arXiv:2305.00258v1 [astro-ph.SR] 29 Apr 2023

# 用于日冕物质抛射到达时间预测的集成学习

**Khalid A. Alobaid** [1] **Jason T. L. Wang** [2]


## 摘要

太阳不断向日球层释放辐射和等离子体。太阳偶尔会发生耀斑和日冕物质抛射（CME）等太阳爆发。日冕物质抛射会携带大量质量和磁通量离开太阳。朝向地球的日冕物质抛射可能对人类社会造成严重后果，破坏电网、管道、卫星和通信系统。因此，准确监测和预测日冕物质抛射对于最大限度减少其对人类社会造成的损害至关重要。本研究提出了一种名为 CMETNet 的集成学习方法，用于预测日冕物质抛射从太阳到达地球的时间。我们收集并整合了1996年至2021年两个太阳活动周（第23太阳活动周和第24太阳活动周）的太阳爆发事件，共计363次地球有效日冕物质抛射。用于预测的数据包括日冕物质抛射特征、太阳风参数，以及通过SOHO/LASCO C2日冕仪获取的日冕物质抛射图像。我们的集成学习框架包含用于数值数据分析的回归算法，以及用于图像处理的卷积神经网络。实验结果表明，CMETNet 的表现优于文献中报道的现有机器学习方法，其皮尔逊积矩相关系数为0.83，平均绝对误差为9.75小时。


## 1. 引言

欧洲空间局（ESA）和美国国家航空航天局发射太阳与日球层探测器（SOHO），使科学家得以拍摄日冕物质抛射（日冕物质抛射；Gopalswamy，2016年）。日冕物质抛射是太阳系中发生的最剧烈、能量最强的现象（Schwenn et al.，2005年；Maloney et al.，2010年；Maloney，2012年）。它们是太阳等离子体的涌流，同时伴有电磁辐射，在能量爆发中从太阳光球喷发，并扩散至行星际环境。日冕物质抛射有三个显著特征：致密的中心、明显的前缘，以及电子密度较低的空腔。日冕仪图像通常显示日冕物质抛射明亮而致密的边缘（Gopal-swamy etal.，2009年）。图1 展示了SOHO观测到的2001年4月10日发生的一次日冕物质抛射的时间演化过程。

日冕物质抛射通常伴随着太阳耀斑（Priest & Forbes, 2002年；Daglis 等, 2004年；Chen 等, 2019年；Liu 等, 2019年；Jiao 等, 2020年；Wang 等人, 2020年；Abdual-lah 等, 2021年；Sun 等, 2022年），尽管日冕物质抛射与耀斑之间的关系仍在积极研究中（Yashiro & Gopalswamy, 2008年；Kawabata 等, 2018年；Liu 等, 2020年；Raheem 等, 2021 年）。太阳表面的磁活动区每天都会发生各种变化，从而引发日冕物质抛射；日冕物质抛射随后穿过行星际环境，有时会在其中产生激波，并且是空间天气预报及其他相关研究的重要因素。任何对太阳相关异常现象的研究，包括但不限于这些激波、日冕物质抛射和耀斑，通常都需要有关太阳磁场、它们彼此之间的联系以及周围行星际介质的信息。因此，展示并理解复杂磁场及其随时间和空间的变化十分重要。太阳活动包含空间天气研究中的关键要素。这些研究在卫星任务管理、空间仪器的布置与维护以及地球上的气候预测流程中发挥着重要作用。此外，日冕物质抛射是近地环境中各种空间天气影响的来源，例如地磁暴。一次强烈的日冕物质抛射可能导致大范围、长期的经济和社会灾难。Showstack（2013年）讨论了空间天气带来的潜在风险。作者指出，如果如今发生一次类似卡林顿事件的地磁暴，可能会使多达4,000万美国人面临停电风险，而停电可能持续数天至数月，并造成巨大的社会和经济影响。

为减轻日冕物质抛射可能造成的损害，人们已经开发出大量方法来预测

[1]新泽西理工学院计算机科学系，美国新泽西州纽瓦克市，邮编07102 [2]新泽西理工学院计算学院，美国新泽西州纽瓦克市，邮编07102。通信作者：Jason T. L. Wang <wangj@njit.edu>。

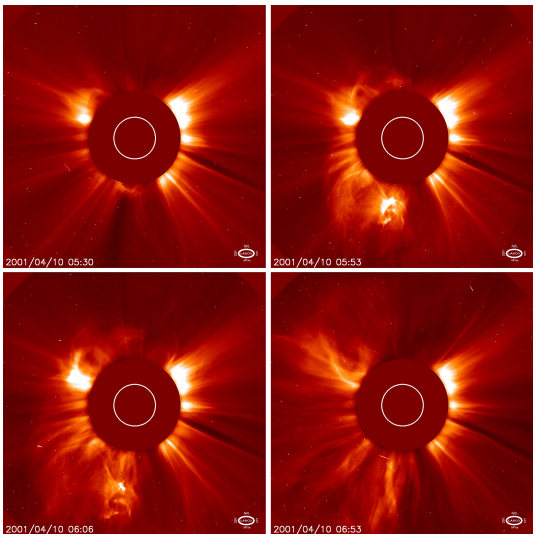



图*1*。2001年4月10日发生的一次日冕物质抛射的时间演变。图像取自 https://soho.nascom.nasa.gov。

日冕物质抛射的到达时间（Zhao 和 Dryer，2014年；Camporeale，2019年；Vourlidas 等，2019年）。其中，两类方法，即基于物理学的方法和基于机器学习的方法，得到了广泛应用。WSA-ENLIL+Cone（WEC）模型（Odstrcil 等，2004年；Riley 等，2018年）是用于预测日冕物质抛射到达时间的一种常用的基于物理学的方法（Riley 等，2018年）。WEC 模型已由多个机构实施并投入使用，其中包括美国国家海洋和大气管理局（NOAA）下属的空间天气预报中心（SWPC），以及美国国家航空航天局戈达德太空飞行中心（GSFC）下属的社区协调建模中心（CCMC）。WEC 模型利用从日冕物质抛射传播过程中获得的参数构建，包括日冕物质抛射速度、密度、位置和传播方向。

