

引用格式: 张金文, 王文龙, 倪荣雨, 等. 基于梯度提升机的中国陆地生态系统土壤异养呼吸预测[J]. 浙江农林大学学报, 2025, 42(4): 774–783. ZHANG Jinwen, WANG Wenlong, NI Rongyu, *et al.* Prediction of soil heterotrophic respiration in China's terrestrial ecosystems based on gradient boosting machine[J]. *Journal of Zhejiang A&F University*, 2025, 42(4): 774–783.

基于梯度提升机的中国陆地生态系统土壤异养呼吸预测

张金文¹, 王文龙¹, 倪荣雨¹, 张彬梅¹, 曾爱聪¹, 郭福涛¹, 苏漳文^{1,2}

(1. 福建农林大学 林学院, 福建 福州 350002; 2. 漳州职业技术学院 石油化工学院, 福建 漳州 363000)

摘要: 【目的】极限梯度提升树 (XGBoost) 与轻量级梯度提升机 (LightGBM) 模型在梯度提升决策树框架下各具优势, 系统对比两者在土壤异养呼吸估算中的性能差异, 有助于深入挖掘梯度提升机在生态系统碳通量预测中的潜力, 并推动该类模型在大尺度碳循环模拟中的优化应用。【方法】基于全球土壤呼吸数据库 (SRDB), 构建了中国陆地生态系统的土壤异养呼吸及环境因子数据库, 利用 XGBoost 和 LightGBM 2 种梯度提升机模型对 2000—2023 年中国陆地生态系统土壤异养呼吸进行估算与对比分析, 并进一步探讨中国陆地生态系统土壤异养呼吸的空间分布趋势及其主要影响因素。【结果】①2 个模型均展现出较高的预测精度 (测试集决定系数均为 0.91), XGBoost 模型在训练集上表现出较强的拟合能力, LightGBM 模型则在测试集上能够更好地控制误差。②在 2000—2023 年, XGBoost 与 LightGBM 模型估算的中国陆地生态系统土壤异养呼吸年平均值分别为 299.57 和 294.60 $\text{g} \cdot \text{m}^{-2} \cdot \text{a}^{-1}$, 年际变化幅度分别为 19.51 和 32.43 $\text{g} \cdot \text{m}^{-2} \cdot \text{a}^{-1}$ 。③中国陆地生态系统土壤异养呼吸呈现南高北低的空间分布特征, 主要受土壤性质和叶面积指数影响。这一空间异质性反映了土壤异养呼吸对环境变化的不同响应。【结论】梯度提升机模型在大尺度土壤异养呼吸建模与预测中表现出良好的适应性, 能够有效捕捉土壤异养呼吸的时空变化特征, 展现出较强的预测能力。图 4 表 3 参 33

关键词: 土壤异养呼吸估算; 陆地生态系统; 极限梯度提升树 (XGBoost) 模型; 轻量级梯度提升机 (LightGBM) 模型
中图分类号: S153 文献标志码: A 文章编号: 2095-0756(2025)04-0774-10

Prediction of soil heterotrophic respiration in China's terrestrial ecosystems based on gradient boosting machine

ZHANG Jinwen¹, WANG Wenlong¹, NI Rongyu¹, ZHANG Binmei¹, ZENG Aicong¹, GUO Futao¹, SU Zhangwen^{1,2}

(1. College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China; 2. College of Petrochemical Technology, Zhangzhou Institute of Technology, Zhangzhou 363000, Fujian, China)

Abstract: [Objective] XGBoost and LightGBM models exhibit differentiated advantages within the gradient boosting decision tree framework. This study aims to systematically compare the performance differences between the two in estimating soil heterotrophic respiration (Rh), which will help to explore the potential of gradient boosting machines in ecosystem carbon flux prediction and promote the optimization of such models in large-scale carbon cycle simulations. [Method] A database of soil Rh and environmental factors in China's terrestrial ecosystems was constructed based on the Global Soil Respiration Database (SRDB). XGBoost and LightGBM models were used to estimate and compare soil Rh from 2000 to 2023. The spatial pattern and main influencing factors of soil Rh in China's terrestrial ecosystems were further investigated. [Result] (1) Both models demonstrated high predictive accuracy ($R^2 = 0.91$ for the test set). XGBoost model showed strong fitting ability on the training set, while LightGBM model performed better in controlling the error on the test set. (2) The annual average values of soil Rh estimated by XGBoost and LightGBM models from 2000 to 2023 were

收稿日期: 2024-12-25; 修回日期: 2025-05-30

基金项目: 国家重点研发计划项目 (2018YFE0207800); 中国博士后科学基金第 76 批面上资助项目 (2024M760460)

作者简介: 张金文 (ORCID: 0009-0002-4238-3567), 从事林火生态学研究。E-mail: 17550153994@163.com。通信作者: 苏漳文 (ORCID: 0000-0002-0890-7198), 讲师, 博士, 从事林火生态学研究。E-mail: suzhangwen@fjzzit.edu.cn

299.57 and 294.60 $\text{g}\cdot\text{m}^{-2}\cdot\text{a}^{-1}$, respectively, with interannual variations of 19.51 and 32.43 $\text{g}\cdot\text{m}^{-2}\cdot\text{a}^{-1}$, respectively. (3) The spatial distribution of soil Rh was high in the south and low in the north, mainly influenced by soil properties and leaf area index. This spatial heterogeneity reflected different responses of soil Rh to environmental changes. [Conclusion] The gradient boosting machine model demonstrates good adaptability in the modeling and prediction of large-scale soil Rh. It can effectively capture the spatiotemporal variations of soil Rh and demonstrate strong predictive ability. [Ch, 4 fig. 3 tab. 33 ref.]

