

基于机器学习与 SHAP 分析的中国寒温带优势树种分类

董 恒¹ 姜宜辰¹ 熊 涛¹ 袁艳斌¹ 玉 山² 都瓦拉³

(1. 武汉理工大学关键非金属矿产资源绿色利用教育部重点实验室, 武汉 430070;

2. 内蒙古师范大学地理科学学院, 呼和浩特 010022;

3. 中国农业科学院草原研究所, 呼和浩特 010010)

摘要: 获取较高精度优势树种分布数据是解析北方森林响应气候变化机制的关键。然而, 相比于北美、欧亚北方森林(特别是受气候变化与人为活动影响显著的南缘地带)的树种分布数据仍显匮乏。为此, 本研究选取位于欧亚北方森林南缘的中国寒温带森林为研究区, 整合光谱波段、纹理特征、地形因子及生物气候变量等多源数据, 应用 XGBoost、Random Forest、MLP 和 Conv1D 4 种机器学习模型, 对白桦(*Betula platyphylla*)、落叶松(*Larix gmelinii*) 和山杨(*Populus davidiana*) 3 种优势树种进行精细识别与绘制。同时, 引入 SHAP 分析从全局和局部视角阐释模型决策机制及特征交互效应。结果表明: XGBoost 模型性能最优, 总体分类精度可达 80.1%, 分类错误率最低, 且山杨在 3 个树种中可分性最佳; SHAP 可解释性分析发现, 降水(Bio14)及与温度变化(Bio7、Bio4、Bio3)相关的变量在模型判别中贡献最显著; 山杨倾向于温度季节性(Bio4)、日较差(Bio2)较大的环境, 在极端气候下具有出色的适应能力, 体现了非线性、物种特异性的气候适应性。研究证实, 耦合机器学习、环境因子与模型可解释性分析, 可有效解决中国寒温带森林优势树种分类精度不足的问题, 增强对气候-树种关系的生态解释能力, 为制定针对性森林管理策略提供现实依据。

关键词: 优势树种分类; 中国寒温带森林; 机器学习; SHAP 分析

中图分类号: K90

文献标识码: A

文章编号: 1000-1298(2026)15-0267-10

OSID:



Dominant Tree Species Classification in Cold-temperate Forests of China Based on Machine Learning and SHAP Analysis

Dong Heng¹ Jiang Yichen¹ Xiong Tao¹ Yuan Yanbin¹ Yu Shan² Du Wala³

(1. Key Laboratory of Green Utilization of Critical Non-metallic Mineral Resources, Ministry of Education, Wuhan University of Technology, Wuhan 430070, China

2. College of Geographical Science, Inner Mongolia Normal University, Hohhot 010022, China

3. Institute of Grassland Research, Chinese Academy of Agricultural Sciences, Hohhot 010010, China)

Abstract: Accurate mapping of dominant tree species distributions is crucial for understanding the response of northern forests to climate change. Compared with North America, species distribution data in Eurasian northern forests, particularly in the southern margin influenced by climate change and human activities, remain scarce. Cold-temperate forests at the southern margin of the Eurasian boreal zone in China were selected as the study area. Multi-source data, including spectral bands, texture features, topographic factors, and bioclimatic variables, were integrated, and four machine learning models, XGBoost, random forest, MLP, and Conv1D were applied to accurately map three dominant species: *Betula platyphylla*, *Larix gmelinii*, and *Populus davidiana*. SHAP analysis was used to interpret model decision mechanisms and feature interactions at both global and local levels. The results indicated that the XGBoost model achieved the highest overall accuracy (80.1%) with the lowest classification error, and *Populus davidiana* was the most distinguishable among the three species; SHAP interpretability analysis

收稿日期: 2026-02-04 修回日期: 2026-03-10

基金项目: 国家自然科学基金项目(42471445、52079101)、内蒙古自治区科技计划项目(2025YFDZ0122)和武汉理工大学自主创新研究基金项目(104972025RSCbs0029)

作者简介: 董恒(1986—),男,副教授,博士,主要从事生态环境遥感研究,E-mail: simondong@whut.edu.cn

通信作者: 熊涛(1990—),男,讲师,博士,主要从事生态学研究,E-mail: xiongt830@pku.edu.cn

revealed that precipitation (Bio14) and temperature-related variables (Bio7, Bio4, Bio3) contributed most to the model discrimination; *Populus davidiana* occurred preferentially in environments with high temperature seasonality (Bio4) and diurnal range (Bio2), exhibiting strong adaptation under extreme climates, reflecting nonlinear, species-specific responses to climate. The research results demonstrated that integrating machine learning, environmental factors, and model interpretability analysis could effectively address the limited accuracy in classifying dominant tree species in China's cold-temperate forests, enhance ecological understanding of climate-species relationships, and provide practical guidance for developing targeted forest management strategies.

Key words: dominant tree species classification; cold-temperate forests in China; machine learning; SHAP analysis

0 引言

北方森林作为全球最大的森林生态系统之一,涵盖针叶林、阔叶林和混交林等类型,在生物多样性保护及全球碳循环等方面发挥着关键作用。优势树种的空间分布蕴含精确且全面的地理信息^[1],能够辅助森林火灾风险防控、生物多样性保护实践^[2]等工作。在全球变暖的大背景下,寒温带针叶林正表现出向高纬度迁移的趋势,部分位于南缘的森林可能面临长期退化甚至消失的风险^[3]。因此,及时获取寒温带优势树种的空间分布信息,对于评估其潜在迁移趋势、制定适应性森林管理策略具有重要意义。

北美已构建高精度北方森林优势树种分布数据集^[4],但欧亚地区尤其是东亚北方森林仍缺乏精确的优势树种分布信息,限制对整个北方森林优势树种格局及气候变化影响的系统认知。当前,亟需解决以下两个关键问题:

(1)构建适用于欧亚北方森林的分类模型。东亚地区广泛分布着落叶针叶林等特有群落,限制北美模型直接迁移的有效性。因此,需综合数据和模型,全面设计具有区域适用性的高精度分类模型。传统森林资源调查手段更新周期长,空间覆盖不连续。近年来,Landsat 和 Sentinel-2 等中高分辨率开放卫星数据凭借时效性强、覆盖范围广的优势,显著提升遥感技术在树种识别中的应用价值^[5]。在铁岭市优势树种分类的研究中,应用了 Sentinel-2、纹理等 120 个特征,分层设计 6 种方案,最高精度达 84.7%^[6],但高维数据易造成信息冗余,需要进行特征筛选。除数据外,识别技术也至关重要,机器学习模型在处理多源数据、捕捉复杂非线性关系方面表现出色,被广泛用于优势树种分类的研究,受自然和人为活动的影响,模型在各个区域的表现性、适用性存在差异。Lim 等^[7]整合多地训练数据,采用的随机森林模型在韩国高城郡分类准确率达 80%,并用于绘制朝鲜高城郡优势树种预测分布图。然而,这些模型往往针对特定区域或树种构建,泛化能力有