在机器学习方法中，Liu 等(2018年) 使用支持向量回归（SVR）算法，估算了1996年至2015年间182次地球有效日冕物质抛射的到达时间。对于每次日冕物质抛射，作者都考虑了相关的太阳风参数，以及从搭载于 SOHO 上的大视场和光谱日冕仪（LASCO）的日冕物质抛射目录中获取的日冕物质抛射特征。 Wang 等人(2019年) 设计了一种卷积神经网络，用于预测日冕物质抛射到达时间。作者使用了1997年至2017年间223次地球有效日冕物质抛射的观测图像，这些图像采集自 SOHO/LASCO。图像总数为1,122张，每次日冕物质抛射对应1至10张图像。

在本研究中，我们重点使用机器学习来预测日冕物质抛射传播时间，即日冕物质抛射主体从太阳行进至地球、在行星际空间中所耗费的时间。到达时间可通过将日冕物质抛射传播时间加到日冕物质抛射的起始时间上计算得出。我们收集并整合了1996年至2021年间两个太阳活动周（第23太阳活动周和第24太阳活动周）的爆发事件，共计363次地球有效日冕物质抛射。用于预测的数据包括日冕物质抛射特征、太阳风参数，以及从 SOHO/LASCO C2日冕仪获取的日冕物质抛射图像。我们提出的方法名为CMETNet，基于集成学习框架，其中包括用于数值/表格数据分析的回归算法，以及用于图像处理的卷积神经网络。实验结果表明，CMETNet表现良好，皮尔逊积矩相关系数为0.83，平均绝对误差为9.75小时。

本文其余部分安排如下：第2节介绍本研究使用的数据。第3 节介绍 CMETNet 并说明其工作原理。第4节展示实验结果。第5节总结全文，并指出未来研究的一些方向。

## 2. 数据

### 2.1. 数据来源

遵循Liu 等(2018年) 的做法，我们采用了四个日冕物质抛射目录：（i）Richardson 和 Cane（RC）目录（Richardson&Cane，2010年），可从以下网址获取：
h
t
t
p://
www.srl.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm；（ii）由中国科学技术大学（USTC）维护的全晕日冕物质抛射目录（Shen等，2014年），可从以下网址获取：
http://space.ustc.edu.cn/dreams/fhcmes/index.php；（iii）乔治梅森大学（GMU）的日冕物质抛射/行星际日冕物质抛射目录（Hess和Zhang，2017年），可从以下网址获取：
http://solar.gmu.edu/heliophysics/index.php/GMU_日冕物质抛射/行星际日冕物质抛射_目录；（iv）由美国国家航空航天局的社区协调建模中心（CCMC）维护的日冕物质抛射记分板，可从以下网址获取：
https://kauai.ccmc.gsfc.nasa.gov/CMEscoreboard/。

我们按照Liu 等(2018年)中所述的流程，对四个目录中的日冕物质抛射进行了合并和清理。随后，我们将这四个目录中的日冕物质抛射与Paouris和Mavromichalaki (2017年)所发布目录中的日冕物质抛射合并。最终得到1996年至2021年的363次日冕物质抛射。对于这363次日冕物质抛射，我们收集了12项相关的日冕物质抛射特征和太阳风参数。此外，对于每次日冕物质抛射，我们都从
h
t
t
p
s://www.swpc.noaa.gov/products/lasco-coronagraph
提供的 SOHO/LASCO C2 日冕仪下载其图像，方法与Wang 等人(2019年)中的做法相同。下文将详细介绍我们的数据整合流程。

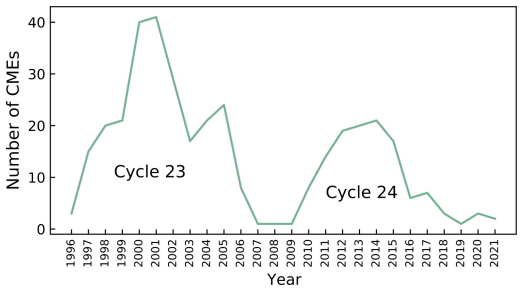



图2. 本研究所考虑的日冕物质抛射年度数量。这些日冕物质抛射发生于1996年8月至2021年5月。第23太阳活动周（1996年 - 2008年）的日冕物质抛射数量多于太阳活动周24（2008年 - 2019年末）。

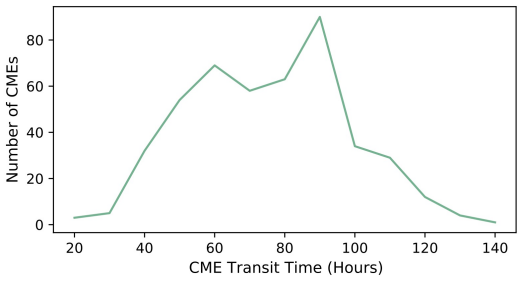



图3. 本研究所考虑的日冕物质抛射传播时间分布。对于这些日冕物质抛射，最短传播时间为18小时，最长为138小时。

## 2.2. 数据整合

对于每个日冕物质抛射列表（RC、USTC、GMU、CCMC），我们使用 Python 编写了一个跟踪程序，以自动收集列表中日冕物质抛射的起始/出现时间和到达时间。随后，我们将四个列表中的日冕物质抛射合并为一个数据集，删除重复项，并扫描数据集以查找发生在同一小时内的事件。对于发生在同一小时内的事件，我们只保留其中一个。根据Liu 等(2018年)中所述的标准，我们最终从 RC、USTC、GMU 和 CCMC 列表中分别得到216、24、38和113次日冕物质抛射。对于收集到的每次日冕物质抛射，我们记录了两个参数：该日冕物质抛射的起始/出现时间和到达时间。除了这四个日冕物质抛射列表（RC、USTC、GMU、CCMC），我们还考虑了Paouris和Mavromichalaki(2017年)发布的目录，其中包含266个日冕物质抛射事件。我们将这266个事件与 RC、USTC、GMU 和 CCMC 列表中的216、24、38和113次日冕物质抛射分别合并，并在数据集中对于发生在同一小时内的事件仅保留一个。

这一数据整合过程筛选出363次地球有效日冕物质抛射。图 2 展示了每年日冕物质抛射的数量。第23太阳活动周（1996年—2008年）有240次日冕物质抛射，而太阳活动周 24（2008年—2019年末）有118次。（2019年末至2021年间还有5次日冕物质抛射，因此本研究考察了1996年至2021年间共363次日冕物质抛射。）在第23太阳活动周中，2001年日冕物质抛射数量最多，共41次。在太阳活动周24中，2014年日冕物质抛射数量最多，共21次。2007年、2008年、2009年和2019年的日冕物质抛射数量最少，这四年每年仅有1次。值得一提的是，日冕物质抛射的峰值年份（即第23太阳活动周的2001年和太阳活动周24的2014年）与两个太阳活动周的太阳黑子峰值时间一致（Raheem 等人，2021年）。图3展示了数据集中363次事件的日冕物质抛射传播时间分布。对于这363次CME事件，最短传播时间为18小时，最长为138小时。