Key words: soil heterotrophic respiration estimation; terrestrial ecosystem; XGBoost model; LightGBM model

土壤异养呼吸 (Rh) 是陆地生态系统碳循环的重要组成部分, 直接影响土壤碳储量与大气温室气体浓度^[1]。作为气候变化反馈机制的关键环节, 土壤异养呼吸的准确估算对于评估碳源、碳汇及全球碳循环具有重要的科学意义^[2]。然而, 由于土壤微生物分解过程的复杂性及对环境变化的敏感性, 土壤异养呼吸的估算面临较大挑战。

传统的土壤异养呼吸估算方法主要依赖于线性或非线性回归模型, 在引入更多变量时可能会引发多重共线性及自相关等问题, 从而削弱模型的稳健性和预测精度^[3-6]。近年来, 机器学习模型凭借其强大的非线性拟合能力, 在生态学领域展现出巨大潜力^[6-8]。例如, TANG 等^[6]和 HE 等^[7]采用随机森林 (RF) 模型对土壤异养呼吸进行估算, 其准确性取得了显著提升。然而, 这些研究在模型应用时未充分考虑环境因子的优化选择及模型结构的进一步优化, 限制了模型性能的进一步提升^[7-8]。相比之下, 极限梯度提升树 (XGBoost) 和轻量级梯度提升机 (LightGBM) 作为 2 种经改进的梯度提升机模型, 通过引入基于模型权重绝对值 (L1) 和平方 (L2) 惩罚项的正则化机制能有效防止过拟合, 同时采用高效的并行计算框架, 显著提高了计算效率, 在处理大规模、多维度数据集时表现出更优的性能, 从而在土壤异养呼吸估算中展现出更为突出和可靠的应用前景^[9-13]。

除估算模型的选择外, 选取高质量的解释变量也是精确预测土壤异养呼吸的基础^[14]。研究表明: 土壤温度、水分、有机质、pH 等土壤性质, 以及降水量、气温等气候因子是影响土壤异养呼吸的关键因素^[6-7]。通过整合高分辨率的多源遥感数据 [如协调世界土壤数据库 (HWSD) 土壤性质数据和 CRU TS 气候数据], 可实现土壤异养呼吸的大尺度估算和动态估算^[12]。然而, 当前针对中国陆地生态系统的土壤异养呼吸估算数据仍较为匮乏, 基于机器学习的土壤异养呼吸估算方法也有待进一步验证与优化。

鉴于此, 本研究基于全球土壤呼吸数据库 (SRDB) 和多源遥感数据, 利用 XGBoost 和 LightGBM 2 种梯度提升机模型, 对 2000—2023 年中国陆地生态系统的土壤异养呼吸进行估算^[1, 6-7, 15]。本研究旨在对比 2 种模型的性能表现, 并探讨中国陆地生态系统的土壤异养呼吸空间分布格局及其环境响应机制^[16], 在当前全球气候变化和生物多样性下降的背景下, 对指导中国碳汇项目的规划、评估和实施具有重要的现实意义。

1 数据获取与处理

1.1 土壤异养呼吸数据

本研究土壤异养呼吸数据集来源于 BOND-LAMBERTY 等^[2]编制的全球土壤呼吸数据库, TANG 等^[6]利用中国知网相关文献数据在此基础上对全球土壤呼吸数据库进行了进一步的完善, 该数据集已于 2022 年 10 月 9 日更新, 获取网址为 <https://github.com/bpbond/srdb/>。综合 TANG 等^[6]和 HE 等^[7]的全球土壤呼吸数据库数据筛选标准, 本研究仅保留符合以下要求的数据: ①测量时间至少为 1 a。②同一地点的多年测量数据被视为独立样本。③提供了明确的地理坐标和测量年份。④数据未经过实验操作干扰。⑤基于碱吸收法的测量结果被排除在外。最终保留了中国区域的 467 个土壤异养呼吸观测数据, 其中草地、城市、耕地、森林分别为 61、33、227、146 个, 湿地和其他土地利用类型因观测数据缺失未纳入统计分析。

1.2 影响因子数据

基于 HE 等^[7]的研究经验, 本研究共选取了 18 个环境因子, 涵盖气候、地形、植被、土壤生物特征

及土壤物理化学性质等多个方面。年平均气温 (MAT) 和年平均降水量 (MAP) 的数据来源于英国东英格利亚大学气候研究中心的 CRU TS v4.07 数据 (<https://crudata.uea.ac.uk/cru/data/hrg>)。该数据集凭借其精细的空间分辨率、长时间序列和全球广泛覆盖等特点,能够准确捕捉区域气候变化特征。基于此数据集定义的土壤干旱指数 (DryMonth)^[17],进一步提升了气候因子对土壤异养呼吸动态变化的响应能力。地形因子中的海拔 (Elevation) 数据来源于中国科学院计算机网络信息中心地理空间数据云平台 (<http://www.gscloud.cn>)^[18],其高精度全球地形数据为本研究提供了可靠支持。植被类型及叶面积指数数据均来源于 ERA5 再分析产品 (<https://cds.climate.copernicus.eu>)^[19],该数据具有高时空分辨率和长期连续性,能够准确反映区域植被的动态变化特征。植被类型作为研究区主要植被群落结构组成的重要指标,能有效揭示生态系统功能及地上生物量的空间分布格局;年均低植被叶面积指数与年均高植被叶面积指数分别代表灌草层与乔木层植被的覆盖度及生产力,是衡量不同植被层级光合效率和碳固定潜力的关键生态参数,对土壤异养呼吸过程具有显著的调控作用。土壤物理化学性质数据来源于协调世界土壤数据库 (HWSD v2.0, <https://www.fao.org/home/en>)^[20],该数据库整合了多种土壤理化指标,如土壤有机碳、pH、总氮密度等,为土壤异养呼吸估算提供了较为全面的栅格数据。本研究选用上述数据集作为模型拟合变量,综合考虑了土壤异养呼吸与不同环境因子影响的关系,为提高模型的预测精度奠定了坚实基础。相关数据的具体描述详见表 1。

表 1 环境因子概况
Table 1 Overview of environmental factors

项目	年平均 气温/K	年平均 降水量/m	植被 类型	含砂量/ %	粉土/ %	黏土/ %	土壤阳离子交 换量/(cmol·kg ⁻¹)	土壤pH	土壤有机碳/ (g·kg ⁻¹)
平均值	276.24	0.02	58.44	41.07	32.63	22.54	12.89	6.27	11.8
最小值	249.67	0	10.00	2.90	5.00	2.10	1.10	0.78	1.1
最大值	296.74	0.14	190.00	89.00	50.00	56.00	32.00	8.40	30.7
中位数	276.51	0.01	50.00	38.00	36.70	22.00	12.00	6.50	11.3
上四分位数	271.21	0	11.00	29.00	24.00	20.00	8.55	5.10	7.4
下四分位数	284.73	0.04	70.00	44.00	41.16	25.00	17.00	7.30	14.6

