Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(3), 876-887
Published Online May 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.153081

基于数据融合的日冕物质抛射传播时间预报

朱华杰，王家豪

南京师范大学商学院，江苏 南京

收稿日期：2026年3月30日；录用日期：2026年5月17日；发布日期：2026年5月28日

摘　要

日冕物质抛射(CMEs)作为极具破坏性的空间天气现象，其传播时间精准预报对保障航天器安全、卫星导航及电力网络稳定至关重要。针对现有研究未充分融合LASCO观测图像与文本物理参数的空白，本文提出一套完整的数据融合方案与预报模型。首先通过时间戳匹配，实现255个日冕物质抛射事件的时序图像与19维文本物理特征(含运动学、源区磁场、背景太阳风参数)合并；随后基于迁移学习的ResNet50模型提取图像深度特征，经降维得到5维图像特征，与文本特征拼接形成24维融合数据集。在此基础上，本文构建了XGBoost、随机森林模型，并进一步沿用二者的堆叠集成学习模型。结果显示，堆叠集成学习模型表现最优，在测试集上的平均绝对误差低至7.8048小时，R^2达0.7613，显著优于两种单一模型，充分验证了数据融合的有效性与集成学习的互补优势。此外，本文还将该学习模型与经典的DBM模型进行对比分析，发现模型学习趋势与实际物理规律契合，从而检验了模型的有效性。

关键词

日冕物质抛射，堆叠集成学习，传播时间预报

Forecast of Coronal Mass Ejection Propagation Time Based on Data Fusion

Huajie Zhu, Jiahao Wang

Business School of Nanjing Normal University, Nanjing Jiangsu

Received: March 30, 2026; accepted: May 17, 2026; published: May 28, 2026

Abstract

Coronal Mass Ejections (CMEs), as highly destructive space weather phenomena, require precise forecasting of their propagation time to ensure the safety of spacecraft, satellite navigation, and the stability of power networks. Addressing the gap in existing research where LASCO observational images and textual physical parameters have not been fully integrated, this paper proposes a comprehensive data fusion scheme and forecasting model. Initially, through timestamp matching, we

文章引用: 朱华杰, 王家豪. 基于数据融合的日冕物质抛射传播时间预报[J]. 人工智能与机器人研究, 2026, 15(3): 876-887. DOI: 10.12677/airr.2026.153081

merge the time-series images of 255 CME events with 19-dimensional textual physical features, encompassing kinematics, source region magnetic field, and background solar wind parameters. Subsequently, we employ a transfer-learning-based ResNet 50 model to extract image deep features, which are then reduced to 5-dimensional image features. These features are concatenated with the textual features to form a 24-dimensional fused dataset. Based on this, this paper constructs XGBoost and Random Forest models, and further employs their stacked ensemble learning model. The results show that the stacked ensemble learning model performs optimally, with an average absolute error as low as 7.8048 hours on the test set and an R^2 of 0.7613, significantly outperforming the two individual models. This fully validates the effectiveness of data fusion and the complementary advantages of ensemble learning. Furthermore, this paper compares and analyzes this learning model with the classic DBM model, finding that the model learning trend aligns with actual physical laws, thus verifying the model's effectiveness.

Keywords

Coronal Mass Ejections, Stacked Ensemble Learning, Propagation Time Forecast

1. 引言

日冕物质抛射(CMEs)是一类破坏力极强的空间天气事件，具体表现为携带着磁场的巨型等离子体团从太阳表面高速抛射至行星际空间。这类现象一旦与地球磁层发生相互作用，就可能诱发强度极高的地磁暴，进而对在轨航天器的运行安全、卫星导航系统的精准度以及地面电力网络的稳定运转等，构成严峻的威胁。通常情况下，日冕物质抛射发生时，还会伴随耀斑爆发、暗条瓦解、射电暴、太阳喷流以及磁流体力学波(MHD)等一系列大范围的太阳活动。当裹挟着磁化等离子体的日冕物质抛射体从太阳出发向外扩散，并抵达地球周边区域时，会与背景太阳风产生相互作用，造成各类物理参数的改变，而所有这些变化过程，都要遵循磁流体力学方程的相关约束。

也正因为如此，借助数值计算方法求解磁流体力学方程是当今较为有效的测度方式。特别是基于观测数据搭建的三维全域磁流体力学仿真模型[1]，能真实复刻行星际背景环境对日冕物质抛射传播过程的作用机制。凭借该特性，磁流体力学模型可以更为精确地预估日冕物质抛射从太阳到地球的传输时长，乃至推算出其抵达地球时的各项核心参数。不过，要将磁流体力学模型应用到复杂多变的行星际空间场景中，仍需攻克不少技术难关，例如进一步提高计算效率与精准度、维持磁场的无散度属性等。而机器学习方法的兴起，恰好为突破这些技术瓶颈提供了新的解决方案。

近年来，随着计算能力的快速提升以及人工智能算法理论的不断完善，机器学习已被引入空间物理学问题的求解过程中。既有研究运用人工神经网络方法预测 1 天文单位处的太阳风速度[2]。有学者借助卷积神经网络(CNN)，直接从活动区磁图中提取预测参数，进而预测 C 级、M 级和 X 级耀斑的发生情况[3]等。在日冕物质抛射从太阳到地球传播时间的预测方面，另有学者利用支持向量机算法，以 182 个具有地磁效应的日冕物质抛射事件作为样本进行预测模型训练[4]。在对整个样本数据集进行 10 万轮交叉验证后，在训练集和测试集的最优分割情况下，取得了约 5.9 小时的最佳平均绝对误差。也有学者采用 CNN 方法，将一系列白光观测数据作为输入来预测日冕物质抛射的到达时间，所获得的平均绝对误差约为 12.4 小时[5]。此前其他相关研究的平均绝对误差约为 10 小时[6]。不难发现，现有相关研究仍未关注