对于数据集中的每个事件，我们从 SOHO LASCO 日冕物质抛射目录获取其五项特征，该目录可通过以下网址访问：https://cdaw.gsfc.nasa.gov/CME_list/index.html，由美国国家航空航天局（NASA）戈达德航天飞行中心协调数据分析研讨会（CDAW）数据中心维护（Gopalswamy 等人，2009年）。这五项日冕物质抛射特征包括角宽度、主位置角（MPA）、线速度、最终高度处的二阶速度以及质量。具体而言，我们以每个事件的起始/出现时间为关键，从 LASCO 日冕物质抛射目录中提取其五项特征，并将这些特征与数据集中该事件的到达时间合并。对于数据集中一个 $E$ 起始/出现时间与 LASCO 日冕物质抛射目录中任何 CME 的起始/出现时间都不匹配的事件，我们选取该目录中时间上最接近且发生在 $E$ 一小时之内的匹配事件的五项特征，并将这五项特征赋给 $E$。在 LASCO 日冕物质抛射目录中，某项特征 $F$ 可能缺失，相应缺失值也会保留在数据集中。我们采用数据插补方法，计算数据集中该特征 $F$ 所有可用值的平均值，并以该平均值代替缺失值。

为获取数据集中每个事件的太阳风参数，我们采用了 Liu 等 (2018年所述的方法，并使用从美国国家航空航天局 OMNIWeb 提取的小时平均数据，网址为 https://omniweb.gsfc.nasa.gov/。我们考虑了七种太阳风参数：$B_x$， $B_z$、氦离子与质子比、流向经度、等离子体压力、流速和质子温度。对于数据集中的每个事件 $E$，我们将 $E$的起始/出现时间 $t$ 作为关键值，从 OMNIWeb 检索时间戳$t+6$处（小时平均分辨率）的太阳风参数值，并将检索到的太阳风参数值分配给事件 $E$。

除五项日冕物质抛射特征和七项太阳风参数外，我们还从 SOHO/LASCO C2 日冕仪下载了每个事件的 FITS 图像。363 个事件中，有 9 个事件的 FITS 图像无法找到。（对于这

9 个事件，在训练和测试 CMETNet 框架时，我们仅考虑其日冕物质抛射特征和太阳风参数。）我们采用 Wang 等人 (2019年) 所述的方法，将数据集中的每个事件与 LASCO 中相应的 FITS 文件进行匹配。具体而言，对于数据集中的每个事件，我们下载了事件起始/出现时间前 10 分钟至事件发生后 2 小时内的 FITS 文件。我们使用 Astropy 合作组 等人，2013年）提供的 Astropy Python 软件包读取每个 FITS 文件，并将其转换为二维（2D）数组形式的日冕物质抛射图像。数据集中图像的尺寸为： $1024 \times 1024$ 像素、 $512 \times 512$ 像素或 $256 \times 256$像素。为保持一致，我们将所有图像的尺寸缩小至$256 \times 256$ 像素。我们舍弃了质量较低的图像。最终共获得 2,281 张图像，每个事件对应 0 至 17 张图像。

最后，我们通过计算事件开始/出现时间与其到达时间之间相差的小时数，得出该 CME 事件的传播时间。在数据集中的 363 个地球有效日冕物质抛射事件中，其传播时间从 18 小时到约 5 天不等（见图 3）。换言之，日冕物质抛射从首次出现在 LASCO 中算起，在太空中从太阳传播到地球需要 18 小时至约 5 天。

## 3. 方法

### 3.1. 概述

我们采用集成学习方法来预测日冕物质抛射的传播时间。该方法使用一组机器学习模型，并将各模型的预测结果合并，得出最终预测（Dietterich，2000）。集成方法通常比构成该集成方法的单个机器学习模型更准确（Dietterich，2000）。构建集成模型有多种技术。本研究采用自助聚合，简称装袋法，该方法最初由 Breiman (1996年) 提出，其工作原理如下：

(1) 从给定的训练集中有放回地随机抽取 $N$ 个训练样本，该训练集包含 $M$ 个样本。(2) 重复步骤 (1)，生成 $L$ 个训练子集： $N_1$、 $N_2$、…、 $N_L$，其中 $L$ 为组成集成方法的学习器数量。请注意，同一个样本可能在训练子集中出现多次。(3) $L$ 个学习器中的每个学习器分别独立地使用一个训练子集进行训练，即 $N_1$、$N_2$、…、 $N_L$，且没有两个学习器使用同一个训练子集进行训练。

（4）每个训练好的学习器都会对给定的测试样本进行预测。集成方法通过组合方法生成测试样本的最终预测结果，通常是对训练好的学习器所做的 $L$ 预测取平均值。 $L$

我们的集成学习框架名为 CMETNet，如图 4 所示，由五个基学习器组成：(i) 支持向量回归（SVR）算法（Cortes & Vapnik, 1995）；(ii) 随机森林（RF）算法（Breiman, 2001）；(iii) 高斯过程（GP）算法（Yuan et al., 2008）；(iv) XGBoost（XGB）算法（Chen & Guestrin, 2016）；以及 (v) 卷积神经网络（CNN）（LeCun et al., 1999）。前四个学习器是回归算法，用于分析本研究涉及的日冕物质抛射特征和太阳风参数相关的数值型／表格数据。这四种回归算法常用于日球物理学和空间天气研究（Liu 等，2017年；Gruet 等，2018年；Inceoglu 等，2018年；Tang 等，2021年；Abduallah 等，2022年）。第五个学习器是用于日冕物质抛射图像处理的卷积神经网络模型。下文将介绍这五个基学习器以及组合这些学习器的集成方法。