项目	土壤总氮 密度/(g·kg ⁻¹)	土壤碳 氮比	土壤含 水量/%	土壤潜在 蒸发/m	土壤真菌 细菌比	土壤线虫密度/ (条·kg ⁻¹)	海拔/m	年均低植被 叶面积指数	年均高植被 叶面积指数
平均值	14.1	11.48	0.53	-0.01	21.84	136.2	790.23	1.62	2.01
最小值	2.2	9.00	0.15	-0.02	8.07	52.7	4.00	0.20	0
最大值	38.9	18.00	0.89	0	38.24	362.5	3 780.00	3.43	4.91
中位数	14.0	11.00	0.53	-0.01	20.84	122.8	437.00	1.48	2.00
上四分位数	11.3	10.00	0.46	0	16.29	96.6	154.00	1.05	0.73
下四分位数	15.4	13.00	0.59	-0.01	28.66	16 300.5	1 329.00	2.10	2.73

说明:年平均气温、年平均降水量、土壤总氮密度、土壤碳氮比、土壤潜在蒸发、土壤真菌细菌比的原始分辨率为 0.50°,植被类型的原始分辨率为 0.05°,含砂量、粉土、黏土、土壤阳离子交换量、土壤 pH、土壤有机碳的原始分辨率为 250 m,土壤含水量的原始分辨率为 0.25°,土壤线虫密度的原始分辨率为 1 km,海拔的原始分辨率为 30 m,年均低植被叶面积指数、年均高植被叶面积指数的原始分辨率为 0.083°。

1.3 数据处理

本研究基于全球土壤呼吸数据库筛选的 467 个观测点,并根据每个样点所属的年份,使用 ArcGIS 10.8 中的“多值提取至点”提取对应表 1 所有的解释变量数据,构建中国区域的土壤异养呼吸及环境因子数据库,用于建模。为了生成年尺度的土壤异养呼吸栅格数据,使用“创建渔网”生成了 10 km×10 km 的中国区域网格数据(约 14 万个网格单元),并使用“以表格显示分区统计”,将表 1 所有的变量栅格数据提取到所有的网格单元中,进而将全球土壤呼吸数据库数据建立的最优模型结合网格数据,估算土壤异养呼吸栅格数据。

2 研究方法

2.1 梯度提升机

梯度提升模型是一种迭代的决策树算法，通过逐步训练多个弱学习器以提升模型的预测性能^[21]。极限梯度提升树 (XGBoost) 和轻量级梯度提升机 (LightGBM) 是梯度提升机中 2 种具有独特优势的模型，凭借基于直方图的分裂算法和高效处理稀疏数据的能力，显著提升了计算效率和预测准确度^[22]。因此，本研究选择 XGBoost 和 LightGBM 模型构建土壤异养呼吸的预测模型。

XGBoost 通过在预测过程中逐步添加新的回归树，每一轮迭代都针对前一轮的残差进行拟合，从而逐步减少预测值与真实值之间的误差，进而提高模型的预测精度^[23]。此外，该模型引入了正则化项以控制模型复杂度，并通过增量式训练和精细调整特征分裂点来优化变量选择与权重参数，进一步提升模型性能^[24-25]。

LightGBM 基于直方图算法 (histogram) 寻找最佳分裂点，并通过单边梯度采样算法 (GOSS) 和互斥特征绑定算法 (EFB) 来降低训练学习过程中样本数量和特征数量^[26-27]。这些创新可使 LightGBM 具备训练快速和内存占用低的优点，能有效处理高维数据和大规模数据集。

本研究采用 R 4.4.1 软件中的“xgboost”包和“lightgbm”包来实现 XGBoost 和 LightGBM 模型的构建，并利用 xgb.cv 函数、lgb.cv 函数与 rBayesianOptimization 包进行交叉验证和贝叶斯优化，以调整模型参数。

2.2 模型评估与验证

本研究采用决定系数 (R^2)、均方误差 (MSE)、均方根误差 (RMSE)、平均绝对误差 (MAE)^[28-31] 和赤池信息量准则 (AIC) 对模型的拟合结果进行评估。决定系数反映了模型对数据变异的解释能力，其值范围为 0~1，值越接近 1，表示模型的拟合效果越好；均方误差用于评估模型预测结果与实际结果之间的平均平方差，值越小，表明模型的预测误差越小；均方根误差是均方误差的平方根，进一步量化了模型预测结果与实际值之间的偏差程度；平均绝对误差用于衡量模型预测值与实际值之间的平均绝对差异，其数值越小，表示模型的预测精度越高；赤池信息量准则考虑了模型的复杂度和拟合优度，用于评估模型的相对质量，值越小，表示模型越优。XGBoost 和 LightGBM 模型的预测精度不会受到样本数据集大小的显著影响^[24, 26]，但考虑到 2 种模型在提高效率的同时可能会出现过拟合和对噪声敏感等问题，为确保模型的稳健性和预测精度，使用 80% 数据用于训练集，20% 用于测试集的划分方式，并通过 5 次随机划分生成不同训练样本集，以提高模型的泛化能力。此外，为识别对土壤异养呼吸影响最显著的环境因子，本研究基于训练好的 LightGBM 和 XGBoost 模型，分别通过 lgb.importance 和 xgb.importance 函数计算变量重要性，用以筛选关键影响因子^[5]。

为全面评估 XGBoost 和 LightGBM 在预测中国陆地生态系统土壤异养呼吸方面的性能，本研究选取了 TANG 等^[6]、HE 等^[7] 和张梅等^[30] 3 项具有代表性的研究作为外部基准。这些研究分别采用随机森林和传统统计方法估算土壤异养呼吸，覆盖广泛的地理区域。其中，TANG 等^[6] 和 HE 等^[7] 的数据以栅格图层形式在文章附件处下载，张梅等^[30] 的数据则以表格形式在文章中展示。通过对比本研究与上述研究的结果，不仅可验证所构建模型的有效性，还能揭示不同方法在表征土壤异养呼吸时空动态方面的差异。