限,尤其在对气候变化敏感的欧亚北方森林南缘地区,相关研究仍匮乏。基于以上问题,亟需针对欧亚北方森林南缘构建最优模型进行树种分类。

(2)明确影响优势树种分布的环境因素。特别是在受气候变暖等因素影响剧烈的东亚北方森林南缘,其生态响应机制具有显著的指示意义。在对中国东北地区未来优势树种的预测研究中,有学者明确指出气候变化是该区域优势树种落叶松、白桦、红皮云杉漂移的主要驱动因素,但未进一步探讨树种的气候适应性^[8]。并且,XGBoost、神经网络等模型常存在“黑箱”问题,需结合可解释性方法量化特征贡献、分析内部决策机制^[9]。近年来,SHAP (Shapley additive explanations) 等模型解释方法逐渐引入,有助于理解模型判别逻辑、评估变量相互作用^[10]。将其与树种研究相结合,不仅可以提升模型的解释能力,还能深化对变量间复杂交互与树种生态机制的理解。

为解决上述 2 个关键问题,本研究选择位于东亚北方森林南部、广泛分布着落叶针叶林的中国寒温带森林为研究区域。综合利用光谱波段、纹理特征、地形因子及生物气候变量等多源数据,构建一种区域尺度的树种分类框架,反演白桦、落叶松和山杨 3 种典型优势树种的空间分布,重点揭示优势树种与气候环境之间的具体关系。通过比较多种集成机器学习方法的分类性能,借助 SHAP 分析探讨各变量的贡献特征和生态意义,提高遥感树种识别的准确性和可解释性,全面了解树种的生长习性,以期为整个欧亚北方森林优势树种的识别提供数据支持和方法参考。

1 研究区域和数据

1.1 研究区

中国寒温带森林(图 1)分布在黑龙江大兴安岭地区及内蒙古东北部^[11],以山地为主,地势西南高东北低,属于寒温带大陆性季风气候,年均温低于 0℃,总降水量约 500 mm^[12],植被以针叶林为主,大

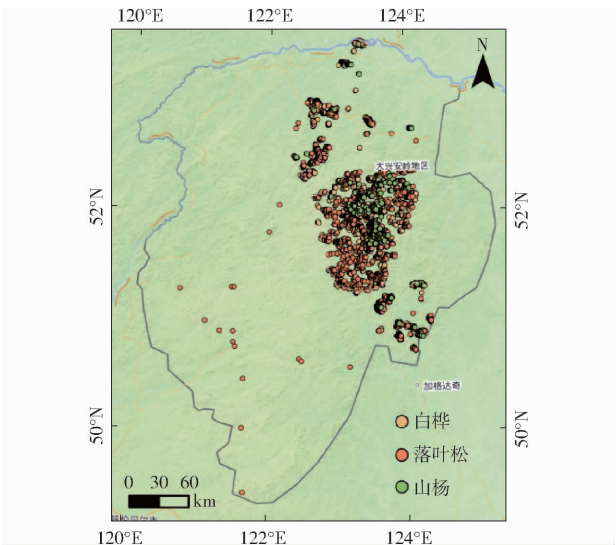


图 1 研究区及样本点分布

Fig. 1 Study area and distribution of sample points

致分为落叶针叶林和常绿针叶林两类,属于典型的欧亚北方森林,对全球气候变化敏感,是优势树种格局发生转变的关键区域。

1.2 数据

1.2.1 卫星数据

哨兵 1 号卫星 (Sentinel) 是欧洲航天局 (欧空局) 根据哥白尼方案发射的一个合成孔径雷达卫星系统,旨在提供各种天气和光照条件下的高分辨率地球观测数据^[13]。Sentinel-1 配备一个 C 波段 SAR 系统,可在复杂的天气条件下提供高质量的地表观测数据^[14]。本研究选取 2015 年 Sentinel-1 影像,极化方式为 VV + VH,并使用欧空局研发的 SNAP 6.0 进行轨道校正、噪声去除、辐射定标、滤波和地形校正。

Sentinel-2 数据广泛应用于植被、土壤和海洋监测以及分类等任务。根据样本数据获取时间,选择云量较少、质量较好且处于植被生长期的影像。本研究选取 2016—2018 年 6—8 月 L2A 影像,并选择与植被信息高度相关的 b2 ~ b12 共 11 个波段参与分析,其空间分辨率有 10 m、20 m 和 60 m^[15]。基于灰度共生矩阵 (GLCM) 提取每个波段的 8 种纹理特征:对比度、相关性、相异性、信息熵、同质性、均值、二阶矩、方差。同时,利用上述波段提取 NDVI、NDVI_{re}、mNDVI、mNDVI_{re}、RVI、EVI、CI_{re}、CI_{green}、SAVI 9 个植被指数,如表 1 所示。

1.2.2 气候与环境数据

除卫星数据外,本研究还综合利用多源地理信息数据以支持研究目标。地形数据采用分辨率为 30 m 的 ASTER 全球数字高程模型 (GDEM),并在此基础上提取坡度和坡向信息。气候数据来源于国家青藏高原科学数据中心,获取空间分辨率约为 1 km 的 23

个生物气候变量^[16] (表 2)。此外,土壤类型数据来源于国家地球系统科学数据中心。以上数据共同用于分析气候、土壤条件对优势树种分布及生态适宜性的影响 (表 1)。为实现多源数据的空间对齐,本文将上述数据 (包括卫星遥感影像) 重采样至 30 m,其中连续型数据采用双线性插值法,而非连续型数据 (土壤) 则采用最近邻插值法进行重采样 (表 3)。

表 1 原始变量及描述

Tab. 1 Original variables and descriptions

原始变量	描述
后向散射系数	S1_VV、S1_VH
光谱波段	b2、b3、b4、b5、b6、b7、b8、b8a、b9、b11、b12
植被指数	CI _{green} 、CI _{re} 、EVI、mNDVI、mNDVI _{re} 、NDVI、NDVI _{re} 、RVI、SAVI
纹理特征	b2_Con、b2_Cor、b2_Dis、b2_Ent、b2_Hom、b2_Mean、b2_SeM、b2_Var、…、b12_Con、b12_Cor、b12_Dis、b12_Ent、b12_Hom、b12_Mean、b12_SeM、b12_Var (共 88 个)
生物气候变量	Bio1、Bio2、Bio3、Bio4、Bio5、Bio6、Bio7、Bio8、Bio9、Bio10、Bio11、Bio12、Bio13、Bio14、Bio15、Bio16、Bio17、Bio18、Bio19、Bio20、Bio21、Bio22、Bio23
地形因子	DEM、aspect、slope
土壤类型	soil