### 3.2. 基学习器

我们的支持向量回归模型取自 Python 的 Sklearn 库（Pedregosa 等人，2011年）。我们使用 Sklearn 中的 GridSearchCV 函数进行参数调优，以找到支持向量回归模型的最优超参数。我们使用径向基函数（RBF）核，并将核缓存大小设为 200。我们的随机森林算法在训练期间构建多棵决策树，并在测试期间输出这些树所作预测的均值。我们使用 Sklearn 实现随机森林模型，并利用 Sklearn 的 RandomizedSearchCV 函数优化模型参数。树的数量设为 800；树中用于划分节点的特征数量设为特征总数的平方根（本研究所用特征共 12 个，包括五个日冕物质抛射特征和七个太阳风参数）。高斯过程是一种非参数模型，利用高斯过程的先验进行数据回归。它是一种概率方法，能够给出预测结果的置信度（G¨ortler 等人，2019年）。我们使用 Sklearn 实现了高斯过程模型，并采用两种核：Whitekernel 和 RBF。Whitekernel 的参数噪声设为 0.5。极端梯度提升是一种可扩展的树提升系统。我们利用 Chen 和 Guestrin (2016年) 提出的 XGBoost 包实现了极端梯度提升模型。我们将 XGBoost 包中的 XGBRegressor 类导入 CMETNet 框架，并使用默认参数。

图5展示了我们卷积神经网络模型的架构。该模型用于根据输入的 CME LASCO C2 图像预测日冕物质抛射传播时间。模型首先经过一个二维卷

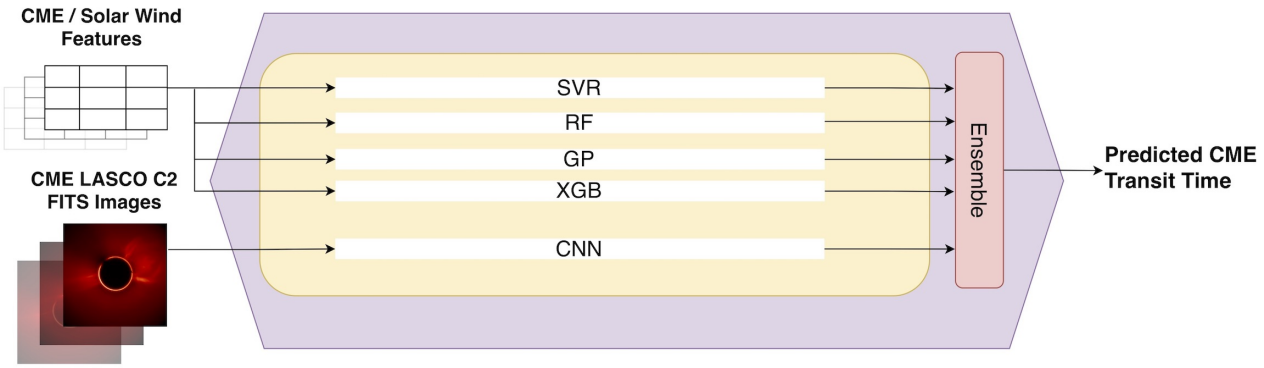



图 4. CMETNet框架示意图。该框架由五种基于机器学习的回归模型组成：支持向量回归、随机森林、高斯过程、极端梯度提升和卷积神经网络。对于CME事件 $E$，前四种模型（支持向量回归、随机森林、高斯过程、极端梯度提升）以 $E$的日冕物质抛射特征和太阳风参数作为输入，而第五种模型（卷积神经网络）以 $E$的日冕物质抛射图像作为输入。五种回归模型之后接入一种集成方法，该方法汇总五种模型的结果，并生成对 $E$传播时间的预测。

积层，包含 64 个大小为 $11 \times 11$ 且步幅为 1 的滤波器，后接带泄漏的修正线性单元（LeakyReLU）层。随后将 LeakyReLU 层的输出传入五个卷积块。每个卷积块都包含一个二维卷积层，其滤波器数量 F = 64 依次为 128、128、256、256，步幅 S = 2依次为 1、2、1、2，滤波器大小 K = $11 \times 11$。二维卷积层后接批归一化层，再接 LeakyReLU 层。最后一个卷积块输出的特征图被传入两个全连接层，分别含有 1024 个神经元和 1 个神经元。最后，模型通过线性激活函数输出传播时间。卷积神经网络模型使用平均绝对误差（MAE）损失函数作为回归损失函数（Berk, 1992）。训练卷积神经网络模型时，我们采用自适应矩估计（Adam）优化器（Goodfellow 等，2016年）来寻找模型的最优参数/权重。请注意，如图5所示，卷积神经网络模型根据单张 CME 图像预测传播时间。另一方面，我们的 CMETNet 框架接收与某个 CME 事件相关的所有图像，利用集成学习方法对 CMETNet 中卷积神经网络模型预测的多个传播时间进行处理并整合，如图4 所示，从而得出该 CME 事件最终的预测传播时间。下文第3.3 节将详细介绍集成学习方法。

### 3.3. 集成学习

我们将第3.1 节中介绍的原始自助聚合法应用于 CMETNet 框架。令 $T_r$ 表示训练集，令 $T_t$ 表示测试集。我们的数据集包含数值/表格数据（包括日冕物质抛射特征和太阳风参数），由回归算法处理；还包含日冕物质抛射图像，由卷积神经网络模型处理。我们分别介绍回归算法和卷积神经网络模型的训练流程。

对于回归算法，我们在训练过程中执行以下步骤。

（1）从 $T_r$中有放回地随机选取200个训练样本。每个样本对应一个CME事件。

（2）重复步骤（1），生成四个不同的训练子集，$N_1$、$N_2$、$N_3$、$N_4$。请注意，同一个样本可能在训练子集中出现多次。我们还获得了四个相应的验证集 $V_1$、$V_2$、$V_3$、$V_4$，其中 $V_i = T_r - N_i$、$1 \leq i \leq 4$。验证集中的每个样本也对应一个CME事件。

(3) 四种基学习器/回归算法——SVR、随机森林、高斯过程、极端梯度提升——分别由一个训练子集单独训练，$N_1$，$N_2$，$N_3$，$N_4$，且没有两个学习器使用同一个训练子集进行训练。每个训练好的学习器都会预测从相应验证集中抽取的每个样本的传播时间。相应验证集上的预测准确率以平均绝对误差（MAE）衡量（Berk，1992）。