3 结果与分析

3.1 模型拟合结果评估

表 2 显示了 5 个样本数据的拟合与预测结果。从训练集来看，XGBoost 和 LightGBM 模型的决定系数均接近 0.9，表明出良好的拟合能力。两者的误差指标 (均方误差为 $8\,602.76\sim 10\,833.86\text{ g}\cdot\text{m}^{-2}$ ，平均绝对误差为 $52.86\sim 64.54\text{ g}\cdot\text{m}^{-2}$ ，均方根误差为 $92.75\sim 104.09\text{ g}\cdot\text{m}^{-2}$) 相近，但 XGBoost 模型的赤池信息量准则值略低，表明其对训练集的拟合能力更为紧凑。然而，在测试集上，LightGBM 模型的指标多优于 XGBoost 模型，决定系数仍保持在 0.8 以上，平均绝对误差为 $80.74\sim 108.18\text{ g}\cdot\text{m}^{-2}$ ，均方误差为 $5\,467.00\sim 30\,514.14\text{ g}\cdot\text{m}^{-2}$ ，均方根误差为 $124.37\sim 167.75\text{ g}\cdot\text{m}^{-2}$ ，赤池信息量准则为 $1\,138.18\sim 1\,176.23$ 。

为全面评估变量组合对模型性能的影响，本研究以 18 个环境因子为基础，按变量数量从 1~18 分

表 2 XGBoost 和 LightGBM 模型的 5 组样本拟合结果

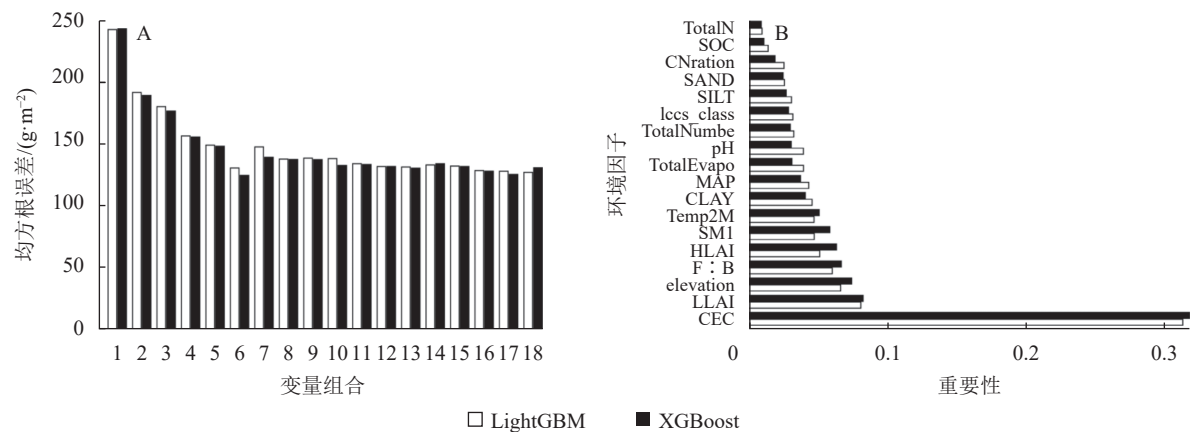
		样本1					样本2				
模拟对象	模型	均方 误差/(g·m ⁻²)	均方根 误差/(g·m ⁻²)	平均绝对 误差/(g·m ⁻²)	决定 系数	赤池信息 量准则	均方 误差/(g·m ⁻²)	均方根 误差/(g·m ⁻²)	平均绝对 误差/(g·m ⁻²)	决定 系数	赤池信息 量准则
训练集	LightGBM	10 833.86	104.09	64.54	0.90	4 483.09	10 617.39	103.04	63.96	0.90	4 471.22
	XGBoost	9 467.11	97.30	57.48	0.91	4 447.99	10 707.57	103.48	65.29	0.90	4 467.09
测试集	LightGBM	17 668.85	132.92	96.37	0.83	1 148.12	15 467.00	124.37	96.31	0.88	1 139.91
	XGBoost	17 070.74	130.66	94.23	0.84	1 149.85	15 768.83	125.57	97.12	0.87	1 142.06

		样本3					样本4				
模拟对象	模型	均方 误差/(g·m ⁻²)	均方根 误差/(g·m ⁻²)	平均绝对 误差/(g·m ⁻²)	决定 系数	赤池信息 量准则	均方 误差/(g·m ⁻²)	均方根 误差/(g·m ⁻²)	平均绝对 误差/(g·m ⁻²)	决定 系数	赤池信息 量准则
训练集	LightGBM	8 602.76	92.75	55.33	0.91	4 405.46	9 321.87	96.55	57.19	0.92	4 436.45
	XGBoost	9 321.09	96.55	59.88	0.91	4 423.38	8 715.14	93.35	53.65	0.92	4 418.05
测试集	LightGBM	28 140.85	167.75	108.18	0.82	1 170.28	16 644.90	129.02	98.81	0.85	1 150.71
	XGBoost	30 514.14	174.68	110.16	0.80	1 176.23	17 520.18	132.36	100.95	0.85	1 156.04

		样本5				
模拟对象	模型	均方 误差/(g·m ⁻²)	均方根 误差/(g·m ⁻²)	平均绝对 误差/(g·m ⁻²)	决定 系数	赤池信息 量准则
训练集	LightGBM	8 737.64	93.48	52.86	0.92	4 428.41
	XGBoost	8 841.29	94.03	53.37	0.92	4 429.43
测试集	LightGBM	15 611.71	124.95	80.74	0.86	1 139.56
	XGBoost	15 741.19	125.46	81.21	0.86	1 138.18

组，自由组合各组变量并分别建模，选取每组中预测误差最小的模型进行性能比较^[30]。结果表明：当变量数量为 6 时，模型均方根误差最低，预测性能最佳。基于此，进行结合变量重要性排序，最终筛选出阳离子交换量 (CEC)、年均低植被叶面积指数 (LLAI)、年均高植被叶面积指数 (HLAI)、海拔 (elevation)、土壤真菌细菌比 (F:B) 和年平均降水量 (MAP) 6 个关键影响因子作为最优模型拟合的数据 (图 1)，模型预测精度也得到了进一步提升，能更精准地反映土壤异养呼吸的变化趋势。

基于筛选出的 6 个影响因子进行模型拟合，建立了最优模型 (表 3)。在训练集上，XGBoost 和



A. 筛选关键影响因子组合。横轴数字表示每组变量中包含的变量个数。B. 根据重要性对影响因子排序。CEC. 土壤阳离子交换量; LLA1. 年均低植被叶面积指数; elevation. 海拔; F : B. 土壤真菌细菌比; HLAI. 年均高植被叶面积指数; SM1. 土壤含水量; Temp2M. 年平均气温; CLAY. 粉土; MAP. 年平均降水量; TotalEvapo. 土壤潜在蒸发; pH. 土壤 pH; TotalNumbe. 土壤线虫密度; lcss_class. 植被类型; SILT. 黏土; SAND. 含砂量; CNration. 土壤碳氮比; SOC. 土壤有机碳; TotalN. 土壤总氮密度。