LASCO 观测所得图像数据与文本数据的融合处理问题，并进行针对性预测。对于这一研究空白，本文旨在完成两类数据的融合，并构建基于融合数据的预报模型，以提升日冕物质抛射到达时间的预测精度。

基于上述讨论，本文提出一套文本与图像数据的融合方案，旨在整合数据的互补信息以提升预报性能。首先通过时间戳匹配实现 LASCO 时序图像与日冕物质抛射物理参数的关联对齐；随后利用 ResNet 50 提取图像深度特征，并结合主成分分析降维，与 19 维物理特征融合构建统一数据集；后续则通过对比 XGBoost、随机森林及堆叠集成模型，验证融合特征的在时间预报方面的有效性。

2. 数据说明

为实现两种数据的有效融合，先介绍数据存储情况：

1) 文本数据：以 Excel 表格形式存储，记录了每次日冕物质抛射事件的物理参数，共包含 19 个特征。如表 1 所示，其主要类别包括：

Table 1. Text data
表 1. 文本数据

日冕物质抛射运动学参数	源区与磁场参数	背景太阳风参数
角宽度	磁压力平衡参数	太阳风等离子体温度
线性速度	磁场强度 x 分量	太阳风质子密度
初始高度处的二阶速度	磁场强度 y 分量	太阳风等离子体速度
最终高度处的二阶速度	磁场强度 z 分量	太阳风等离子体经度角
二阶速度		太阳风等离子体纬度角
加速度		氦核与质子的数量比
		太阳风流动压力
		等离子体贝塔参数

2) 图像数据：以 PNG 格式存储，为 LASCO 等日冕仪观测的日冕物质抛射时序图像。每个事件对应一个独立的文件夹，内含从日冕物质抛射爆发到传播至一定高度的多张追踪图像。具体示例如图 1 所示：

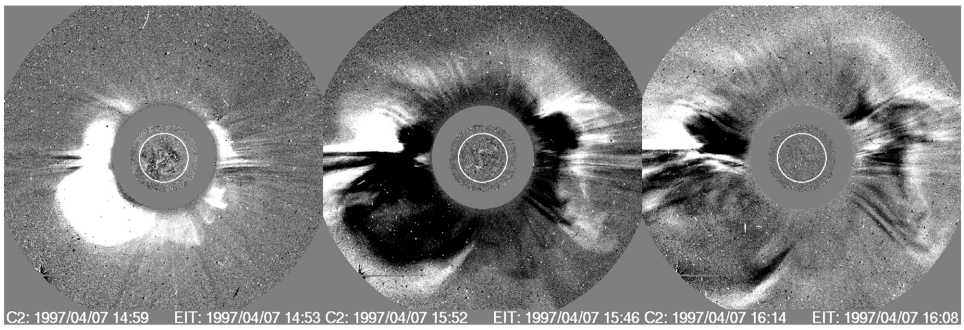

Figure 1. Image data example
图 1. 图像数据示例

3. 文本与图像的融合

为实现数据的有效融合融合，本文采用如下三步：

Step 1：将非结构化的图像文件夹与结构化的表格数据行进行精确匹配。通过从文本数据中提取事件

初发时间，再根据图像文件夹的命名规则(通常包含观测日期和时间)，生成两种可能的时间格式候选，通过遍历图像根目录，为每个文本事件匹配到唯一对应的图像文件夹。

Step 2：图像特征深度提取，为从图像中提取比手工特征更丰富的语义信息，本课题采用迁移学习策略：采用在 ImageNet 上预训练的 ResNet50 模型，移除其最后的分类层，替换为一个输出维度为 2048 的全连接层，构成一个深度特征提取器。并取消参数冻结，让其在本文的图像识别领域进行自适应微调，为使其特征更适应太阳日冕图像，使用关联成功的图像数据对提取器进行无监督微调，通过最小化输出特征的方差使其对太阳图像更具判别性。此外，如图 2 所示，本文绘制了卷积的示意图。