(4) 将步骤 (1) 至 (3) 重复 100 次，为每个基学习器找出在相应验证集上预测准确率最高的最佳训练子集/模型。保存四种基学习器各自的最佳模型。

由于四种基学习器 SVR、随机森林、高斯过程和极端梯度提升仅适用于数值/表格数据，因此步骤 (1) 至

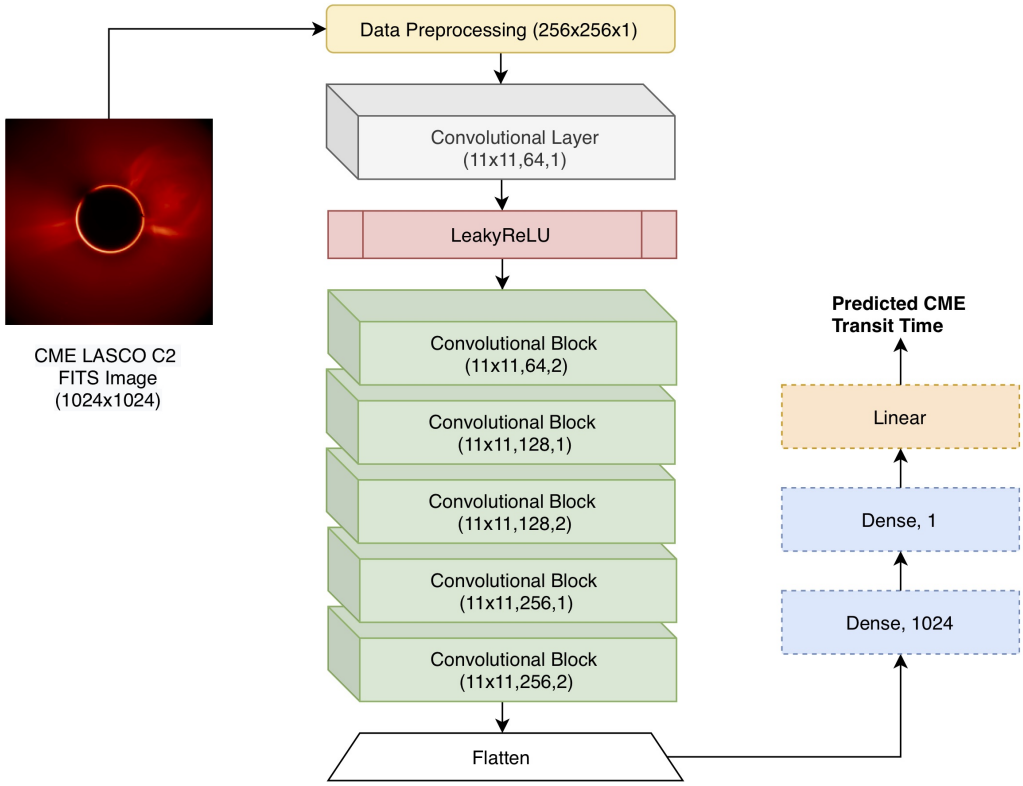



图 *5.* CMETNet 中卷积神经网络（CNN）模型的示意图，该模型用于根据输入的日冕物质抛射 LASCO C2 图像预测日冕物质抛射传播时间。(a) CNN 模型的整体架构。模型首先通过一个包含 64 个大小为 $11 \times 11$、步幅为 1 的滤波器的二维卷积层，随后接入带泄漏的修正线性单元（LeakyReLU）层。然后将 LeakyReLU 层的输出送入五个卷积块。最后一个卷积块输出的特征图将送入两个全连接层，分别包含 1024 个神经元和 1 个神经元。最后，模型使用线性激活函数输出预测的日冕物质抛射传播时间。(b) 卷积块的详细配置。卷积块包含一个具有 F 个滤波器的二维卷积层，其中 F 分别为 = 64、128、128、256 256 = 2 1 2 1 2 = $11 \times 11$ 2 ，步幅为 S，其中 S 分别为、、、，滤波器大小为 K。该二维卷积层之后依次接入批归一化层和 LeakyReLU 层。

上述（4）中，四种基学习器所用的特征是与每个样本/CME事件相关的日冕物质抛射特征和太阳风参数。

我们以略有不同的方式单独训练卷积神经网络模型。在这里，每个样本对应一张日冕物质抛射图像，而不是一个CME事件。一个CME事件可能有多张图像，且这些图像具有相同的日冕物质抛射传播时间。因此，我们将这些日冕物质抛射图像/样本标记为相同的日冕物质抛射传播时间。随后，我们在训练过程中执行以下步骤。

（1）从训练集 $T_r$ 中随机选取 10% 的样本/日冕物质抛射图像 $T_r$ 并将所选样本分配到验证集 $V$。

（2）使用 $T_r$ 中剩余的 90% 训练样本以及 $V$中的 10% 验证样本训练我们的卷积神经网络模型。验证集 $V$用于寻找最优参数和超参数，以获得最佳的卷积神经网络模型。

训练完成后，我们得到最佳的支持向量回归、随机森林、高斯过程、极端梯度提升和卷积神经网络模型。在测试阶段，我们根据以下两种情况预测测试集 $T_t$中每个测试CME事件 $E$ 的传播时间。

- $E$ 没有日冕物质抛射图像。在这种情况下， $E$ 只有数值/表格数据，包括日冕物质抛射特征和太阳风参数。然后， $E$ 的预测传播时间，是基于 $E$的日冕物质抛射特征和太阳风参数，分别由支持向量回归、随机森林、高斯过程和极端梯度提升这四种回归模型预测的传播时间的中位数。

- $E$ 拥有日冕物质抛射图像、日冕物质抛射特征和太阳风参数。在不失一般性的前提下，假设 $E$ 拥有$k$ 张日冕物质抛射图像。然后，对于每张日冕物质抛射图像，我们使用卷积神经网络模型预测该图像对应的传播时间 $t_1$。此外，我们基于 $E$的日冕物质抛射特征和太阳风参数，分别通过支持向量回归、随机森林、高斯过程和极端梯度提升这四种回归模型，得到四个传播时间 $t_2$，$t_3$，$t_4$，$t_5$。我们取五个预测传播时间 $t_1$，$t_2$，$t_3$，$t_4$，$t_5$的中位数。 $E$ 有 $k$ 张日冕物质抛射图像，因此我们得到 $k$ 个中位数。最后，我们取这 $k$ 个中位数的中位数，作为我们的集成方法（CMETNet）给出的 $E$的预测传播时间。

# 4. 实验与结果

## 4.1. 实验设置与评估指标

美国国家航空航天局社区协调建模中心（CCMC）维护的日冕物质抛射记分板于2013年上线。我们的数据集包含363次发生于1996年至2021年间的日冕物质抛射。为了比较我们的CMETNet所生成的预测与日冕物质抛射记分板上的预测，我们进行了如下的9折交叉验证实验。在每次实验中，将2013年至2021年间某一年的数据用作测试数据，其余所有数据则一并用作训练数据。2013年至2021年共9年，因此我们进行了9次实验／9折。在每次实验中，测试集与训练集互不重叠。我们计算了每次实验的预测准确率，并绘制了9次实验的均值和标准差。