图 1 关键影响因子流程图
Figure 1 Flowchart for key influential factors

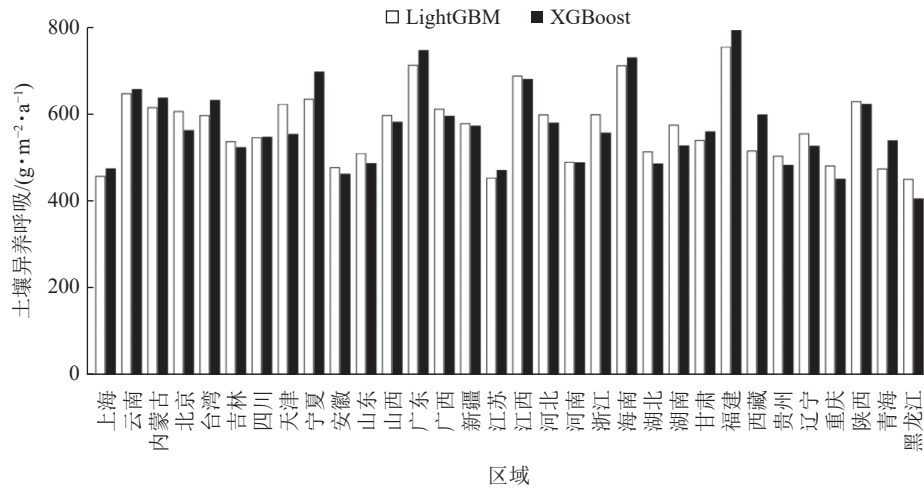
LightGBM 模型的决定系数均为 0.88，表现良好，但 XGBoost 模型略优于 LightGBM 模型，平均绝对误差为 72.09 $\text{g}\cdot\text{m}^{-2}$ ，均方误差为 13 093.28 $\text{g}\cdot\text{m}^{-2}$ 。相比之下，LightGBM 模型在测试集中的表现更为出色，其决定系数达 0.91，平均绝对误差为 68.64 $\text{g}\cdot\text{m}^{-2}$ ，均方误差为 8 388.33 $\text{g}\cdot\text{m}^{-2}$ ，均方根误差为 91.59 $\text{g}\cdot\text{m}^{-2}$ ，赤池信息量准则为 1 087.35。综合来看，在土壤异养呼吸估算中，LightGBM 模型测试集的预测效果略优于 XGBoost 模型，而 XGBoost 模型在训练集中的泛化能力方面表现更佳。

表 3 使用关键影响因子组合进行建模后的结果

Table 3 Results from modeling with key influential factor combinations						
模拟对象	模型	均方误差	均方根误差	平均绝对误差	决定系数	赤池信息量准则
训练集	LightGBM	13 241.05	115.07	73.06	0.88	4 568.84
	XGBoost	13 093.28	114.43	72.09	0.88	4 564.34
测试集	LightGBM	8 388.33	91.59	68.64	0.91	1 087.35
	XGBoost	8 768.15	93.64	70.10	0.91	1 091.89

3.2 模型预测结果

土壤异养呼吸的预测结果显示 (图 2)：中国不同区域土壤异养呼吸水平存在明显差异。其中，福建、海南、广东等表现出较高的土壤异养呼吸水平，反映出这些区域土壤有机质分解活跃，可能与温暖湿润的气候条件和丰富的植被覆盖有关。西藏、黑龙江等寒冷或高海拔地区，土壤异养呼吸水平相对较低，表明土壤微生物活动受限，碳释放速率较慢。此外，内蒙古、新疆、宁夏等干旱半干旱地区也显示出相对较高的土壤异养呼吸水平，可能与局部草原生态系统和土壤水分条件相关。

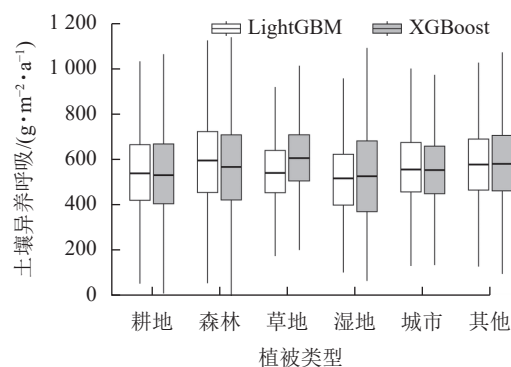


因缺乏有效土壤样本，未对香港、澳门的土壤异养呼吸进行预测。

图 2 中国土壤异养呼吸预测结果分布
Figure 2 Spatial distribution of soil heterotrophic respiration

为了探讨不同类型土壤异养呼吸的情况，本研究对 2000—2023 年 2 种模型预测的不同土地利用类型 (耕地、森林、草地、湿地、城市及其他) 的土壤异养呼吸结果做了进一步分析 (图 3)。结果表明：2 种模型在不同植被类型下的平均预测值总体相近，但仍存在一定差异。其中，草地的平均预测值最高，XGBoost 和 LightGBM 模型分别为 607.39 和 548.78 $\text{g}\cdot\text{m}^{-2}\cdot\text{a}^{-1}$ ；湿地的平均预测值最低，XGBoost 与 LightGBM 模型分别为 528.47 和 514.18 $\text{g}\cdot\text{m}^{-2}\cdot\text{a}^{-1}$ 。此外，2 种模型的最大预测值均出现在草地和森林，最小预测值则集中在湿地和城市地区。总体来看，XGBoost 模型在草地区域的预测值略高于 LightGBM 模型，而在湿地和城市区域则相对较低。

本研究通过对比分析揭示了 2000—2023 年全国土壤异养呼吸的年际变化趋势，比较了 XGBoost、LightGBM、随机森林 (TANG 等^[6] 和 HE 等^[7]) 以及回归模型 (张梅等^[30]) 的估算结果 (图 4)。TANG 等^[6] 的随机森林模型结果表明：土壤异养呼吸在 2008 和 2016 年分别达到峰值 (348.80 和 356.71 $\text{g}\cdot\text{m}^{-2}\cdot\text{a}^{-1}$)，显示出对环境变化的较强敏感性。相比之下，HE 等^[7] 的随机森林模型结果相对平稳，年际变化幅度仅