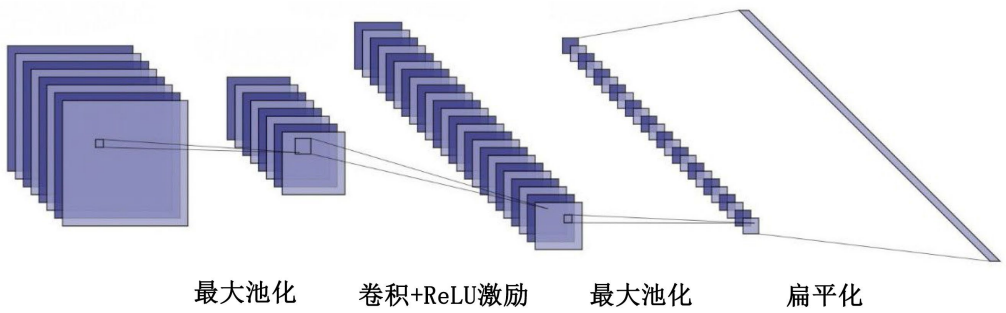

Figure 2. Convolution schematic diagram
图 2. 卷积示意图

该模型在训练集与测试集中的训练误差达 0.0000，表明提取的数据集较为可信。

Step 3：接下来对单个事件文件夹内的多张时序图像，分别提取 2048 维特征向量，然后计算其均值，得到一个能综合表征该事件整体视觉信息的特征向量。为提取最主要体征，采用 PCA 算法：先对所有事件的 2048 维图像特征进行标准化后，执行主成分分析。根据 PCA 负载矩阵的绝对值和排序，筛选出最重要的 5 个主成分特征，极大降低了数据维度与后续模型的复杂度。并将降维后的 5 维图像特征与原始的 19 维文本物理特征直接拼接，形成最终的 24 维融合特征数据集，用于后续的预报模型训练。

4. 预报模型构建

4.1. 模型构建

4.1.1. XGBoost

作为一种典型的集成模型，XGBoost 采用提升算法对多个单一模型进行集成，使算法具有强大的性能。这些集成的单一模型也称为基础估计器。集成算法的核心思想是利用基础估计器完成对数据集的初步学习，根据学习结果调整每个样本事件的权重。然后由下一个估计器对权重调整后的数据集进行重新训练。最后，将不同基础估计器的误差进行加权平均，作为集成模型的最终输出结果。常用的基础估计器是决策树，这是一种基于树结构模型的机器学习算法，由树节点和分支组成。数据集从根节点输入，分支代表不同的条件，负责将数据记录划分到不同的子空间/子节点，并递归构建子节点，直到满足停止条件或每个节点中的样本事件无法再分割，最终得到决策树模型。决策树模型的训练步骤可分为两步：决策树的迭代生成和决策树的剪枝。本文采用分类回归树(CART)，通过 CART 算法构建严格的二叉树，并且针对分类问题和回归问题采用不同的分裂准则。

大多数其他算法利用一阶导数拟合损失函数，而 XGBoost 考虑泰勒展开式的二阶导数以找到更准确的梯度下降方向，并添加正则项以防止模型过拟合。它还基于特征预排序实现了每棵树的并行训练，最终给出近似算法以进一步优化效率。通过采用加法模型，当 XGBoost 不断学习第 m 个基础估计器时，优

化目标会添加基于叶节点数量和权重的正则项，以控制每个基础估计器的复杂度，从而防止模型过拟合。XGBoost 针对第 m 个基础估计器的学习目标函数如下：

$$\underset{f}{\operatorname{argmin}}\sum_{i=1}^{N}L\left(y_i,f_m\left(x_i\right)\right)+\sum_{i=1}^{N}\Omega\left(T\left(x,\theta_i\right)\right) \tag{1}$$

其中，L 为损失函数，$\Omega\left(T\left(x,\theta_m\right)\right)=\gamma\left|T_m\right|+\frac{1}{2}\lambda\sum_{t=1}^{\left|T_m\right|}w_t^2$ 为正则项。表示第 m 个树模型生成的叶节点数量，w_t 为叶节点的输出值，γ 和 λ 为相应的正则化系数。

通过二阶泰勒展开近似损失函数后，学习目标函数变为：

$$\begin{aligned}L^{(m)}&=\sum_{i=1}^{N}L\left(y_i,f_{m-1}\left(x_i\right)+T\left(x,\theta_m\right)\right)+\Omega_m\\&\approx\sum_{i=1}^{N}L\left(y_i,f_{m-1}\left(x_i\right)\right)+L'_{m-1}T\left(x,\theta_m\right)\\&\quad+\frac{1}{2}L''_{m-1}T^2\left(x,\theta_m\right)+\gamma\left|T_m\right|+\frac{1}{2}\sum_{i=1}^{\left|T_m\right|}w_t^2\end{aligned} \tag{2}$$

合并上式中的两项和，得到：

$$L^{(m)}=\sum_{t=1}^{\left|T_m\right|}\left[\left(\sum_{i\in I_t}L'_{m-1}\left(x_i\right)\right)w_t+\frac{1}{2}\left(\sum_{i\in I_t}L''_{m-1}\left(x_i\right)+\lambda\right)w_t^2\right]+\gamma\left|T_m\right| \tag{3}$$

这是叶节点权重的二次函数。当且仅当 $w_i=-\dfrac{\sum_{i\in I_i}L'_{m-1}\left(x_i\right)}{\sum_{i\in I_i}L''_{m-1}\left(x_i\right)+\lambda}$ 时，目标函数取得最小值，如下所示：

$$\tilde{L}^{(m)}=-\frac{1}{2}\sum_{t=1}\left|T_m\right|\frac{\left(\sum_{i\in I_t}L'_{m-1}\left(x_i\right)\right)^2}{\sum_{i\in I_t}L''_{m-1}\left(x_i\right)+\lambda}+\gamma\left|T_m\right| \tag{4}$$

XGBoost 模型将上式作为树结构分裂的评分函数，取代了基尼系数，从而改变了学习器的分裂方式。通过这些操作，XGBoost 更新了基础学习器的学习目标和分裂方式，并在工程实现上进行了大幅优化。为防止过拟合，XGBoost 引入了学习率作为正则化手段。同时，它在训练前对特征进行预排序并存入块结构，实现了最优特征选择的并行化，提升训练效率。此外，XGBoost 还能利用特征排序后的分位数作为分割点，采用近似算法进而加快训练。