我们采用两种评估指标来量化预测模型的性能。第一个指标是平均绝对误差（MAE；Berk，1992），定义如下：

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

此处 $N$ 表示测试集中的日冕物质抛射数量，而 $y_i$(ˆ$y_i$ 分别表示测试集中一次日冕物质抛射的实际传播时间（预测传播时间）。使用该指标，我们计算测试集中日冕物质抛射的实际传播时间与预测传播时间之间绝对误差的平均值。平均绝对误差越小，预测模型越准确，模型性能越好（Liu 等，2020年）。

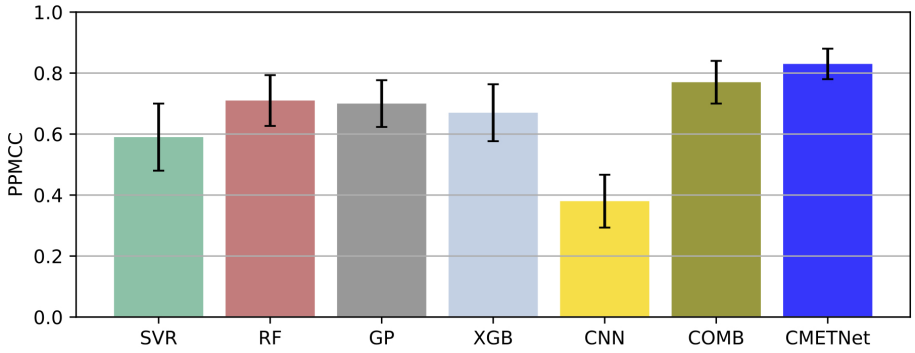
第二个评估指标是 Pearson 乘积矩相关系数（PPMCC；Pearson，1895），定义如下：

$$\mathrm{PPMCC} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \tag{2}$$

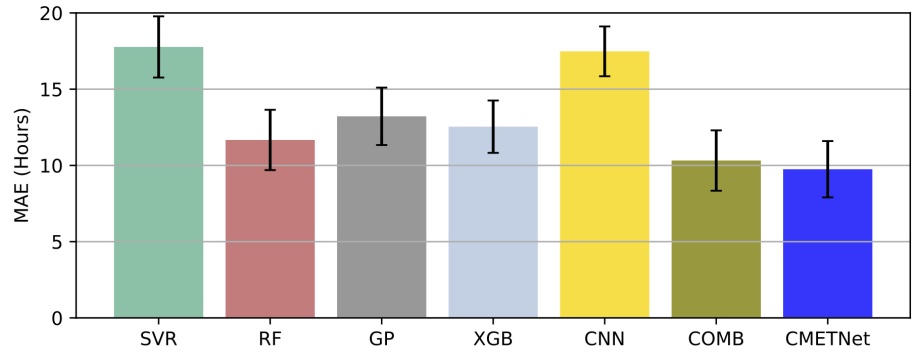
此处 $X$ 和 $Y$ 分别表示测试集中的实际传播时间和预测传播时间； $\mu_X$ 和 $\mu_Y$ 分别是 $X$ 和 $Y$的均值； $\sigma_X$ 和 $\sigma_Y$ 分别是 $X$ 和 $Y$ 的标准差； $E$（·）表示期望值。PPMCC 的取值范围为 −1 至 1。取值为 1 表示$X$ 与 $Y$ 之间存在线性关系，且 $Y$ 随 $X$ 增加而增加。取值为 −1 表示所有数据点都位于一条直线上，其中 $Y$ 随 $X$ 增加而减小。取值为零表示变量 $X$ 与 $Y$ 之间不存在线性相关（Liu 等，2020年）。

## 4.2. 消融研究

在本小节中，我们进行了消融测试，以分析和评估 CMETNet 框架的各个组成部分。我们考察了 CMETNet 的五个子单元：支持向量回归、随机森林、高斯过程、极端梯度提升和卷积神经网络。此外，我们还考察了 COMB，它是由支持向量回归、随机森林、高斯过程和极端梯度提升组成的集成模型。也就是说，COMB 是通过从 CMETNet 中移除卷积神经网络得到的。图 6展示了七种方法——支持向量回归、随机森林、高斯过程、极端梯度提升、卷积神经网络、COMB 和 CMETNet——的 PPMCC 和平均绝对误差结果。图中，每根彩色柱表示 9 次运行的均值，其对应的误差条表示标准




(a)




(b)

图 6. 用于评估 CMETNet 各组成部分（SVR、RF、GP、XGB、CNN）的消融测试结果，其中 COMB 是 SVR、RF、GP 和 XGB 的集成模型。（a）所测试模型的皮尔逊积矩相关系数（PPMCCs）。（b）所测试模型的平均绝对误差（MAEs）。CMETNet 在所有测试模型中表现最佳。

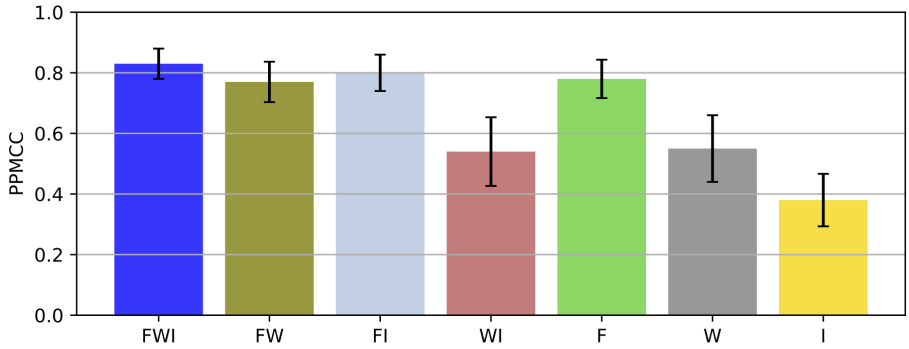
差除以 3（即运行次数的平方根）（Iong等人, 2022年）。从图 6可以看出，CMETNet 在七种方法中表现最佳。CMETNet 的结果优于 COMB，这表明在框架中纳入卷积神经网络模型十分重要。基于 PPMCC 和平均绝对误差得出的结果一致。

我们数据集中的每个事件 $E$ 包含 5 个日冕物质抛射特征（记为 F）、7 个太阳风参数（记为 W）和 0–17 张日冕物质抛射图像（记为 I）。为评估这些特征、参数和图像的有效性，我们进行了额外实验，考察以下七种情况：

