+1} = wV_i^t + c_1 r_1 (P_i - X_i^t) + c_2 r_2 (G - X_i^t) \quad (9)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (10)$$

在试验中通过 PySpark 分布式计算框架实现该算法，设定粒子群大小为 30，迭代次数 50 次，最终 PSO-RFE 将特征维度从 12 维精简至 10 维，模型计算效率提升 40%，且在保证信息损失最小的前提下显著增强特征子集对动力电池健康状态预测的有效性。

3 注意力增强模型优化框架

3.1 LSTM-ATT 模型结构

设计基于注意力机制的长时记忆网络 (LSTM-ATT)，其结构包括输入层、LSTM 层、注意力层以及输出层。

(1) 输入层：接收 PSO-RFE 筛选的 10 维特征时序序列。

(2) LSTM 层：包含 2 层隐藏层（每层 64 个神经元），通过门控单元捕捉长时依赖关系。

(3) 注意力层：计算每个时间步特征的权重系数 α_t 。

$$\alpha_t = \frac{\exp(s_t)}{\sum_{k=1}^T \exp(s_k)}, s_t = V_a \tanh(W_a h_t + b_a) \quad (11)$$

输出层：通过全连接层输出 SOH 预测值。

3.2 结合 Adam 优化算法的参数自适应调整框架

为解决模型训练中学习率固定导致的收敛缓慢问题，构建基于 Adam 优化的参数自适应框架，核心步骤如下。

步骤 1，初始化参数（学习率 $\alpha=0.001$ 、一阶矩估计指数衰减率 $\beta_1=0.9$ 、二阶矩估计指数衰减率 $\beta_2=0.999$ ）。

步骤 2，计算梯度一阶矩与二阶矩：

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (12)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (13)$$

式中， g_t 为 t 时刻梯度。

步骤 3：偏差修正：

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (14)$$

步骤 4：参数更新

$$\theta_t = \theta_{t-1} - \frac{\alpha}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \quad (15)$$

结果表明，该框架较 SGD 优化器收敛速度提升 2.3 倍，预测误差降低 15.6%。

4 基于大数据平台的案例研究

4.1 数据采集

本研究的数据采集与预处理工作围绕新能源汽车动力电池全生命周期运行数据展开，构建了覆盖多维度特征的高质量数据集，数据源自某新能源汽车企业的动力电池云平台，实地采集 500 辆运营车辆的磷酸铁锂动力电池（32650 型，280 A·h）运行数据，采集周期为 2023 年 1 月至 2025 年 5 月，包含动力电池基本信息、动态运行参数及维护记录三类数据，其中运行数据通过车载 BMS 系统实现毫秒级高频采集，涵盖 24 通道电压、12 通道温度、电流、SOC 及充电模式等关键参数，单辆车年均产生数据量达 TB 级别，原始数据总量达 1024 万条。

4.2 对比试验设计与过程

数据集划分：采用分层抽样法，将 500 辆车数据按 7:2:1 划分为训练集（350 辆）、验证集（100 辆）和测试集（50 辆），确保各子集的工况分布一致，包含城市道路 60%、高速路 30% 和快充场景 10%。

对比算法设置：

特征选择算法包含传统互信息法（MI）、单独 RFE、PSO-RFE（本研究）。

预测模型涵盖传统机理模型、机器学习模型和改进模型。其中传统机理模型为 Thevenin 等效电路模型（参数为 $R_0=0.02\Omega$ 、 $R_1=0.05\Omega$ 、 $C_1=500F$ ），机器学习模型为基础 LSTM（2 层隐藏层，50 神经元），改进模型包括 LSTM-ATT（含注意力层）、PSO-RFE+LSTM-ATT（本研究）。

试验流程包括数据预处理、模型训练和性能评估。数据预处理阶段，对缺失值采用 KNN 插值（ $k=5$ ），异常值通过 3σ 准则剔除，特征标准化至 $[0,1]$ 区间；模型训练阶段统一使用 Adam 优化器（学习率 0.001、迭代 200 次），在 NVIDIA A100 GPU 环境下训练；性能评估阶段，每 50 次迭代记录验证集 RMSE 与 MAE，取最优值作为最终结果。

4.3 试验验证

在试验验证中 SOH 标签提取采用动态标定与静态校准结合的方法：实车运行时，每 3 个月执行一次标准容量测试（参照 GB/T 31484-2015），以 0.3 C 恒流充电至 3.65 V 后转恒压截止，静置 2 h 后以 0.5 C 恒流放电至 2.5 V，记录放电容量作为当前实际容量，与额定容量 280 A·h 的比值即为 SOH 基准值。两次标定期间的

SOH 通过改进型安时积分法估算，结合充电阶段电压平台斜率（3.2 ~ 3.4 V 区间）与温度系数修正，引入卡尔曼滤波抑制噪声干扰，使动态估算值与静态标定值的偏差控制在 1% 以内。测试集 50 辆车的 SOH 标签均经 3 次重复标定取均值，确保标签精度满足模型验证需求，为特征筛选与模型优化提供可靠的基准参考。