表 2 生物气候指标及定义

Tab. 2 Bioclimatic indicators and definitions

指标	定义	量纲
Bio1	年平均温度	℃
Bio2	平均日较差 (月均最高温与月均最低温差值)	℃
Bio3	等温性 (Bio2 与 Bio7 比值)	
Bio4	温度季节性变异 (100 倍标准差)	
Bio5	最暖月最高温度	℃
Bio6	最冷月最低温度	℃
Bio7	温度年较差 (Bio5 与 Bio6 差值)	℃
Bio8	最湿季平均温度	℃
Bio9	最干季平均温度	℃
Bio10	最暖季度平均温度	℃
Bio11	最冷季度平均温度	℃
Bio12	年降水量	mm
Bio13	最湿月降水量	mm
Bio14	最干月降水量	mm
Bio15	降水量季节性变异 (变异系数)	
Bio16	最湿季降水量	mm
Bio17	最干季降水量	mm
Bio18	最暖季度降水量	mm
Bio19	最冷季度降水量	mm
Bio20	日均温高于 0℃ 的累积生长积温	℃
Bio21	日均温大于 5℃ 的累积生长积温	℃
Bio22	年潜在蒸散量	mm
Bio23	湿润指数 (Bio12 与 Bio22 比值)	

1.2.3 样本数据处理与区域筛选

本研究基础样本来源于 2015 年林业规划小班

表 3 数据属性

Tab.3 Introduction to data properties

数据源	空间分辨率	时间尺度	数据匹配方法
Sentinel-1	5 m × 20 m	2015 年	双线性插值法
Sentinel-2	10 m、20 m 和	2016—2018 年	
	60 m	6—8 月	
生物气候变量	1 km	1991—2020 年均值	
地形因子	30 m	2015 年	最近邻插值法
土壤类型	1 700 m × 1 700 m	2018 年	

数据和大兴安岭呼中林区林相图。小班数据由壮志、奋斗、长山、廿八站等多个林场实地调查获取,涵盖小班尺度的空间位置(样本点经纬度)、林分年龄、生物量以及树种等关键信息;在已公开发布的呼中林区林相图内随机提取部分样本点,弥补林业规划数据中空间分布相对稀疏的区域。同时,针对山杨样本数量相对不足的问题,系统整理文献[17]中可靠山杨样本点作为补充。

鉴于多源数据可能存在同一栅格对应多个树种标签、样本点重叠等情况,采用多数投票原则对样本进行筛选;对于同一栅格单元,仅当某一树种占据多数时,才将其确定为该栅格的最终优势树种;若出现平票的情况,则全部剔除。经过上述筛选过程,最终确定样本有:白桦 1 764 个、落叶松 1 570 个、山杨 919 个(图 1)。

2 数据处理与模型建构

本文选取中国寒温带森林的优势树种作为研究对象,根据 MCD12Q1 V6 产品的土地利用数据将非林地地区进行掩膜,综合光谱波段、纹理特征、地形因子及生物气候变量等多源数据,使用集成学习方法建立优势树种识别模型,并引入模型可解释性分析以提升结果的可靠性与生态解释力。整体流程主要包括数据处理与筛选、模型构建、特征解释 3 部分。

2.1 特征筛选

为降低高维纹理特征可能带来的信息冗余与计算负担,基于共线性分析方法对初始纹理特征进行筛选。首先,在研究区内随机抽取 50 000 个像元样本,提取其对应的 88 维特征值,计算各特征间的 Pearson 相关系数矩阵。随后采用逐步迭代剔除策略进行特征筛选:在每一轮迭代中,计算当前特征与其他特征间的平均相关系数,并移除平均相关性最高的变量,直至保留特征数量降至 15 个,该方法能够在尽量保留信息的同时有效降低特征间的冗余。

最终结合 15 个纹理特征(表 4)、11 个光谱波段(b2 ~ b8a、b9、b11、b12)、Sentinel-1 的 VV + VH

2 个极化通道、9 个植被指数、23 个生物气候变量、3 个地形因子(DEM、aspect、slope)、1 个土壤类型(soil)变量,共同构建 64 维多源特征集,用于模型训练与优势树种分布预测分析。

表 4 筛选后的纹理特征及描述

Tab.4 Filtered texture features and descriptions

变量	描述
纹理特征	b2_Cor、b2_Ent、b4_Con、b4_Cor、b4_Ent、b5_Hom、
	b7_Mean、b8_Mean、b8a_Mean、b9_Con、b9_Hom、
	b9_Var、b11_Hom、b12_Con、b12_Var

2.2 模型构建

2.2.1 模型分析

考虑到遥感影像及生态数据的高维度与非线性特征,选择 XGBoost、Random Forest、多层感知机(MLP)和一维卷积神经网络(Conv1D)4 种具有代表性的机器学习与深度学习模型进行对比分析。其中,XGBoost、Random Forest 作为经典的机器学习算法,具备较强的泛化能力与鲁棒性,是遥感分类任务中广泛采用的方法。XGBoost(Extreme gradient boosting)采用层次数生长策略,在传统提升树算法基础上引入正则项,通过约束模型复杂度有效缓解过拟合问题,准确度优于大多数机器学习模型,可扩展性强^[18]。Random Forest 通过构建多棵随机决策树并集成投票,有效降低单一模型的方差,增强对高维数据中非线性关系的捕捉能力^[19]。多层感知机(MLP)以及一维卷积神经网络(Conv1D)均为适用于光谱数据的经典网络结构,在近几年备受关注。其中 MLP 是人工神经网络(ANN)的重要分支,采用分层前馈拓扑结构,通过激活函数建立输入与输出变量间的非线性关系,进而提升模型输出稳健性^[20]。Conv1D 是卷积神经网络(CNN)的基础网络层类型,在一维空间中执行卷积操作,可有效捕捉数据局部特征,深入关注输入特征间的相互作用^[21]。

2.2.2 模型训练与评估

数据集采用分层抽样方式构建训练集与测试集,比例为 8:2。此外,在 XGBoost、Random Forest 训练过程中采用 5 折分层交叉验证,并结合随机搜索方式对模型关键超参数进行调优,以加权 F1 分数为优化目标,最优参数组合见表 5。MLP 采用 6 层全连接前馈结构(768、512、384、256、128 和 64 个神经元),各层均使用 ReLU 激活函数。Conv1D 网络包含 3 组卷积模块,每组由 2 个卷积层组成,卷积核尺寸由 7 逐步减小至 3,通道数从 128 逐级增加至 512。MLP、Conv1D 模型均引入 Dropout 层抑制过拟