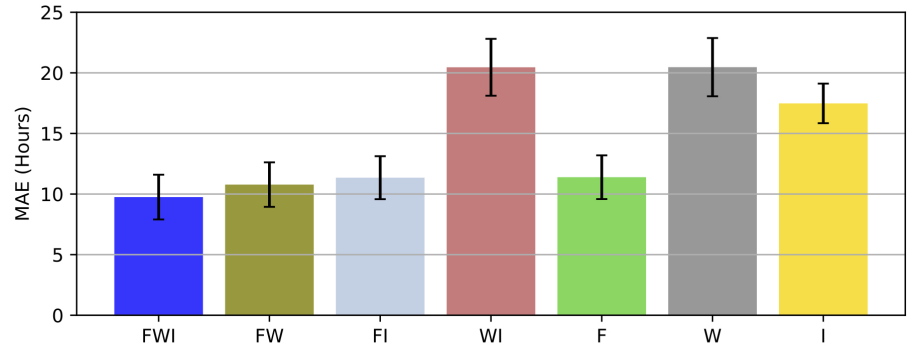
- $E$ 包含所有日冕物质抛射特征、太阳风参数和日冕物质抛射图像（记为 FWI）；
- $E$ 仅包含日冕物质抛射特征和太阳风参数（记为 FW）；
- $E$ 仅包含日冕物质抛射特征和日冕物质抛射图像（记为 FI）；
- $E$ 仅包含太阳风参数和日冕物质抛射图像（记为 WI）；
- $E$ 仅包含日冕物质抛射特征（记为 F）；
- $E$ 仅包含太阳风参数（记为 W）；
- $E$ 仅包含日冕物质抛射图像（记为 I）。

在每种情况下，我们都将 CMETNet 应用于现有数据。请注意，FWI 等同于 CMETNet。在 FW 情况下，CMETNet 相当于前述的 COMB 模型。在 I 情况下，CMETNet 相当于前述的卷积神经网络模型。

图7 展示了七种情况的 PPMCC 和平均绝对误差结果。从图7 可以看出，FWI 的结果最准确，PPMCC 为 0.83，平均绝对误差为 9.75 小时，这表明将三类数据（日冕物质抛射特征、太阳风参数和日冕物质抛射图像）结合起来能够取得最佳性能。仅结合两类数据时，FI 的 PPMCC 最高




(a)




(b)

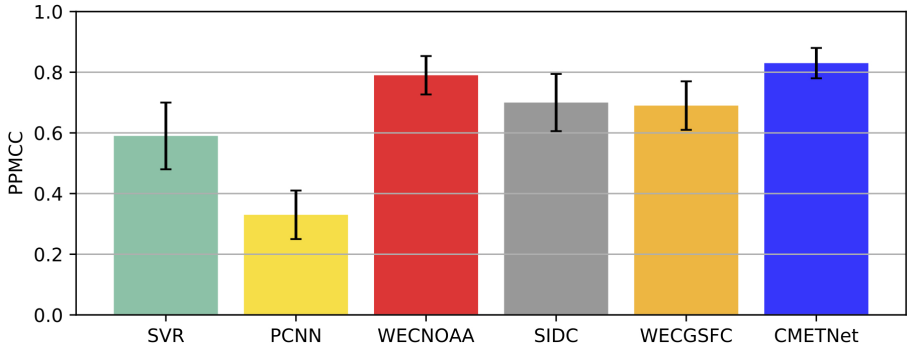
图7。评估七种情况（FWI、FW、FI、WI、F、W、I）的消融测试结果。其中，FWI 表示日冕物质抛射特征、太阳风参数和日冕物质抛射图像的组合；FW 表示日冕物质抛射特征与太阳风参数的组合；FI 表示日冕物质抛射特征与日冕物质抛射图像的组合；WI 表示太阳风参数与日冕物质抛射图像的组合；F 表示日冕物质抛射特征；W 表示太阳风参数；I 表示日冕物质抛射图像。(a) 各测试情况的皮尔逊积矩相关系数（PPMCCs）。(b) 各测试情况的平均绝对误差（MAEs）。在所有测试情况中，FWI 的表现最佳。

为 0.80；另一方面，FW 的平均绝对误差最低，为 10.32 小时。分别单独使用三类数据时，F 在两项评估指标上均取得最佳结果，PPMCC 为 0.78，平均绝对误差为 11.39 小时。因此，日冕物质抛射特征分别比太阳风参数和日冕物质抛射图像具有更强的预测能力。

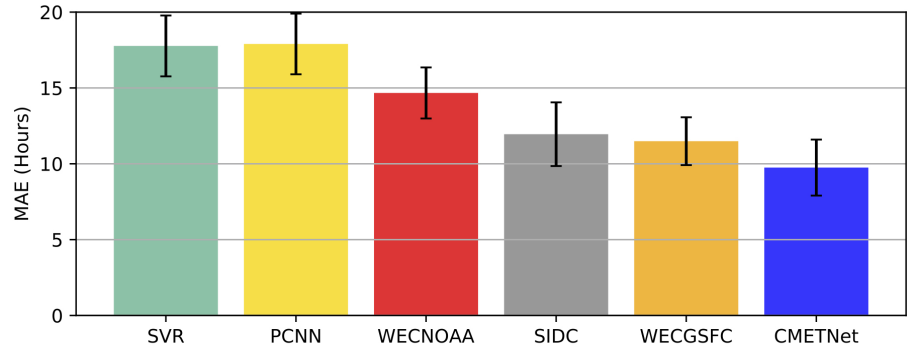
### 4.3. 与相关方法的比较

在本实验中，我们将 CMETNet 与此前发表的用于日冕物质抛射到达时间预测的机器学习方法进行了比较。Liu 等（2018年）利用本文所采用的日冕物质抛射特征和太阳风参数，开发了一种基于支持向量回归的方法。该方法采用的支持向量回归模型与 CMETNet 中的相同。我们将他们的方法简称为支持向量回归。Wang 等人（2019年）利用本文所采用的日冕物质抛射图像，开发了一种卷积神经网络模型。他们的卷积神经网络模型与 CMETNet 中使用的模型不同。我们将他们的卷积神经网络模型称为 PCNN（即先前的卷积神经网络）。需要指出的是，尽管支持向量回归和 PCNN 分别使用的特征/参数和图像与 CMETNet 所用的相同，但本文所介绍的数据集比先前研究中的数据集（Liu 等，2018年；Wang 等人，2019年）规模大得多，也更加全面。因此，本文报告的预测结果与Liu 等（2018年）和Wang 等人（2019年）给出的结果并不完全相同。