箱子上框线为上四分位数, 中间线为平均数, 下框线为下四分位数。

图3 不同植被类型下土壤异养呼吸预测结果

Figure 3 Statistics of soil heterotrophic respiration under different land use types

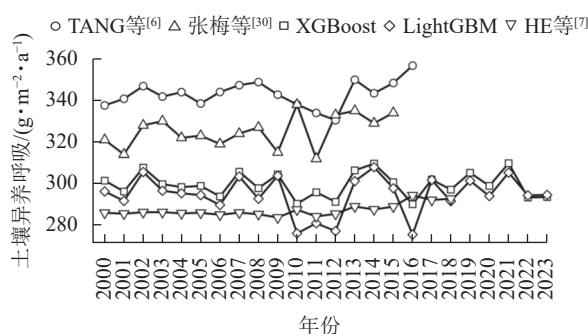


图4 不同模型预测的土壤异养呼吸年平均值对比图

Figure 4 Comparison of annual average soil heterotrophic respiration values predicted by different models

为 $10.68 \text{ g} \cdot \text{m}^{-2} \cdot \text{a}^{-1}$, 表明其对误差的控制较好, 但可能未能充分捕捉极端年份的波动。张梅等^[30] 基于回归模型的估算结果则显示: 在 2009—2012 年, 土壤异养呼吸呈现明显波动, 峰值出现在 2010 年, 可能与该时期的气候异常有关。本研究中, XGBoost 和 LightGBM 模型均能够较好地捕捉土壤异养呼吸的年际变化趋势, 尤其在 2010—2012 年和 2016 年表现出明显的波动, 最低值分别为 290.16 和 $275.54 \text{ g} \cdot \text{m}^{-2} \cdot \text{a}^{-1}$, 与 TANG 等^[6] 估算的土壤异养呼吸的变化趋势相一致。这表明 XGBoost 和 LightGBM 模型能够较好反映极端年份的变化, 且在 2000—2023 年, XGBoost 和 LightGBM 模型的土壤异养呼吸年均值分别为 299.57 和 $294.60 \text{ g} \cdot \text{m}^{-2} \cdot \text{a}^{-1}$, 年际变化幅度分别为 19.51 和 $32.43 \text{ g} \cdot \text{m}^{-2} \cdot \text{a}^{-1}$, 显示出两者在捕捉年际变化幅度上的差异。

4 讨论

经过筛选, 最终确定模型拟合的关键变量为土壤阳离子交换量、年均低植被叶面积指数、年均高植被叶面积指数、海拔、土壤真菌细菌比以及年平均降水量。尽管湿地及其他植被类型的土壤异养呼吸观测数据存在缺失, 但考虑到模型未直接引入植被类型变量, 同时本研究采用了 10 km 高分辨率网格进行模拟, 这 6 个关键变量在一定程度上亦能反映湿地及其他生态系统的特性, 从而提升土壤异养呼吸估算的精度。在这些变量中, 土壤阳离子交换量的重要性最高, 这可能是由于其在调控土壤养分动态及促进有机质分解过程中对微生物活性产生了决定性影响, 从而显著影响了土壤异养呼吸。叶面积指数则紧随其后, 可能通过增加有机碳输入及调节土壤温湿度与微生物活性等, 间接驱动土壤异养呼吸的变化^[7]。此外, 海拔、土壤真菌细菌比以及年平均降水量虽然贡献度相对较低, 但分别代表了气候条件、微生物群落结构与水分状态, 对土壤异养呼吸的空间异质性和动态变化仍发挥着不可忽视的调节作用^[3]。

相比于张梅等^[30] 采用的传统统计方法, XGBoost 和 LightGBM 2 种梯度提升机模型更精准地刻画了非线性关系及高维变量间的复杂关系, 显著提升了预测精度。相较于 TANG 等^[6] 和 HE 等^[7] 所使用的机器学习方法, 本研究进一步结合贝叶斯优化与交叉验证, 在模型评估中综合考量平均绝对误差、赤池信息量准则等多项指标, 以确保评估结果的全面性与稳健性。此外, 本研究通过优化模型的关键变量, 并结合 10 km 高分辨率网格数据, 提升了对区域异质性的解析能力, 进一步增强了模型的泛化能力。2 种模型通过梯度提升策略和正则化机制显著提升了预测精度, 其中 LightGBM 模型凭借其独特的直方图算法在测试集上的表现相对更为优越, 而 XGBoost 模型则在训练集上展现出更强的泛化能力。考虑到土壤异养呼吸数据的观测误差, 高达 90% 的解释率可能部分来源于对训练数据噪声的过度学习, 存在潜在的过拟合风险。尽管贝叶斯优化和交叉验证在一定程度上缓解了这一问题, 但数据的不确定性依然难以完全消除。因此, 本研究进一步对网格预测值进行了不确定性分析, 结果显示: XGBoost 模型在误差控制方面表现出更为优越的特性, 这可能与更强的抗噪能力和鲁棒性相关。

基于 XGBoost 和 LightGBM 模型预测的中国陆地生态系统的土壤异养呼吸结果, 进一步分析不同植

被类型的土壤异养呼吸估算结果后发现：土壤异养呼吸的最大值主要出现在草地和森林区域。这 2 类植被在南方低纬度地区(如云南、广西和福建)分布广泛，与 2 种模型预测的土壤异养呼吸空间分布模式高度一致。进一步的空间分析表明：2000—2023 年，中国土壤异养呼吸较高的区域集中在东南沿海的低纬度地区，而北方高纬度地区土壤异养呼吸较低，这一趋势与张梅等^[30]的研究结果相吻合。土壤异养呼吸南北差异的形成主要受气候、土壤性质及植被类型的共同影响。在高纬度地区，寒冷气候抑制微生物活性，限制有机质分解，导致土壤异养呼吸较低^[32]。在南方，较高的气温、充足的降水以及较高的植被覆盖共同促进了土壤有机碳的分解和土壤异养呼吸^[33]。