表 5 机器学习最优参数组合
Tab.5 Optimal parameter combinations for machine learning

模型	参数	数值
XGBoost	subsample	0.88
	reg_lambda	0.03
	reg_alpha	0.015
	min_child_weight	1
	n_estimators	380
	max_depth	13
	learning_rate	0.01
	gamma	0.23
Random Forest	colsample_bytree	0.88
	n_estimators	380
	min_samples_split	2
	min_samples_leaf	1
	max_samples	0.9
	max_features	0.8
	max_depth	20

合,同时结合 Early Stopping 策略监控验证集性能。

为系统评估模型在优势树种识别中的分类能力,选用多种评价指标,包括总体准确率 (Accuracy)、混淆矩阵 (Confusion Matrix)、分类报告等。其中,分类报告进一步细化为精确率 (Precision)、召回率 (Recall) 与 F1 分数 (F1-score)^[22],衡量模型在不同类别上识别能力的优劣性。本研究引入多次随机划分的重复实验(5 次),对各模型进行多次训练与测试,计算准确率均值 ± 标准差。

2.3 特征重要性与可解释性分析

SHAP 值来源于 Shapley 值,用于公平地分配合作中的收益或成本,量化特征对模型输出特定树种类别的边际贡献。在全局尺度上,通过对所有树种类别、样本的 SHAP 取绝对值并计算均值,实现特征全局重要性排序;在局部尺度上,结合 SHAP 依赖图,分析特征交互对模型判别树种的影响机制。对于目标树种而言,若特征 SHAP 值为正,表明该特征推动模型将样本判别为该树种;若 SHAP 值为负,则表明该特征阻碍模型判别为该树种^[23]。在模型 f 中,输入特征为 $x = (x_1, x_2, \cdots, x_D)$,其中 D 是特征的总数。对于特征 i ,其 Shapley 值 $\phi_i(f, x)$ 表示为^[24]

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (D - |S| - 1)!}{D!} (f_x(S \cup \{i\}) - f_x(S))$$

(1)

式中 N ——所有特征的集合
 S —— N 中不包含特征 i 的任意子集

$f_x(S)$ ——模型在特征子集 S 上的条件期望值
 $|S|$ ——子集 S 中特征的数量

3 结果分析

3.1 整体性能比较

在本研究构建的模型框架中,传统机器学习模型的准确率整体高于神经网络模型(表 6)。XGBoost 模型总体精度(OA)可达 80.1% (78.2% ± 1.9%),Xi 等^[25]的研究也表明 XGBoost 在多数场景中的预测性能最佳。XGBoost 在准确性上优于大多数机器学习算法。其次是 Random Forest (77.3% ± 1.1%),尽管 Random Forest 和 XGBoost 均能有效捕捉高维特征之间的非线性关系,但 Random Forest 对训练样本类别均衡性要求较高,在本研究样本不平衡的条件下,其性能优势未能有效发挥^[19]。相比之下,深度神经网络 (MLP、Conv1D) 表现欠佳,OA 分别为 65.5% ± 2.4%、65.7% ± 2.9%。本研究样本量相对有限,网络训练难以充分优化大量参数,因而深度学习未表现出明显优势^[26]。

表 6 不同机器学习模型对优势树种分类精度评估
Tab.6 Evaluation of classification accuracy for dominant tree species using different machine learning models

		%			
机器学习模型	树种	精确率	召回率	F1 分数	总体精度 (OA)
XGBoost	白桦	79	79	79	78.2 ± 1.9
	落叶松	78	76	77	
	山杨	85	88	86	
Random Forest	白桦	78	79	78	77.3 ± 1.1
	落叶松	77	76	76	
	山杨	85	84	84	
MLP	白桦	69	67	68	65.5 ± 2.4
	落叶松	70	66	68	
	山杨	66	74	70	
Conv1D	白桦	70	69	70	65.7 ± 2.9
	落叶松	67	72	69	
	山杨	79	72	75	

在优势树种的分类效果上,山杨的可辨识度最高。XGBoost 模型对山杨的 F1 分数可达 86%,精确率与召回率亦优于白桦和落叶松,后两者的 F1 分数分别为 79% 和 77%;Random Forest 模型也表现出相近的判别性能。对于神经网络模型,Conv1D 的分类结果同样显示出山杨具有较高的可分性;相比之下,MLP 的整体分类性能略低,但从 F1 分数的分布来看,3 种树种之间的差异较小,白桦、落叶松与山杨的 F1 分数差值均不超过 2 个百分点。

图 2 给出了 4 种模型在测试集上的分类混淆矩

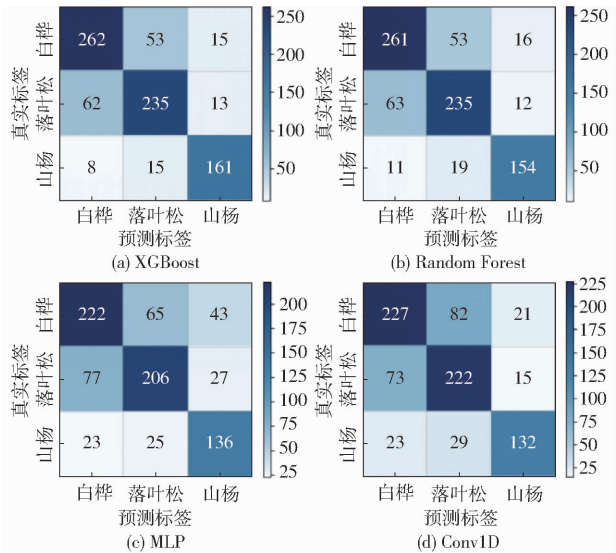


图 2 分类混淆矩阵对比

Fig. 2 Comparison of classification confusion matrices

阵。总体来看,XGBoost 模型在山杨的分类上表现较突出,正确分类的样本有 161 个;Random Forest 在落叶松上的分类表现相对稳定,正确分类的样本与 XGBoost 相同;对于非传统机器学习模型,Conv1D 总体略高于 MLP,白桦和落叶松预测标签符合实际的分别为 227、222 个,尽管 MLP 在山杨上的正确识别样本数较多,但其对白桦与山杨之间的误分现象较为明显。总体而言,白桦与落叶松之间的混淆在各模型中均较为突出,有研究也表示这种混淆可能与光谱特征、冠层结构相似性以及叶片的生理生化参数等有关,需要结合具体的生理生化参数进行深入探讨^[27]。