XGBoost 模型在选择基础估计器和调整参数方面具有更高的灵活性。在模型调优中，可以调整以下超参数：基础估计器数量(n-estimators)、学习率(Learning-rate)、基础学习器的最大深度(Max_depth)、子采样率(Subsample)、列子采样率(Colsample-bytree)、节点最小权重(Min_child_weight)、最小分裂增益(Gamma)以及正则化系数(Reg-alpha 和 Reg-lambda)。

4.1.2. 随机森林

作为一种经典的集成模型，随机森林(Random Forest, RF)采用 Bagging 算法对多棵决策树进行集成，使模型具备优异的预测性能与鲁棒性。随机森林的核心思想是通过双重随机化机制构建具有强多样性的基础估计器集合，再通过集成融合策略整合各基础估计器的预测结果，最终实现模型性能的提升与过拟合风险的有效抑制。

同 XGBoost 一样，随机森林常用的基础估计器是决策树，并同样采用分类回归树(CART)，通过 CART 算法构建严格二叉树，针对回归任务采用均方误差作为节点分裂准则，针对分类任务采用基尼系数作为节点分裂准则。基尼系数的计算公式为：

$$Gini(D)=1-\sum_{k=1}^{K}p_k^2 \tag{5}$$

其中，D 为当前节点的样本集合，p_k 为该节点中第 k 类样本所占的比例。基尼系数越小，表明节点样本的类别分布越集中(纯度越高)；反之则表明样本类别越混杂(纯度越低)。

在节点分裂过程中，随机森林通过计算某一特征在某一分裂点下，左右子节点的基尼系数加权和，进而得到该分裂方式的基尼增益。最终，随机森林会选择基尼增益最大的特征–分裂点组合，完成当前节点的分裂，并递归执行该过程直至满足停止条件(如节点样本数达到最小值、树深度达到上限等)。与单一决策树易过拟合的特性不同，随机森林通过样本随机化与特征随机化的双重正则化设计，有效弥补了单一决策树的缺陷。样本随机化通过 Bootstrap 有放回抽样实现，为每一棵基础决策树生成独立的训练子数据集，使各决策树在不同的数据分布上完成学习；特征随机化则在决策树节点分裂时，随机选取部分特征子集进行最优分裂点筛选，打破强相关特征对分裂结果的主导性，进一步增强基础估计器之间的多样性。

随机森林采用加法平均的集成融合策略，当构建完成 T 棵基础决策树后，对于回归任务，将所有决策树的预测结果进行算术平均，作为集成模型的最终输出；对于分类任务，采用多数投票法确定最终预测类别。此策略保留决策树对非线性关系的强拟合能力，使随机森林在复杂任务中展现出优于单一决策树的泛化性能。随机森林的学习目标是最小化集成模型的经验损失与基学习器复杂度正则项之和，即：

$$\underset{\{h_t\}_{t=1}^{T}}{\operatorname{argmin}}\sum_{i=1}^{N}L\left(y_i,\frac{1}{T}\sum_{t=1}^{T}h_t(x_i)\right)+\sum_{t=1}^{T}\Omega(h_t) \tag{6}$$

对于日冕物质抛射传播时间的回归预测任务，随机森林的最终预测模型采用算术平均融合规则，将所有基学习器的预测结果进行平均，数学表达式为：

$$H(x)=\frac{1}{T}\sum_{t=1}^{T}h_t(x;\Theta_t) \tag{7}$$

其中 $h_t(x;\Theta_t)$ 表示第 t 棵基础决策树的预测函数，Θ_t 为其对应的模型参数。这种融合方式能够有效抵消单棵决策树的随机预测误差，同时保留决策树对非线性关系的强拟合能力，使模型在处理日冕物质抛射传播过程中复杂的物理参数关联时表现出良好的适应性。

为量化基学习器多样性对集成性能的影响，可对总损失进行代数分解：

$$\begin{aligned}H(x_i)&=\frac{1}{T}\sum_{t=1}^{T}h_t(x_i)L_{\text{total}}=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\frac{1}{T}\sum_{t=1}^{T}h_t(x_i)\right)^2+\frac{1}{T}\sum_{t=1}^{T}\Omega_t(h_t)L_{\text{total}}\\&=\frac{1}{T}\sum_{t=1}^{T}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-h_t(x_i)\right)^2}_{L_t}-\frac{1}{T^2}\sum_{t\neq t'}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-h_t(x_i)\right)\left(y_i-h_{t'}(x_i)\right)}_{Cov(h_t,h_t')}\end{aligned} \tag{8}$$

其中，L_t 为单棵决策树的预测损失，$Cov(h_t,h_t')$ 为不同决策树预测结果的协方差。该式表明，随机森林的总损失由单棵树的平均损失与树间协方差共同决定，通过增强基学习器多样性降低协方差，可进一步优化模型性能。

随机森林的性能可通过超参数调优进一步优化，核心可调超参数包括：基学习器数量(n_estimators)、最大树深度(max_depth)、最小叶节点样本数(min_samples_leaf)、节点分裂的最小不纯度减少量