目前，全球土壤呼吸数据库中关于中国区域的土壤异养呼吸观测数据不足，未能实现对研究区的全覆盖，这可能会削弱模型外推能力并增加过拟合风险；同时世界土壤数据库数据缺乏年尺度的土壤质地与理化性质的栅格信息，限制了对土壤异养呼吸复杂动态特征的精确刻画。此外，当环境因子数值超出观测点分布范围时，模型外推的不确定性将显著增加。针对上述问题，未来研究应在扩充观测数据空间覆盖的同时，系统引入更多具备时空动态特征的环境因子，并通过生态系统建模集成策略，优化模型在未知区域的准确性与适用性。

5 结论

XGBoost 和 LightGBM 模型在拟合精度方面均表现优异(训练集决定系数为 0.88，测试集决定系数为 0.91)，且在误差指标上也取得了较好的结果，表明梯度提升机模型在土壤异养呼吸估算中展现了较高的稳定性。通过深入分析土壤异养呼吸的影响因素，本研究发现：土壤阳离子交换量、年均低植被叶面积指数、年均高植被叶面积指数、海拔、土壤真菌细菌比以及年平均降水量是决定土壤异养呼吸变化的主要因素。空间分布分析揭示：中国土壤异养呼吸量较高的区域集中在东南沿海的低纬度地区，北方高纬度地区的土壤异养呼吸相对较低。这一发现为土壤异养呼吸预测和相关领域的决策支持提供了坚实的基础。未来研究可以进一步探讨不同生态系统和气候条件下土壤异养呼吸的变化规律，优化模型的普适性与准确性，以更好地应对气候变化。

6 参考文献

- [1] BOND-LAMBERTY B, BAILEY V L, CHEN Min, *et al.* Globally rising soil heterotrophic respiration over recent decades [J]. *Nature*, 2018, **560**(7716): 80–83.
- [2] BOND-LAMBERTY B, THOMSON A. A global database of soil respiration data [J]. *Biogeosciences*, 2010, **7**(6): 1915–1926.
- [3] BOND-LAMBERTY B, WANG Chuankuan, GOWER S T. A global relationship between the heterotrophic and autotrophic components of soil respiration? [J]. *Global Change Biology*, 2004, **10**(10): 1756–1766.
- [4] 范志平, 王红, 邓东周, 等. 土壤异养呼吸的测定及其温度敏感性影响因子[J]. *生态学杂志*, 2008, **27**(7): 1221–1226.
FAN Zhiping, WANG Hong, DENG Dongzhou, *et al.* Measurement methods of soil heterotrophic respiration and key factors affecting the temperature sensitivity of the soil heterotrophic respiration [J]. *Chinese Journal of Ecology*, 2008, **27**(7): 1221–1226.
- [5] LIU Jiangnan, HU Junguo, LIU Haoqi, *et al.* Global soil respiration estimation based on ecological big data and machine learning model[J/OL]. *Scientific Reports*, 2024, **14**: 13231[2024-11-25]. DOI: [10.1038/s41598-024-64235-w](https://doi.org/10.1038/s41598-024-64235-w).
- [6] TANG Xiaolu, DU Jie, SHI Yuehong, *et al.* Global patterns of soil heterotrophic respiration—a meta-analysis of available dataset[J/OL]. *CATENA*, 2020, **191**: 104574[2024-11-25]. DOI: [10.1016/j.catena.2020.104574](https://doi.org/10.1016/j.catena.2020.104574).
- [7] HE Yue, DING Jinzhi, DORJI T, *et al.* Observation-based global soil heterotrophic respiration indicates underestimated turnover and sequestration of soil carbon by terrestrial ecosystem models [J]. *Global Change Biology*, 2022, **28**(18): 5547–5559.
- [8] YAO Yitong, CIAIS P, VIOVY N, *et al.* A data-driven global soil heterotrophic respiration dataset and the drivers of its inter-annual variability[J/OL]. *Global Biogeochemical Cycles*, 2021, **35**(8): e2020GB006918[2024-11-25]. DOI: [10.1029/2020GB006918](https://doi.org/10.1029/2020GB006918).
- [9] CHEN Lin, XING Minfeng, HE Binbin, *et al.* Estimating soil moisture over winter wheat fields during growing season using