3.2 优势树种空间分布

3 个优势树种的空间分布特征在 4 种集成模型基本一致(图 3)。白桦主要分布于研究区西部;落叶松在东部、中部和南部均有大面积分布,表现出较强的区域主导性;山杨主要以零星斑块状分布于中部和东北部,其空间连续性相对较弱。不同模型之间的差异主要体现在东北部山杨和南部地区落叶松的识别上:Conv1D 模型在东北部几乎未识别出山杨,Random Forest 模型在该区域预测的山杨分布范围较小;MLP 模型在南部地区对落叶松的识别面积小于其他模型,结合其混淆矩阵中较高的误分率,可以推断 MLP 模型可能过度依赖训练样本已有的光谱细节,从而导致泛化能力不足。

为进一步评估不同模型预测结果的空间一致性,对 4 种模型的分类结果进行逐像元比较,统计每个像元中预测一致的模型数量,生成模型空间一致性图(图 4,图中 2、3、4 分别表示有 2、3、4 个模型预测结果一致)。结果表明,大部分区域具有较高的

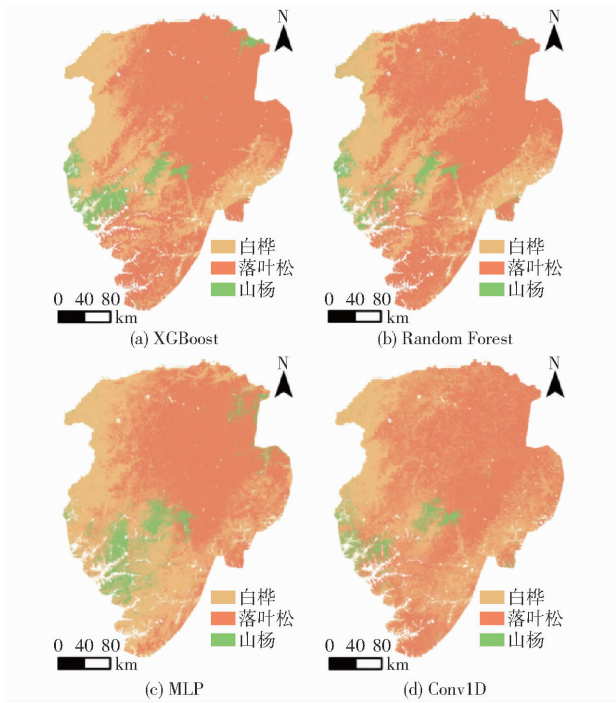


图 3 优势树种分类结果对比

Fig. 3 Comparison of classification results for dominant tree species

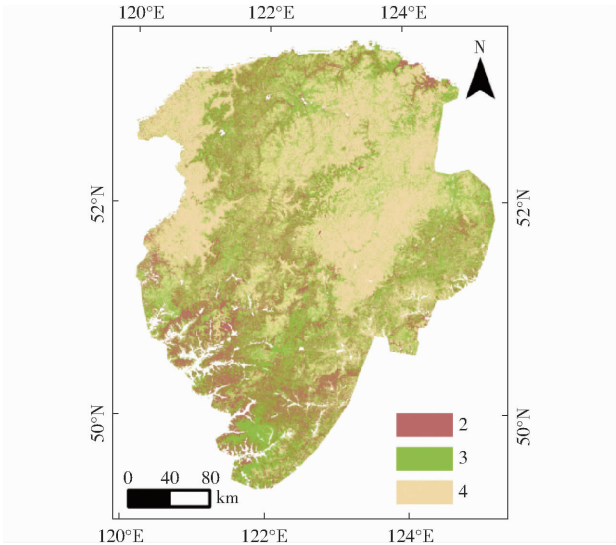


图 4 4 种模型预测结果的空间一致性

Fig. 4 Spatial consistency of predictions among four models

一致性,3、4 个模型预测结果相同;仅在少数区域出现 2 个模型一致的情况。总体而言,不同模型在研究区优势树种分布预测上具有较高的一致性和稳定性。

3.3 XGBoost 模型可解释性分析

本文引入 SHAP 方法,从全局与局部 2 个层面量化特征对 XGBoost 输出的边际贡献,系统解析变量在分类决策中的作用方向与影响强度(为方便展示,仅显示 20 个最大值)。研究表明,温度年较差(Bio7)、温度季节性(Bio4)、最干月降水量(Bio14)

和等温性 (Bio3) 是模型区分树种最重要的 4 个特征 (图 5)。其中, Bio7 和 Bio4 反映了研究区大陆性气候特征, 说明该地树种对较大的年温差和季节性温度变化具有较强适应能力。中国寒温带森林位于欧亚北方森林南缘过渡区, 对水热条件变化敏感, 这一发现进一步凸显了气候的波动对树种分布的关键调控作用。其具体机制还需要结合 SHAP 蜂巢图进行对比分析。