(min_impurity_decrease)、随机特征子集大小(max_features)以及正则化系数(ccp_alpha)。

4.1.3. 堆叠集成学习

堆叠集成学习(Stacking，又称堆叠泛化)是高阶集成学习范式，核心通过分层训练融合异质基学习器预测能力，突破单一模型性能上限。与 Bagging (如随机森林)的平行集成不同，其采用基学习器层与元学习器层，能充分挖掘模型预测互补性，在日冕物质抛射传播时间这类复杂非线性回归任务中，具备更优预测精度与鲁棒性。本文构建的堆叠集成模型，之所以选 XGBoost 与随机森林作为异质基学习器，以线性回归为元学习器，核心源于二者算法特性与任务适配性的高度互补。

从算法原理来看，XGBoost 作为梯度提升树改进版，串行迭代构建决策树，以修正前序模型残差为目标，结合正则化与二阶泰勒展开，拟合能力更强，擅长捕捉局部细微规律；随机森林则基于决策树 Bagging 集成，通过 Bootstrap 重采样构建多棵独立决策树，以投票或平均输出结果，优势是抗噪性强、不易过拟合，能高效捕捉数据全局模式。二者误差来源亦不同，随机森林误差偏方差，XGBoost 误差偏偏差，独立性较强，为元学习器融合奠定基础。

针对日冕物质抛射传播时间预测的任务特性，输入特征兼具太阳物理参数高维相关性与特征噪声干扰。随机森林可保留特征交互信息并规避过拟合，适配高维数据全局模式；XGBoost 对局部细微规律敏感，在小样本区间预测更精准。二者结合能全方位建模复杂非线性关系，弥补单一模型短板。线性回归元学习器则通过最小二乘法最优加权，融合二者优势，最终显著提升模型泛化能力与预测精度。

1) 基学习器层：选取 XGBoost 与随机森林两种具有互补性的基学习器，两类模型分别基于不同的集成思想构建，对日冕物质抛射传播时间的特征捕捉能力各有侧重，能够有效提升基预测结果的多样性。

对于第 k 个基学习器，首先使用训练集 D_{train} 进行训练，得到拟合后的模型 $M_k^*(x)$。

利用训练好的基模型 $M_k^*(x)$，分别对训练集 X_{train} 与测试集 X_{test} 进行预测，得到对应的预测结果：

$$\hat{y}_{train,k} = M_k^*\left(X_{train}\right), \hat{y}_{test,k} = M_k^*\left(X_{test}\right) \tag{9}$$

其中 $\hat{y}_{train,k} \in \mathbb{R}^{n_{train}\times 1}$， $\hat{y}_{test,k} \in \mathbb{R}^{n_{test}\times 1}$。

将所有 K 个基模型的训练集预测结果按列拼接，形成训练元特征矩阵 Z_{train}；将所有 K 个基模型的测试集预测结果按列拼接，形成测试元特征矩阵 Z_{test}，即：

$$\begin{aligned} Z_{train} &= \left[\hat{y}_{train,1}, \hat{y}_{train,2}, \cdots, \hat{y}_{train,K}\right] \in \mathbb{R}^{n_{train}\times K} \\ Z_{test} &= \left[\hat{y}_{test,1}, \hat{y}_{test,2}, \cdots, \hat{y}_{test,K}\right] \in \mathbb{R}^{n_{test}\times K} \end{aligned} \tag{10}$$

元特征矩阵的每一列对应一个基模型的预测结果，每一行对应一个样本的元特征向量，实现了从原始特征到高阶元特征的转换。

2) 元学习器层：对训练集进行学习后，生成基模型预测结果，将其作为元特征输入至第二层元学习器。元学习器层则选取线性回归作为元学习器，以真实日冕物质抛射传播时间标签为目标，通过线性拟合学习各基模型预测结果的最优加权组合，实现预测性能的进一步提升。此外，元学习器以训练元特征矩阵 Z_{train} 为输入，以真实标签 y_{train} 为目标进行训练，本文选用线性回归作为元学习器，其数学模型为：

$$G(z) = \omega_0 + \sum_{k=1}^{K} \omega_k z_k \tag{11}$$

其中，$z = [z_1, z_2, \cdots, z_K]^{\mathrm{T}}$ 为元特征向量，ω_0 为偏置项 $\{\omega_k\}_{k=1}^{K}$ 为元特征的权重系数，对应线性回归的模型参数。

元学习器的训练目标是最小化训练元特征上的均方损失，即：

$$\underset{\omega_0,\{\omega_k\}_{k=1}^{k}}{\operatorname{argmin}} \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \left(y_{train,i} - G\left(z_{train,i} \right) \right)^2 \tag{12}$$

其中，$z_{train,i}$ 为训练元特征矩阵 Z_{train} 的第 i 行(第 i 个样本的元特征向量) $y_{train,i}$ 为对应的真实标签。训练完成后，得到最优元学习器 $G^*(z)$，将测试元特征矩阵 Z_{test} 输入至 $G^*(z)$，即可得到堆叠集成学习模型的最终预测结果：

$$\hat{y}_{final,test} = G^*\left(Z_{test} \right) \tag{13}$$

综合两层架构的数学关系，堆叠集成学习模型的整体预测表达式为：

$$\hat{y}_{final} = G^*\left(\left[M_1^*(x), M_2^*(x), \cdots, M_K^*(x) \right]^{\mathrm{T}} \right) \tag{14}$$

在本文场景下，基学习器选取 XGBoost+随机森林、元学习器为线性回归时，代入可得：

$$\hat{y}_{final} = \omega_0^* + \omega_1^* M_{XGB}^*(x) + \omega_2^* M_{RF}^*(x) \tag{15}$$

