森林地上生物量（Above-Ground Biomass, AGB）是反映森林碳储量和碳循环过程的关键指标，对评估森林碳汇能力、监测气候变化背景下森林生态系统响应具有重要意义。然而，传统野外调查方法耗时、耗力且难以获得大范围连续观测数据，难以满足长期、区域乃至全球尺度的监测需求。基于遥感数据的 AGB 估计提供了一种高效、非破坏性的解决方案，被广泛应用于不同空间尺度的森林生态研究之中[1]。

遥感 AGB 估计方法通常分为基于光学、多光谱影像、雷达和 LiDAR 等多源数据的特征提取以及后续的回归建模两个核心环节。Lu et al.（2014）[2]对森林生态系统中基于遥感的生物量估计方法进行了系统综述，指出多源遥感数据融合能够显著克服单一传感器饱和效应问题；同时，模型不确定性和尺度效应仍是亟待解决的关键技术挑战。近年来，有学者[3]在新疆区域研究中通过 Boruta 算法筛选变量，并结合 XGBoost、RF、SVM 等机器学习方法，取得了省级尺度四大林型 AGB 空间映射的良好结果，验证了基于机器学习的模型在复杂生态系统中的可行性与优势。

在机器学习算法方面，随机森林（Random Forests, RF）以其对高维数据的鲁棒性和内置重要性评估功能被广泛应用于 AGB 估计；其核心思想来源于 Bagging 和随机子空间采样，能够有效降低模型方差并提升泛化能力[4]。XGBoost 则通过加权量化概括（weighted quantile sketch）和稀疏感知算法实现了可扩展的梯度提升树（Gradient Boosting Tree）系统，为大规模数据场景下的树模型训练提供了高效支持[5]。LightGBM 同样作为另一种高效梯度提升框架，以其基于直方图（Histogram-based）分裂决策和 Leaf-wise 策略获得了更快的训练速度与更低的内存消耗。

在特征选择方面，递归特征消除与交叉验证（RFECV）结合了包裹式（wrapper）和过滤式（filter）思想，通过递归地消除对模型贡献度最低的特征并在交叉验证中评估模型性能，有效自动化确定最优特征子集[6]。此过程能最大限度地减少冗余变量的干扰，同时保证模型复杂度与预测性能的平衡。

尽管上述方法各自取得了显著进展，但在多年份、跨尺度的 AGB 估计问题中，各阶段特征筛选策略、模型选择与参数调优流程仍存在脱节，导致整体估计性能难以达到最优。为此，本文通过多级特征筛选、多模型超参数随机搜索与嵌套交叉验证、以及模型堆叠技术的综合应用，实现了不同年份遥感-实测数据集上 AGB 估计。该工作不仅为大范围森林碳监测提供了可靠工具，也为遥感大数据环境下的机器学习建模方法论贡献了可推广的方案。

在本研究中，设计实现了一套流程，用于基于遥感样本点光谱数据的森林地上生物量（AGB）回归估计。该流程包括数据预处理、缺失值填补、多级特征选择、模型候选与超参数调优、模型集成与评估五大模块，具体方法如下。

首先，对原始 CSV 格式的样本数据进行统一读取与基本校验。对于每个年份的数据集，程序会检查是否包含样本标识列（ID）与目标变量列（AGB），并在缺失时自动生成或报错补全。自变量矩阵 X 由除 ID 与 AGB 外的所有特征列构成，因子类型特征可在后续扩展时接入独热编码或嵌入式编码。

针对 X 中可能存在的缺失值，采用中位数填补策略，即对每个特征列 j 计算其中位数 $\tilde{x}_j$，并将缺失位置统一替换为 $\tilde{x}_j$。该方法对异常值具有一定鲁棒性，能够避免平均值填补对分布偏态数据的失真。

预处理完成后，数据集按 80%：20% 的比例随机划分为训练集与测试集，划分过程同时保留样本 ID 以便后续结果追踪。划分使用固定随机种子以保证实验可重复性。划分后，将训练集进一步用于特征选择与模型调优，测试集仅在最终模型评估阶段使用。

特征选择模块分为三阶：首先基于训练集计算每个候选特征 x_j 与 AGB 之间的皮尔逊相关系数

$$\rho_j = \frac{\mathrm{cov}(x_j, y)}{\sigma_{x_j}\,\sigma_y}$$

并取其绝对值排序，保留与目标变量绝对相关系数 $|\rho_j|$ 大于预设阈值（0.1）的初步特征子集；若此阈值筛后特征数过少，则取相关性最高的前 10 个特征以保证后续步骤的稳定性。

在初步筛选基础上，进行基于随机森林的快速重要性过滤。具体地，训练一个包含 100 棵树的随机森林，对每棵树中各特征的划分贡献（Gini 或方差减少量）累加求均值，得到特征重要性评分 I_j。以所有初步特征重要性中位数 $\tilde{I}$ 作为阈值，选取 $I_j \geq \tilde{I}$ 的特征进入下一阶段；若此步骤后特征数少于 3，则同样回退至相关性排序前 5 的特征以确保模型可训练性。

最终精筛环节采用递归特征消除结合交叉验证。在该过程中，令当前特征子集为 F，对其训练一个基学习器（随机森林），并依据特征重要性每次剔除对模型性能贡献最小的单个特征，然后通过五折交叉验证评估剔除后模型在负均方误差指标下的平均表现，直至性能最优。此算法可形式化为：

1. 初始 F_0 快速过滤后特征集；

2. 对 F_t 训练模型并计算每个 $j \in F_t$ 的重要性得分 $I_j^{(t)}$；

3. 从 F_t 中移除$\arg\min_j I_j^{(t)}$，得到 F_{t+1}；

4. 若 $\mathrm{CV\_score}(F_{t+1}) < \mathrm{CV\_score}(F_t)$，则更新最优子集；否则终止。

至此，获得最终特征子集 F^*，用于后续模型训练。

在模型候选与超参数调优阶段，选取包括随机森林（RF）、XGBoost、LightGBM、以及多层感知机回归（MLPRegressor）在内的四类算法作为基模型。针对每一基模型，定义了涵盖树个数、树深、学习率、叶子数、正则化强度、隐藏层结构与激活函数等维度的超参数搜索空间，采用随机搜索进行 20 次采样。每次评估内部嵌套五折交叉验证，以负均方误差为优化目标函数，确保在超参数调优过程中有效控制过拟合风险并获取稳定性能估计。调优完成后，对每一基模型使用相同折数的交叉验证统计出 RMSE、MAE 与 R^2 的分布，以便模型间公平比较。

若基模型数量多于两个，进一步构建堆叠回归器，其底层使用表现最优的两种基学习器，终级学习器采用带正则化的岭回归（Ridge）。堆叠模型通过交叉验证学习各基模型预测输出的线性最优组合，从而充分利用不同算法的互补性以提升泛化性能。

在测试集评估环节，对所有候选模型及堆叠模型在保留特征 F^* 上进行预测，并计算 RMSE、MAE 与 R^2 三项指标。最终选择测试 RMSE 最小的模型作为年度最佳模型，并对其在测试集上绘制真实值—预测值散点图、残差分布直方图、以及基于训练集的学习曲线，以直观呈现模型的拟合效果与稳定性。