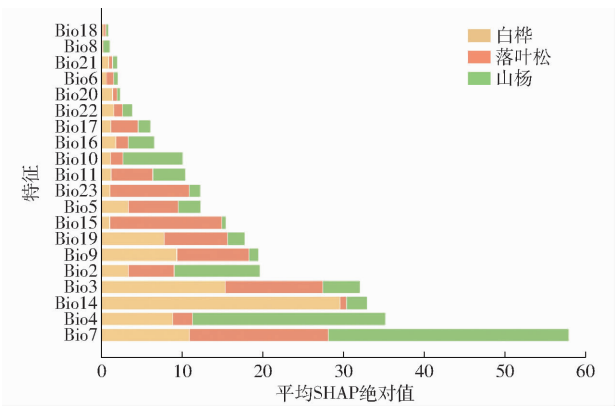


图 5 全局特征变量重要性的可解释性

Fig. 5 Interpretability of global feature importance

图 6 展示了 3 类优势树种的特征变量重要性分

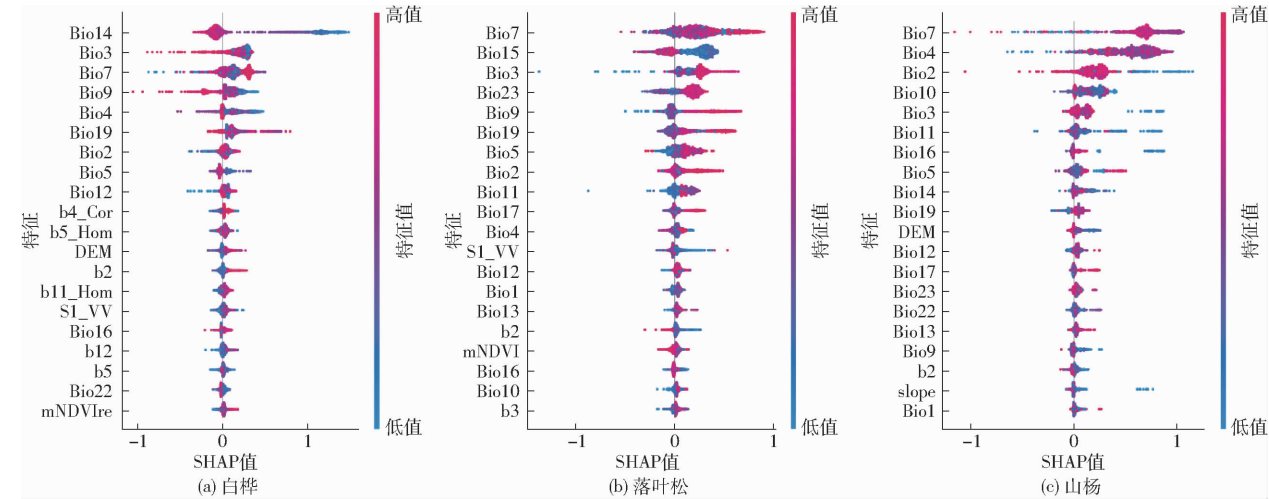


图 6 局部特征变量重要性的可解释性

Fig. 6 Interpretability of local feature importance

图 7 进一步展示了各树种依赖性最强的 4 对特征组合, 以刻画气候因子之间的交互作用, 图中 (+) 区域为正向交互作用, (-) 区域为反向交互作用。对白桦而言 (图 7a), Bio3 与最冷季度降水 (Bio19) 呈正向交互作用, 当 Bio3 介于 0.53 ~ 0.55 且 Bio19 处于 14 ~ 17 mm 之间时, SHAP 值为正, 即该变量组合对白桦的预测结果具有正向贡献。表明在年温差较大的背景下, 冷季降水有助于改善越冬水分供给、减少低温冻害, 同时也突出了白桦耐寒的特点^[30]。在落叶松中 (图 7b), 湿润指数 (Bio23) 与等温性 (Bio3) 在一定条件下呈正向交互, 当 Bio23

布, 不同变量在高值或低值区间对模型分类的贡献存在差异, 图中每个点代表样本的 SHAP 值, 并根据特征值的大小进行颜色编码 (红色表示高值, 蓝色表示低值)。对白桦而言 (图 6a), 较高的日较差 (Bio2) 和较低的 Bio14 对应正的 SHAP 值, 可提升模型对白桦的识别概率, 表明白桦对较大的昼夜温差及季节性干旱具有一定耐受性, 已有研究也指出桦树在快速变化的环境中具有出色的适应能力, 和其本身的高可塑性及基因型变异有关^[28]。在落叶松中 (图 6b), 降水季节分配较均匀 (Bio15) 且冬季较温和 (Bio11) 的区域更易被识别为落叶松, 已有研究结果也证明在干旱的条件下, 降水增加能缓解压力, 促进落叶松径向生长^[29]; 而降水过于集中, 达到饱和时, 可能导致根系缺氧, 反而成为限制因素, 此外, 冬季相对温和的环境可以降低低温胁迫对其存活的影响。对于山杨 (图 6c), 本研究发现 Bio4 大的环境下, 模型更易识别出山杨, 反映了其对寒冷冬季和温暖夏季这种典型气候的适应。上述结果不仅揭示了关键特征在树种分类中的贡献方向与强度, 也反映了不同树种对生态环境条件的偏好。

大于 0.75 且 Bio3 大于 0.555 时, SHAP 值为正, 表明落叶松更倾向于分布在温度波动较小且水分条件相对稳定的环境, 均衡的气候条件有助于缓解干旱或低温胁迫, 为落叶松的存活、生长提供更有利的环境条件^[31]。对于山杨 (图 7c), Bio4 与 Bio2 同样表现出正向交互关系, 当 Bio4 介于 1 650 ~ 1 700 之间时, 较高的 Bio2 值通常对应正的 SHAP 值, 提升模型对山杨的判别能力, 说明在年内温度波动较大的背景下, 适度的温度梯度可能增强山杨的竞争优势, 闫沛斌等^[32]研究也表明温度季节变化会影响山杨的呼吸速率, 促进有机质的积累。优势树种分布变