其中，ω_0^*、ω_1^*、ω_2^* 为线性回归元学习器的最优参数，$M_{XGB}^*(x)$、$M_{RF}^*(x)$ 分别为训练完成的 XGBoost 与随机森林基模型的预测函数。

4.2. 模型评价指标

在后续部分，选择四个常用的评价指标来评估模型的预测能力，分别是：(1) 决定系数 R^2；(2) 均方根误差(RMSE)；(3) 平均绝对误差(MAE)；(4) 探测概率(POD)。其公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{17}$$

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{18}$$

$$\mathrm{POD} = \frac{\mathrm{Hits}}{\mathrm{Hits} + \mathrm{Misses}} \tag{19}$$

其中，y_i 和 $\hat{y}_i$ 分别为实际值和预测值，$\bar{y}$ 为实际值的平均值。在式(19)中，Hits 和 Misses 分别表示误差小于和大于平均绝对误差的事件数量。此外，补充定义平均绝对误差的均值、最小平均绝对误差、平均决定系数与最大决定系数，以期更准确的刻画训练过程。

$$\text{average MAE}_n = \frac{1}{n} \sum_{k=1}^{n} \mathrm{MAE}_k \tag{20}$$

$$\min \mathrm{MAE}_n = \underset{1 \le k \le n}{\operatorname{argmin}} \mathrm{MAE}_k \tag{21}$$

$$\text{average } R_n^2 = \frac{1}{n} \sum_{k=1}^{n} R_k^2 \tag{22}$$

$$\max R_n^2 = \underset{1 \le k \le n}{\operatorname{argmax}} R_k^2 \tag{23}$$

其中，MAE_k 和 R_k^2 分别表示第 k 次训练获得的平均绝对误差和决定系数。

5. 实验结果与分析

从描述性统计分析表中可以看出，不同特征的数值大小差异显著，这极大地增加了优化步长选择和迭代收敛的难度。因此，对数据集进行标准化处理，公式如下：

$$X' = \frac{X - \bar{X}}{\sigma_X} \tag{24}$$

其中，$\bar{X}$ 为每个特征的平均值，σ_X 为标准差。为避免每次测试的偶然性，进行了多次训练。本文尝试了诸多比例，发现 8:2 是本模型的最佳分割方式。故数据集按此比例随机分割，每次训练相互独立，即每次训练开始前重新初始化模型。

由于不同特征的数值大小差异显著，这极大地增加了优化步长选择和迭代收敛的难度。因此，本文选择对数据集进行标准化处理。且尝试了诸多比例，发现 8:2 是本模型的最佳分割方式。故数据集按此比例随机分割，每次训练相互独立，即每次训练开始前重新初始化模型。

模型性能对比

由于样本量相对较小，未预先进行特征选择，而是将所有特征输入模型进行训练，并得到了如表 2 的训练结果：

Table 2. Model training results
表 2. 模型训练结果

模型选取	指标	训练集	测试集
XGBoost	MAE	7.6797	8.6153
	R^2	0.7595	0.7048
	MSE	99.5276	113.1710
	RMSE	9.9763	10.6382
	POD	60.10%	54.90%
随机森林	MAE	8.4143	8.7765
	R^2	0.7407	0.6910
	RMSE	10.3601	10.8831
	MSE	107.3323	118.4424
	POD	56.65%	52.94%
XGboost 与随机森林集成学习	MAE	7.0511	7.8048
	R^2	0.8079	0.7613
	MSE	79.5222	91.4911
	RMSE	8.9175	9.5651
	POD	61.08%	52.94%

基于模型训练结果，本文得到如下结论：

1) XGBoost 模型在测试集的 MAE 为 8.6153，R^2 为 0.7048，体现了较好的拟合能力。随机森林模型测试集 MAE 为 8.7765，R^2 为 0.6910；相比 XGBoost，其 MAE 略高、R^2 略低，整体预测精度稍逊于

XGBoost。XGBoost 与随机森林的集成学习模型，测试集 MAE 降至 7.8048，是所有模型中 MAE 最低的；同时测试集 R^2 达到 0.7613，既优化了误差指标，也保持了较好的拟合优度，展现了集成方法的互补优势。

2) 训练集的 MAE 普遍略低于测试集、R^2 则略高于测试集，但指标差异幅度较小。本文认为，这一现象与数据集规模有限密切相关：由于数据量不足，模型在学习核心有效规律的同时，不可避免地拟合了训练数据中的部分噪声成分，进而导致轻微过拟合。即便存在这一局限，模型仍保持了较强的解释力。

3) 从各项指标的综合表现看，集成学习模型展现了较为显著的互补增益：其测试集 MAE (7.8048)不仅低于 XGBoost (8.6153)和随机森林(8.7765)，POD 指标(52.94%)也与随机森林持平、仅略低于 XGBoost。这说明集成方法通过融合不同模型的优势，既进一步压缩了预测误差，又维持了稳定的性能表现，表明了堆叠集成学习有助于提升预测精度与鲁棒性。