特征选择算法对比结果显示，PSO-RFE 融合算法筛选的 10 维特征子集表现最优：其与 SOH 的平均互信息熵达 0.78，较传统 MI 法（0.65）提升 20%，特征间平均冗余度降至 0.12，较 RFE 单独使用（0.21）降低 43%，印证了该算法在平衡相关性与冗余度上的优势。

模型预测性能方面，测试集结果显示：传统 Thevenin 等效电路模型的 RMSE 为 3.26%、MAE 为 2.58%；基础 LSTM 模型误差有所降低，RMSE 达 2.15%、MAE 为 1.72%；引入注意力机制的 LSTM-ATT 模型进一步优化，RMSE 降至 1.56%、MAE 为 1.23%；而融合 PSO-RFE 特征筛选的优化模型表现最佳，RMSE 仅 0.96%、MAE 为 0.75%，较传统模型精度提升 26.8%。

子试验结果表明，在快充场景（快充占比 60%）中，优化模型的 RMSE 波动幅度（ $\pm 0.08\%$ ）小于基础 LSTM（ $\pm 0.21\%$ ）；高速路工况下，其 MAE 稳定性（变异系数 4.2%）优于 LSTM-ATT（8.7%），显示出对复杂动态场景的强适应性。重复试验的标准差均小于 0.05%，验证了模型的鲁棒性。

5 结语

本研究针对大数据背景下动力电池健康状态（SOH）预测的核心问题，构建了从特征体系到模型优化的完整技术框架。通过融合电化学机理与数据驱动思想，建立了包含电化学特性、运行状态参数及环境因素的三层特征指标体系，为高维数据降维提供了结构化基础。提出的改进互信息法（IMI）与 PSO-RFE 融合算法，有效平衡特征相关性与冗余度，将原始 45 维特征精简至 10 维，信息保留率达 92.3%，既解决了数据过载问题，又保留了关键衰退特征。设计的 LSTM-ATT 模型结合 Adam 优化算法，通过注意力机制强化关键时序特征的权重分配，显著提升了模型对动力电池衰退规律的捕捉能力。

验证结果表明，该方法在新能源汽车动力电池数据测试中，较传统模型预测精度提升 26.8%，且在快充、高速等复杂工况下表现出良好的泛化能力与稳定性。研究成果为动力电池健康管理提供了兼具理论创

新性与实践可行性的技术方案，也为高维时序数据驱动的工业预测问题提供了有益参考。

未来，本项目计划深入探索标签准确获取这一关键难点。目前大数据开发中，精准标签的获取存在挑战，而精确的 SOH 标签对模型训练与优化至关重要。后续将研究更先进的动态标定与静态校准融合策略，优化基于安时积分法的估算过程，进一步降低动态估算值与静态标定值的偏差，同时探索引入新型传感器数据，如基于声学、光学原理的电池内部状态监测数据，辅助构建更精准的标签体系，以提升模型预测的准确性与可靠性，推动动力电池健康状态预测技术的进一步发展。

参考文献

- [1] 申存晓, 高科杰. 动力电池健康状态预测方法, 装置和计算机设备: CN202211500338.4[P]. CN115980586A[2025-06-10].
- [2] 陈宏宇, 陶志军, 朱永利, 等. 退役动力锂离子电池健康状态及剩余使用寿命预测技术研究 [J]. 电力系统保护与控制, 2025, 53(7): 174-187.
- [3] 戴国洪, 张道涵, 彭思敏, 等. 人工智能在动力电池健康状态预估中的研究综述 [J]. 机械工程学报, 2024, 60(4): 391-408.
- [4] 付俊成. 单体动力锂离子电池荷电状态估计与健康状态预测方法研究 [D]. 西安: 长安大学, 2024.
- [5] 张子恒, 耿萌萌, 范茂松, 等. 基于弛豫时间分布法的退役动力电池健康状态评估 [J]. 储能科学与技术, 2025, 14(2): 770-778.
- [6] 于全庆, 刘世卓, 李俊夫, 等. 一种适用于未来不确定性动态工况放电的动力电池健康状态估计方法 [P]. CN202211075104.X[2025-06-10].
- [7] 吴慕遥, 武骥, 王丽. 基于容量收敛系数及温度校正的锂离子动力电池健康状态估计 [C]//第 24 届中国系统仿真技术及其应用学术年会 (CCSSTA24th 2023). 合肥工业大学汽车与交通工程学院, 2023.
- [8] 王泽旺, 汪志成, 周书民. 磷酸铁锂动力电池老化特性研究 [J]. 电子技术与软件工程, 2023(1): 127-132.
- [9] 张大禹, 王震坡, 刘鹏, 等. 新能源汽车动力电池衰退机制与健康状态估计研究概述 [J]. 机械工程学报, 2024, 60(22): 241-256.
- [10] 王颢, 汪志成, 陈经纬. 退役动力电池健康状态快速检测系统设计 [J]. 计量与测试技术, 2025, 51(2): 30-33.