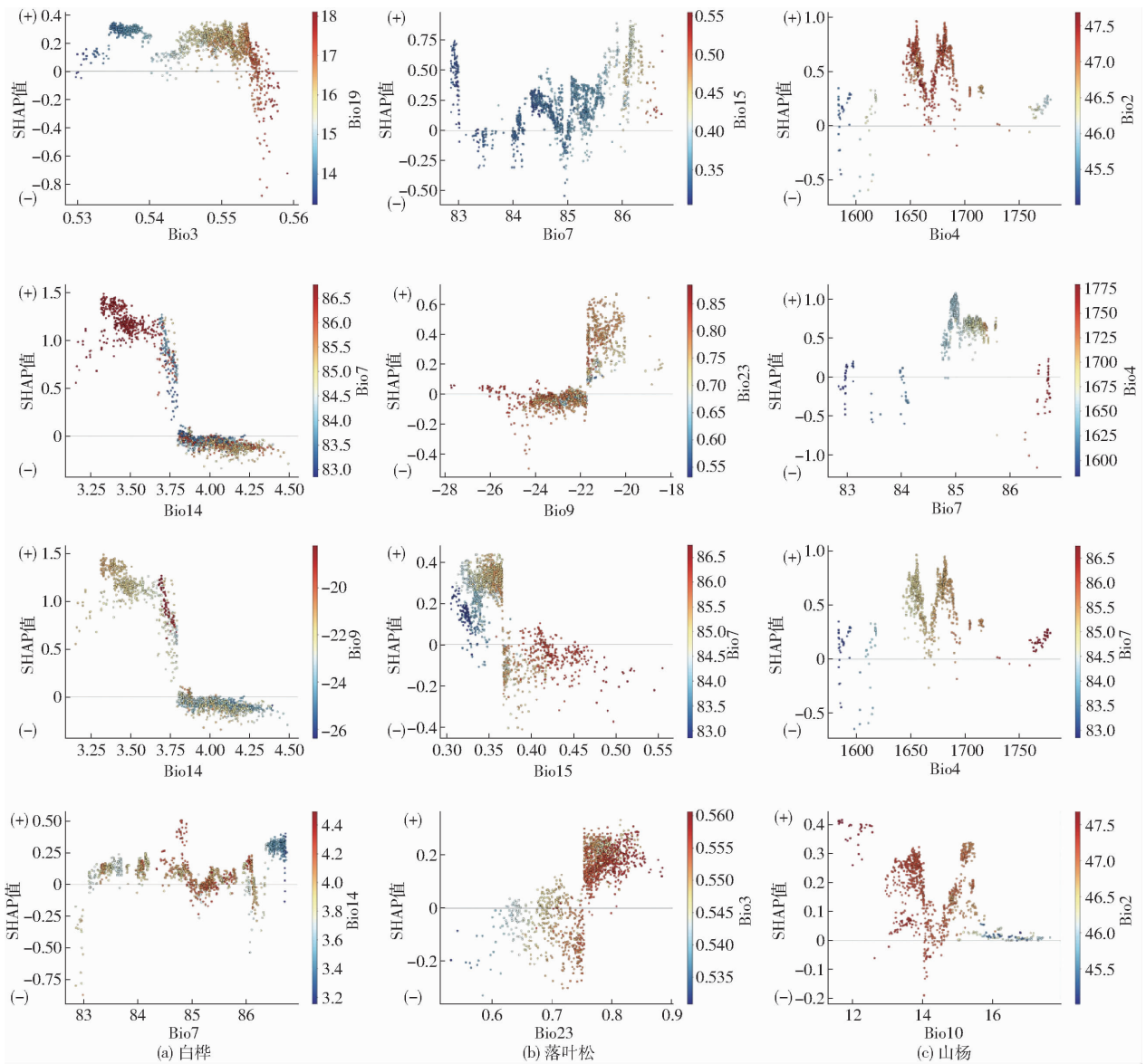


图 7 关键特征对交互依赖性
Fig. 7 Key feature interaction dependence of key features

化与气候因子之间的关系并非简单的线性关系,这与物候对气候响应的相关研究结果一致^[33]。这种非线性交互关系强调了单一因子分析的局限性,在预测未来气候变化对树种分布的影响时,必须考虑关键气候变量间的协同交互效应。

3.4 分析讨论

目前,XGBoost 模型已在中国寒温带地区取得了较高的预测精度。未来的研究将进一步评估该模型在整个欧亚北方森林区域的泛化能力与鲁棒性,从而构建整个北方森林数据集。然而,现有的气候数据原始空间分辨率大多为 1 km,且站点数据有限,重采样至更高分辨率仍难以充分反映栅格尺度的空间差异,有可能导致生物气候变量与光谱、纹理特征之间的尺度不匹配,使模型结果存在潜在偏差。此外,本研究样本数量仍存在一定差异,山杨样本相对较少,可能会在一定程度上影

响模型对优势树种学习的平衡性。尽管 XGBoost 通过梯度提升与正则化机制,可缓解样本不平衡带来的偏置问题,但在更大空间尺度或包含更多树种类别的研究中,样本不平衡问题将更为突出。未来研究可考虑引入更高精度环境数据、扩充样本规模、优化样本空间分布,提升模型在欧亚北方森林复杂环境下的预测性能。

同时,本研究构建的适用于寒温带森林复杂环境的机器学习分类模型,融合了多源遥感数据,可以显著提升过渡带优势树种的识别能力;并且本研究中可解释性分析框架具有良好的扩展性,在未来可推广至树种动态监测、潜在分布预测等研究。

4 结论

(1)在寒温带森林优势树种的分类中,XGBoost 模型取得最高总体精度(OA 为 80.1%),整体误分

率最低。其中,山杨在各模型中均表现出较高的可分性。

(2)生物气候变量在模型判别中起主导作用。降水 (Bio14) 及与温度变化 (Bio7、Bio4、Bio3) 相关的变量贡献显著高于光谱、纹理及地形因子,反映了寒温带森林优势树种对气候条件变化的高度敏感性。

(3)生物气候变量存在复杂的非线性交互作用。山杨更倾向于分布在温度季节性 (Bio4) 和平均日较差 (Bio2) 较大的区域,表明其在极端气候条件下具备较强的适应能力。这种非线性的、物种独特的气候响应特征强调在优势树种分布研究中,相比单变量分析,需优先考虑多变量交互效应。

参 考 文 献

[1] Guo Y, Zhou J, Dong Q, et al. Refined high-definition map model for roadside rest area[J]. *Transportation Research Part A: Policy and Practice*, 2025, 195: 104463.

[2] Felton A, Felton A M, Wam H K, et al. Forest biodiversity and ecosystem services from spruce-birch mixtures: the potential importance of tree spatial arrangement[J]. *Environmental Challenges*, 2022, 6: 100407.

[3] Feng M, Sexton J O, Wang P, et al. Northward shift of boreal tree cover confirmed by satellite record[J]. *Biogeosciences*, 2026, 23(3): 1089–1101.

[4] Hermosilla T, Wulder M A, White J C, et al. Characterizing long-term tree species dynamics in Canada’s forested ecosystems using annual time series remote sensing data[J]. *Forest Ecology and Management*, 2024, 572: 122313.

[5] Axelsson A, Lindberg E, Reese H, et al. Tree species classification using Sentinel – 2 imagery and Bayesian inference[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2021, 100: 102318.

[6] 刘晟屹, 温一博, 武锦炜, 等. 基于 Sentinel – 2 多维特征融合的铁岭市优势树种分类[J]. *森林工程*, 2025, 41(6): 1101–1115.

Liu Shengyi, Wen Yibo, Wu Jinwei, et al. Classification of dominant tree species in Tieling based on Sentinel – 2 multi-dimensional feature fusion[J]. *Forest Engineering*, 2025, 41(6): 1101–1115. (in Chinese)

[7] Lim J, Kim K-M, Kim E-H, et al. Machine learning for tree species classification using Sentinel – 2 spectral information, crown texture, and environmental variables[J]. *Remote Sensing*, 2020, 12(12): 2049.

[8] 晏寒冰, 彭丽潭, 唐旭清. 基于气候变化的东北地区森林树种分布预测建模与影响分析[J]. *林业科学*, 2014, 50(5): 132–139.

Yan Hanbing, Peng Litan, Tang Xuqing. Modeling and impact analysis on distribution prediction of forest tree species in Northeast China based on climate change[J]. *Scientia Silvae Sinicae*, 2014, 50(5): 132–139. (in Chinese)

[9] Jiang S, Sweet L, Blougouras G, et al. How interpretable machine learning can benefit process understanding in the geosciences [J]. *Earth’s Future*, 2024, 12(7): e2024EF004540.

[10] Qian Q, Jia X. Seasonal forecast of winter precipitation over china using machine learning models[J]. *Atmospheric Research*, 2023, 294: 106961.

[11] 王芳, 李炳元, 田思雨, 等. 中国生态地理区划更新和优化[J]. *地理学报*, 2024, 79(1): 3–16.

Wang Fang, Li Bingyuan, Tian Siyu, et al. Updated scheme for eco-geographical regionalization in China [J]. *Acta Geographica Sinica*, 2024, 79(1): 3–16. (in Chinese)

[12] Rao G, Yan S-Z, Song W-L, et al. Distribution, assembly, and interactions of soil microorganisms in the bright coniferous forest area of China’s cold temperate zone[J]. *Science of the Total Environment*, 2023, 897: 165429.