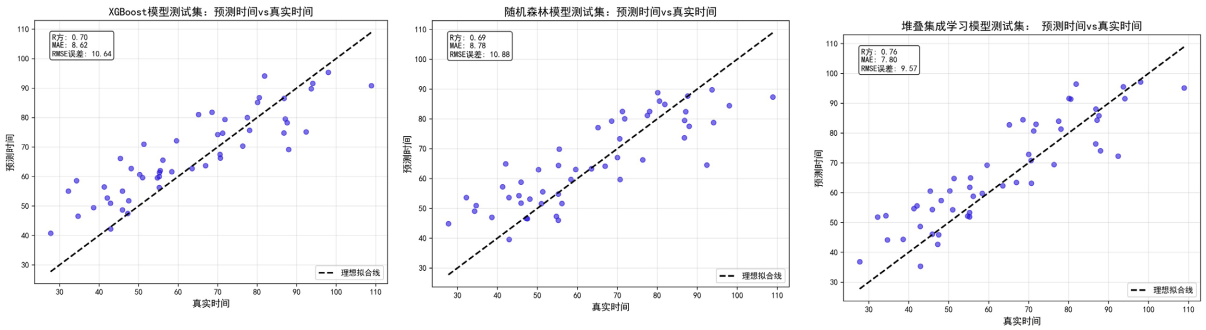

Figure 3. Comparison chart of predicted propagation time and actual propagation time of coronal mass ejections in model test sets
图 3. 模型测试集中日冕物质抛射的预测传播时间与实际传播时间对比图

图 3 为三种模型测试集中日冕物质抛射的预测传播时间与实际传播时间对比图，横轴代表日冕物质抛射实际传播时间，纵轴代表模型预测传播时间(小时)，图中对角线为基准线，代表预测值等于实际值的理想线。数据点分布显示，堆叠集成学习模型的散点更密集地贴近基准线，偏离程度最小，效果最优；XGBoost 模型散点分布次之，随机森林模型部分数据点偏离基准线较远。通过对比不同模型散点与基准线的贴合度，清晰展现了集成模型在预测精度上的优势，以及单一模型在极端值预测中的偏差。