- machine-learning methods [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, **14**: 3706–3718.
- [10] ZHANG Linqi, LIU Yi, REN Liliang, *et al.* Analysis of flash droughts in China using machine learning [J]. *Hydrology and Earth System Sciences*, 2022, **26**(12): 3241–3261.
- [11] 付平凡, 杨晓静, 苏志诚, 等. 基于集成学习的土壤含水量预测研究——以辽西地区为例[J]. *土壤*, 2023, **55**(3): 671–681.
- FU Pingfan, YANG Xiaojing, SU Zhicheng, *et al.* Prediction of soil moisture content based on ensemble learning: a case study of western Liaoning Province [J]. *Soils*, 2023, **55**(3): 671–681.
- [12] 李志生, 梁锡冠, 金字凯, 等. 基于树模型的北京市 PM_{2.5} 预测效果对比分析[J]. *环境工程*, 2021, **39**(6): 106–113.
- LI Zhisheng, LIANG Xiguan, JIN Yukai, *et al.* A comparative study on edictive effect of PM_{2.5} in Beijing based on tree models [J]. *Environmental Engineering*, 2021, **39**(6): 106–113.
- [13] 朱珈莹, 安俊琳, 冯悦政, 等. 基于轻量级梯度提升机的南京大气臭氧浓度预测[J]. *环境科学*, 2023, **44**(7): 3685–3694.
- ZHU Jiaying, AN Junlin, FENG Yuezheng, *et al.* Atmospheric ozone concentration prediction in Nanjing based on LightGBM [J]. *Environmental Science*, 2023, **44**(7): 3685–3694.
- [14] HASHIMOTO S, CARVALHAIS N, ITO A, *et al.* Global spatiotemporal distribution of soil respiration modeled using a global database [J]. *Biogeosciences*, 2015, **12**(13): 4121–4132.
- [15] LU Haibo, LI Shihua, MA Minna, *et al.* Comparing machine learning-derived global estimates of soil respiration and its components with those from terrestrial ecosystem models[J/OL]. *Environmental Research Letters*, 2021, **16**(5): 054048[2024-11-25]. DOI: [10.1088/1748-9326/abf526](https://doi.org/10.1088/1748-9326/abf526).
- [16] 展小云, 于贵瑞, 郑泽梅, 等. 中国区域陆地生态系统土壤呼吸碳排放及其空间格局——基于通量观测的地学统计评估[J]. *地理科学进展*, 2012, **31**(1): 97–108.
- ZHAN Xiaoyun, YU Guirui, ZHENG Zemei, *et al.* Carbon emission and spatial pattern of soil respiration of terrestrial ecosystems in China: based on geostatistic estimation of flux measurement [J]. *Progress in Geography*, 2012, **31**(1): 97–108.
- [17] HARRIS I, JONES P D, OSBORN T J, *et al.* Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 Dataset [J]. *International Journal of Climatology*, 2014, **34**(3): 623–642.
- [18] 焦鑫宇, 龙梅, 刘志雄. 历史地理信息系统视角下野生蕙兰时空分布及其影响因素[J]. *浙江农林大学学报*, 2023, **40**(6): 1261–1272.
- JIAO Xinyu, LONG Mei, LIU Zhixiong. Spatiotemporal distribution and influencing factors of wild *Cymbidium faberi* from the perspective of historical geographic information system [J]. *Journal of Zhejiang A&F University*, 2023, **40**(6): 1261–1272.
- [19] BUONTEMPO C, BURGESS S N, DEE D, *et al.* The Copernicus climate change service: climate science in action [J]. *Bulletin of the American Meteorological Society*, 2022, **103**(12): E2669–E2687.
- [20] NACHTERGAELE F, van VELTHUIZEN H, VERELST L, *et al.* *Harmonized World Soil Database Version 2.0*[M]. Rome and Laxenburg: FAO, 2023.
- [21] 黄昇. 基于多源数据与人工智能算法的湖南省森林地上生物量估算[D]. 长沙: 中南林业科技大学, 2023.
- HUANG Sheng. *Above-ground Biomass Estimation in Hunan Province Based on Multi-source Data and Artificial Intelligence Algorithms*[D]. Changsha: Central South University of Forestry & Technology, 2023.
- [22] WANG Hao, SHAO Wei, HU Yunfeng, *et al.* Assessment of six machine learning methods for predicting gross primary productivity in grassland[J/OL]. *Remote Sensing*, 2023, **15**(14): 3475[2024-11-25]. DOI: [10.3390/rs15143475](https://doi.org/10.3390/rs15143475).
- [23] GUO Xu, GUI Xiaofan, XIONG Hanxiang, *et al.* Critical role of climate factors for groundwater potential mapping in arid regions: insights from random forest, XGBoost, and LightGBM algorithms[J/OL]. *Journal of Hydrology*, 2023, **621**: 129599[2024-11-25]. DOI: [10.1016/j.jhydrol.2023.129599](https://doi.org/10.1016/j.jhydrol.2023.129599).
- [24] 陈涵, 张超, 余树全. 基于 Stacking 模型集成算法的莲都区南方红豆杉潜在分布区[J]. *浙江农林大学学报*, 2019, **36**(3): 494–500.
- CHEN Han, ZHANG Chao, YU Shuquan. Potential distribution area of *Taxus chinensis* var. *mairei* in Liandu District based on a Stacking algorithm [J]. *Journal of Zhejiang A&F University*, 2019, **36**(3): 494–500.

- [25] 晏红波, 梁雨豪, 卢献健, 等. 基于 XGBoost 融合多维度时空数据的干旱遥感建模及应用研究[J]. *地球信息科学学报*, 2024, **26**(6): 1531–1546.
YAN Hongbo, LIANG Yuhao, LU Xianjian, *et al.* Remote sensing modeling and applications in drought monitoring based on XGBoost and fusion of multi-dimensional spatiotemporal data [J]. *Journal of Geo-Information Science*, 2024, **26**(6): 1531–1546.
- [26] KE Guolin, MENG Qi, FINLEY T, *et al.* Lightgbm: a highly efficient gradient boosting decision tree[C]//von LUXBURG U, GUYON I. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 2017: 3149–3157.
- [27] 付佳龙. 基于 LightGBM 模型的径流预报研究[D]. 武汉: 华中科技大学, 2022.
FU Jialong. *Research on Runoff Forecast Based on LightGBM Model*[D]. Wuhan: Huazhong University of Science and Technology, 2022.
- [28] 咸阳, 宋江辉, 王金刚, 等. 基于环境变量筛选与机器学习的土壤养分含量空间插值研究[J]. *农业机械学报*, 2024, **55**(10): 379–391.
XIAN Yang, SONG Jianghui, WANG Jingang, *et al.* Spatial interpolation of soil nutrients content based on environmental variables screening and machine learning [J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2024, **55**(10): 379–391.
- [29] LU Ruhua, ZHANG Pei, FU Zhaopeng, *et al.* Improving the spatial and temporal estimation of ecosystem respiration using multi-source data and machine learning methods in a rainfed winter wheat cropland[J/OL]. *Science of the Total Environment*, 2023, **871**: 161967[2024-11-25]. DOI: 10.1016/j.scitotenv.2023.161967.
- [30] 张梅, 黄贤金, 揣小伟, 等. 中国净生态系统生产力空间分布及变化趋势研究[J]. *地理与地理信息科学*, 2020, **36**(2): 69–74.
ZHANG Mei, HUANG Xianjin, CHUAI Xiaowei, *et al.* Spatial distribution and changing trends of net ecosystem productivity in China [J]. *Geography and Geo-Information Science*, 2020, **36**(2): 69–74.
- [31] 谢薇, 陈书涛, 胡正华. 中国陆地生态系统土壤异养呼吸变异的影响因素[J]. *环境科学*, 2014, **35**(1): 334–340.
XIE Wei, CHEN Shutao, HU Zhenghua. Factors influencing the variability in soil heterotrophic respiration from terrestrial ecosystem in China [J]. *Environmental Science*, 2014, **35**(1): 334–340.
- [32] HAN Guangxuan, ZHOU Guangsheng, XU Zhenzhu, *et al.* Biotic and abiotic factors controlling the spatial and temporal variation of soil respiration in an agricultural ecosystem [J]. *Soil Biology and Biochemistry*, 2007, **39**(2): 418–425.
- [33] PAN Qilong, HARROU F, SUN Ying. A comparison of machine learning methods for ozone pollution prediction [J/OL]. *Journal of Big Data*, 2023, **10**(1): 63[2024-11-25]. DOI:10.1186/s40537-023-00748-x.