[13] 王康, 万轶炜. 基于哨兵 1 号数据的鄱阳湖水体面提取及应用[J]. *江西测绘*, 2024(2): 19–22.

Wang Kang, Wan Yiwei. Water extraction of Poyang Lake based on Sentinel – 1 data[J]. *Jiangxi Surveying and Mapping*, 2024(2): 19–22. (in Chinese)

[14] 张超, 陈畅, 徐海清, 等. 基于 XGBoost 算法的多云多雾地区多源遥感作物识别[J]. *农业机械学报*, 2021, 53(4): 149–156.

Zhang Chao, Chen Chang, Xu Haiqing, et al. Multi-source remote sensing crop identification based on XGBoost algorithm in cloudy and foggy area[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2021, 53(4): 149–156. (in Chinese)

[15] Zhang B, Wang Z, Liang B, et al. A lightweight spatiotemporal classification framework for tree species with entropy-based change resistance filter using satellite imagery[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2025, 138: 104449.

[16] 王博一, 董灵波, 刘兆刚. 不同气候情景下小兴安岭 3 个珍稀易危树种的潜在分布预测[J]. *中南林业科技大学学报*, 2024, 44(2): 29–40.

Wang Boyi, Dong Lingbo, Liu Zhaogang. Prediction of potential distribution of three protective tree species in Xiaoxing’an Mountain under different climatic scenarios[J]. *Journal of Central South University of Forestry and Technology*, 2024, 44(2): 29–40. (in Chinese)

[17] Du S, Wang Z, Ingvarsson P K, et al. Multilocus analysis of nucleotide variation and speciation in three closely related

- Populus* (*Salicaceae*) species[J]. *Molecular Ecology*, 2015, 24(19): 4994–5005.
- [18] 侯轶, 张霞, 郑海光, 等. 综合光谱多特征与 XGBoost 模型的土壤高光谱遥感分类方法[J]. *农业机械学报*, 2025, 56(11): 612–620.
Hou Yi, Zhang Xia, Zheng Haiguang, et al. Hyperspectral remote sensing classification method for soil by combining spectral multi-feature and XGBoost model[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2025, 56(11): 612–620. (in Chinese)
- [19] Belgiu M, Drăguț L. Random forest in remote sensing: a review of applications and future directions[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016, 114: 24–31.
- [20] Ayodele B V, Mustapa S I, Kanthasamy R, et al. Modeling the prediction of hydrogen production by co-gasification of plastic and rubber wastes using machine learning algorithms[J]. *International Journal of Energy Research*, 2021, 45(6): 9580–9594.
- [21] Liu Q, Liu Y, Son H. The memory cycle of time-series public opinion data: validation based on deep learning prediction[J]. *Information Processing & Management*, 2025, 62(4): 104168.
- [22] Yu K, Hao Z, Post C J, et al. Comparison of classical methods and Mask R-CNN for automatic tree detection and mapping using UAV imagery[J]. *Remote Sensing*, 2022, 14(2): 295.
- [23] 杨佳, 斛如媛, 赵海平, 等. 基于 XGBoost-SHAP 的三峡库区生态屏障区生态系统服务权衡及驱动力分析[J]. *水土保持学报*, 2025, 39(6): 127–138.
Yang Jia, Hu Ruyuan, Zhao Haiping, et al. Analysis of trade-offs and driving forces of ecosystem services in ecological barrier zone of Three Gorges Reservoir Area based on XGBoost-SHAP model[J]. *Journal of Soil and Water Conservation*, 2025, 39(6): 127–138. (in Chinese)
- [24] 张宏鸣, 侯贵河, 孙志同, 等. 基于 XGBoost-Shapley 的玉米不同生育期 LAI 遥感估算[J]. *农业机械学报*, 2022, 53(7): 208–216.
Zhang Hongming, Hou Guihe, Sun Zhitong, et al. Remote sensing estimation of maize leaf area index at different growth periods based on XGBoost-Shapley algorithm[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2022, 53(7): 208–216. (in Chinese)
- [25] Xi Y, Ren C, Tian Q, et al. Exploitation of time series Sentinel-2 data and different machine learning algorithms for detailed tree species classification[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 7589–7603.
- [26] Benkendorf D J, Hawkins C P. Effects of sample size and network depth on a deep learning approach to species distribution modeling[J]. *Ecological Informatics*, 2020, 60: 101137.
- [27] 何诗静. 基于叶片和冠层级别的高光谱城市树种识别[D]. 武汉: 华中农业大学, 2012: 58.
He Shijing. Urban tree species discrimination based on leaf and canopy level hyperspectral data[D]. Wuhan: Huazhong Agricultural University, 2012: 58. (in Chinese)
- [28] Oksanen E. Birch as a model species for the acclimation and adaptation of northern forest ecosystem to changing environment[J]. *Frontiers in Forests and Global Change*, 2021, 4: 682512.
- [29] Kirdeyanov A V, Proushkin A S, Tabakova M A. Tree-ring growth of Gmelin larch under contrasting local conditions in the north of Central Siberia[J]. *Dendrochronologia*, 2013, 31(2): 114–119.
- [30] Zhang Q, Pei X, Xu L, et al. Genetic improvement of *Betula platyphylla* Suk. in China: a review[J]. *Phyton-International Journal of Experimental Botany*, 2022, 91(8): 1585–1599.
- [31] 孙振静, 赵慧颖, 朱良军, 等. 大兴安岭北部不同降水梯度下兴安落叶松生长对升温的响应差异[J]. *北京林业大学学报*, 2019, 41(6): 1–14.
Sun Zhenjing, Zhao Huiying, Zhu Liangjun, et al. Differences in response of *Larix gmelinii* growth to rising temperature under different precipitation gradients in northern Daxing'an Mountains of northeastern China[J]. *Journal of Beijing Forestry University*, 2019, 41(6): 1–14. (in Chinese)
- [32] 闫沛斌, 王子乔, 王翠英, 等. 甘肃兴隆山油松林和山杨-白桦林土壤呼吸特征及其影响因素[J]. *甘肃林业科技*, 2025, 50(3): 42–48.
Yan Peibin, Wang Ziqiao, Wang Cuiying, et al. Soil respiration characteristics and its influencing factors of two forest stands in Xinglong Mountain in Gansu Province[J]. *Journal of Gansu Forestry Science and Technology*, 2025, 50(3): 42–48. (in Chinese)
- [33] Xiong T, Du S, Zhang H, et al. Satellite observed reversal in trends of spring phenology in the middle-high latitudes of the northern hemisphere during the global warming hiatus[J]. *Global Change Biology*, 2023, 29(8): 2227–2241.