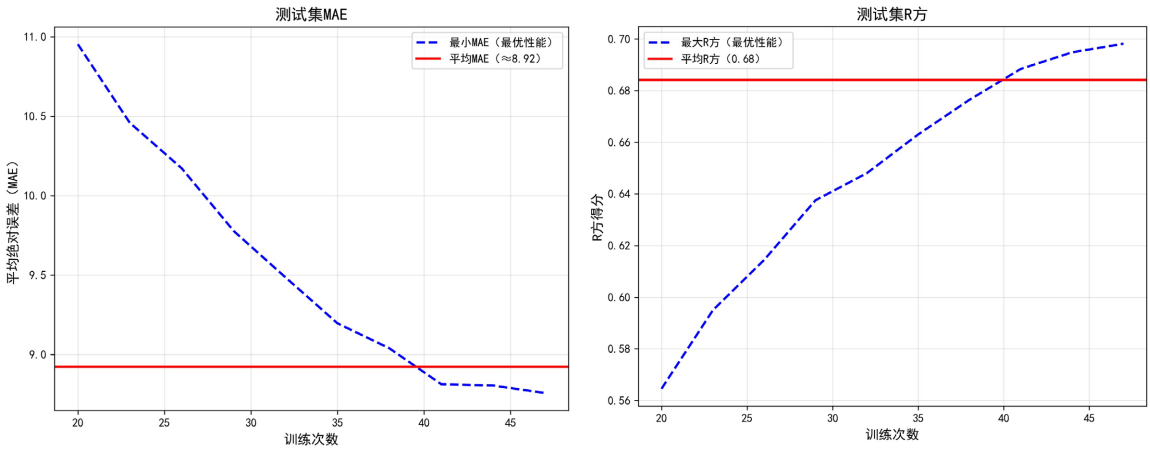

Figure 4. XGBoost training process
图 4. XGBoost 训练过程

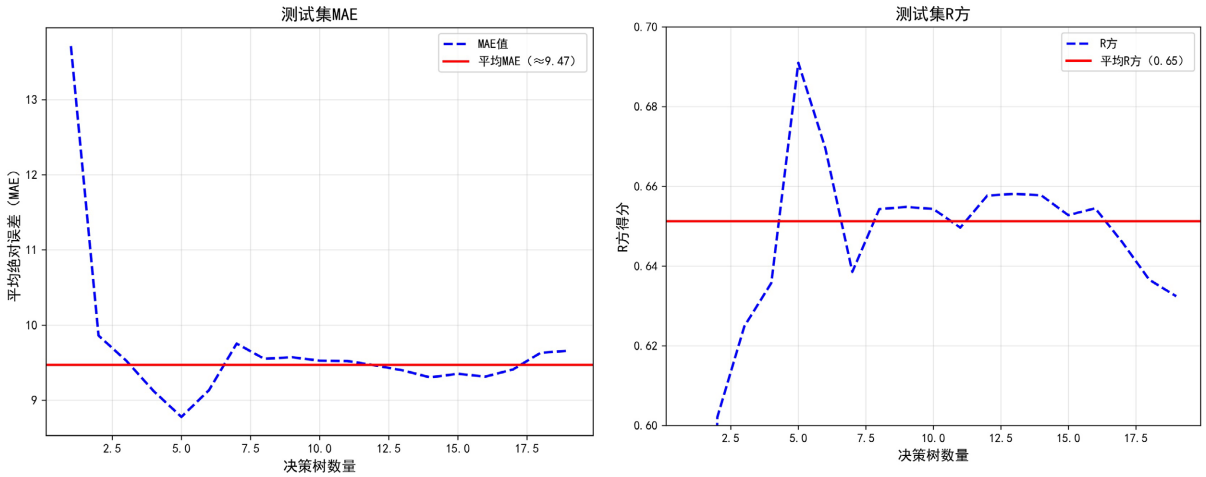

Figure 5. Random forest training process
图 5. 随机森林训练过程

图 4、图 5 以折线图形式呈现 XGBoost 与随机森林模型的训练动态过程。横轴为训练迭代次数，纵轴为模型训练评价指标(MAR 与 R^2)。曲线走势显示，随着迭代次数增加，两种模型的训练误差均逐步下降并趋于稳定，其中 XGBoost 模型的误差下降速率最快，最终稳定在最低水平；随机森林表现有一定波动，但测试集上最优的预报模型二者性能相近。

6. 结论与展望

本文针对现有研究未充分融合 LASCO 观测图像与文本物理参数的空白，提出了一套完整的数据融合方案与日冕物质抛射传播时间预报模型，取得了显著成果。通过时间戳匹配实现 255 个 CME 事件的时序图像与 19 维文本物理特征的有效关联，借助迁移学习的 ResNet50 模型提取图像深度特征，经 PCA 降维得到 5 维图像特征，与文本特征拼接形成 24 维融合数据集。对比 XGBoost、随机森林及二者堆叠集成学习模型的实验结果表明，堆叠集成学习模型性能最优，测试集平均绝对误差(MAE)达 7.8048 小时，决定系数(R^2)为 0.7613，显著优于单一模型，充分验证了数据融合的有效性与集成学习的互补优势。需指出的是，所有模型均仅存在轻微过拟合，推测与数据集规模有限相关，整体模型可靠性较强。与物理模型的对比结果显示，本文提出的模型在预测精度与计算效率上表现良好，为 CME 传播时间预报提供了更精准、高效的技术方案。

未来研究可从以下四大方向进一步深化与拓展，推动空间天气中日冕物质抛射相关预测技术的迭代升级：

其一，扩充数据来源，提升数据质量与多样性。当前数据集在样本规模、覆盖范围和事件类型上存在局限，影响了模型的泛化能力。可补充事件从爆发到近地响应的完整演化数据。当前模型受限于数据量有限，保留了所有的有效数据。因而在有一定数据量后，当在数据预处理阶段采取措施，通过噪声过滤、异常值检测等方法减少观测误差与数据缺失的影响。

其二，深化文本与图像数据的融合。目前图像与文本数据的融合仍以浅层拼接为主，未能充分提取其内在的语义与因果联系。下一步可引入注意力机制与其他的深度学习模型，构建更精细的跨模态融合框架。通过交叉注意力机制，模型可自主聚焦于关键的图像区域(与文本特征。

其三，拓展预报参数，支撑空间天气综合预报。现有模型多集中于日冕物质抛射到达时间等单一指标的预测，尚未充分发挥多参数预报的潜力。未来可将框架扩展至日冕物质抛射到达地球时的多种关键

物理参数预报，如磁场强度、等离子体密度、温度等，这些参数直接影响空间天气对技术系统的影响评估。

其四，融合物理机制，构建可解释的混合预测模型。当前机器学习模型虽具有一定精度优势，缺乏物理过程的可解释性，限制了其在实际业务中的应用。未来需深入结合日冕物质抛射爆发与传播的物理机理，尝试将磁流体力学方程作为损失函数的约束项。通过将磁流体力学等理论作为模型训练的约束条件，确保预测结果符合物理规律；同时引入物理模型的中间变量作为特征输入，增强模型对日冕物质抛射演化的捕捉能力。此类模型既可保持机器学习的高效拟合能力，又具备物理模型的可靠性与可解释性，有较为广阔的应用前景。

参考文献

[1] Odstrcil, D. (2003) Modeling 3-D Solar Wind Structure. *Advances in Space Research*, **32**, 497-506. https://doi.org/10.1016/s0273-1177(03)00332-6

[2] Yang, Y. and Shen, F. (2019) Modeling the Global Distribution of Solar Wind Parameters on the Source Surface Using Multiple Observations and the Artificial Neural Network Technique. *Solar Physics*, **294**, Article No. 111. https://doi.org/10.1007/s11207-019-1496-5

[3] Huang, X., Wang, H., Xu, L., Liu, J., Li, R. and Dai, X. (2018) Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-Of-Sight Magnetograms. *The Astrophysical Journal*, **856**, Article No. 7. https://doi.org/10.3847/1538-4357/aaae00

[4] Liu, J., Ye, Y., Shen, C., Wang, Y. and Erdélyi, R. (2018) A New Tool for CME Arrival Time Prediction Using Machine Learning Algorithms: Cat-puma. *The Astrophysical Journal*, **855**, Article No. 109. https://doi.org/10.3847/1538-4357/aaae69

[5] Wang, Y., Liu, J., Jiang, Y. and Erdélyi, R. (2019) CME Arrival Time Prediction Using Convolutional Neural Network. *The Astrophysical Journal*, **881**, Article No. 15. https://doi.org/10.3847/1538-4357/ab2b3e

[6] Zhao, X. and Dryer, M. (2014) Current Status of CME/Shock Arrival Time Prediction. *Space Weather*, **12**, 448-469. https://doi.org/10.1002/2014sw001060