此外，我们将 CMETNet 与三种基于物理的模型进行了比较，包括：(i) 由美国国家海洋和大气管理局（NOAA）空间天气预报中心实现的 WSA-ENLIL+Cone（日冕物质抛射，WEC）模型（Odstrcil 等人，2004年；Riley 等人，2018年），记为 WECNOAA；(ii) 比利时皇家天文台太阳影响数据分析中心（SIDC）开发的模型，可在 https://www.sidc.be/ 获取，记为 SIDC；以及 (iii) 由+美国国家航空航天局戈达德太空飞行中心（GSFC）的社区协调建模中心（CCMC）实现的 WSA-ENLIL Cone（日冕物质抛射，WEC）模型




(a)




(b)

图 *8.* 六种方法在日冕物质抛射到达时间预测方面的性能比较。CMETNet 是本文提出的集成学习框架。SVR 和 PCNN 是此前发表的机器学习方法。WECNOAA、SIDC 和 WECGSFC 是美国国家航空航天局 CCMC CME 评分榜上展示的基于物理的模型。(a) 六种方法的皮尔逊积矩相关系数（PPMCC）。(b) 六种方法的平均绝对误差（MAE）。在 PPMCC 和 MAE 两项指标上，CMETNet 均优于其他五种方法。

，记为 WECGSFC。这三种基于物理的模型被认为是最佳的物理模型，其预测结果最为准确，相关结果已提交至美国国家航空航天局的 CCMC 日冕物质抛射排行榜，网址为 https://kauai.ccmc.gsfc.nasa.gov/CMEscoreboard/（Riley 等人，2018年）。

图 8 展示了六种方法（SVR、PCNN、WECNOAA、SIDC、WECGSFC 和 CMETNet）的 PPMCC 和 MAE 结果。总体而言，CMETNet 的结果最准确，PPMCC 为 0.83，MAE 为 9.75 小时。在三种基于物理的方法中，WECNOAA 的 PPMCC 最高，为 0.79；而 WECGSFC 的 MAE 最低，为 11.49 小时。CMETNet 和这三种基于物理的方法的表现均优于此前发表的机器学习方法 SVR 和 PCNN（Liu 等，2018年；Wang 等人，2019年）。

## 5. 结论与未来工作

本文提出了一种用于预测日冕物质抛射事件到达时间的集成框架（CMETNet）。每个事件包含 5 个日冕物质抛射特征（角宽度、主位置角、线速度、最终高度处的二阶速度和质量）、7 个太阳风参数（$B_x$、$B_z$、氦离子与质子比、流向经度、等离子体压强、流速和质子温度）以及 0–17 张日冕物质抛射图像。CMETNet 框架由五种机器学习模型组成，包括支持向量回归（SVR）、随机森林（RF）、高斯过程（GP）、XGBoost（XGB）和卷积神经网络（CNN）。实验结果表明，CMETNet 的表现优于各个单独的机器学习模型，PPMCC 为 0.83，MAE 为 9.75 小时。此外，同时使用所有日冕物质抛射特征、太阳风参数和日冕物质抛射图像，比仅使用其中一部分取得了更好的表现。

随后，我们将 CMETNet 与两种先前发表的机器学习方法进行比较（Liu 等人，2018年；Wang

等人，2019年）以及三种基于物理的方法（WECNOAA、SIDC 和 WECGSFC）。这三种基于物理的方法被认为是性能最佳的物理模型，在提交至美国国家航空航天局的 CCMC CME 评分榜的结果中，其预测准确度最高。实验结果表明，与两种先前发表的机器学习方法相比，CMETNet 和这三种基于物理的方法均能得出更准确的结果；在所有方法中，CMETNet 的表现最佳。

CMETNet 的成功归功于整合了来自多个来源的数据，并结合了多种机器学习模型。尽管如此，准确预测日冕物质抛射的到达时间仍然是一项艰巨的任务，因为 CMETNet 的平均绝对误差（9.75 小时）仍然较大，说明仍有很大的改进空间。提升预测模型性能的一种可能方法，是像近期一项研究那样将机器学习与基于物理的方法相结合（Tiwari 等人，2021年）。Dumbovi´等人（2021年）回顾了多种用于预测日冕物质抛射到达时间和速度的基于阻力模型（DBM）工具。未来，我们计划探索将这些 DBM 工具与不同机器学习模型结合的方法，以更准确地预测日冕物质抛射到达时间。

未来研究的另一个方向是开发基于机器学习的业务化近实时日冕物质抛射到达时间预测系统。日冕物质抛射特征和太阳风参数在这里并不适用，因为它们无法实时获取。我们计划在日冕物质抛射发生并被 LASCO C2 图像捕获后（
h
t
t
p
s://www.swpc.noaa.gov/products/lasco-coronagraph）或 STEREO COR2（
https://stereo-ssc.nascom.nasa.gov/browse/）观测到后立即预测其传播时间；这些图像可在数小时内获取。该近实时预测系统需要新的先进卷积神经网络模型，从日冕物质抛射图像中学习潜在特征或表征。开发此类先进卷积神经网络模型还需要更多研究。

## 参考文献


Abduallah, Y., Wang, J. T. L., Nie, Y., Liu, C., 和Wang, H. DeepSun：用于太阳耀斑预测的机器学习即服务。《天文学与天体物理学研究》，21 (7):160，2021年8月。

Abduallah, Y., Jordanova, V. K., Liu, H., Li, Q., Wang, J. T. L., 和 Wang, H. 利用SDO/HMI矢量磁场数据产品和双向长短期记忆网络预测太阳高能粒子。《天体物理学杂志增刊》，260(1):16，2022年5月。

Astropy 合作组、Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M.,Bray,E., Aldcroft,T., Davis, M., Ginsburg, A., Price-Whelan, A. M., Kerzen- dorf, W. E., Conley, A., Crighton, N., Barbary, K., Muna, D., Ferguson, H., Grollier, F., Parikh, M. M., Nair, P. H., Unther, H. M., Deil, C., Woillez, J., Conseil, S., Kramer, R., Turner, J. E. H., Singer, L., Fox, R., Weaver, B. A., Zabalza, V., Edwards, Z. I., Azalee Bostroem, K., Burke, D. J., Casey, A. R., Crawford, S. M., Dencheva, N., Ely, J., Jenness, T., Labrie, K., Lim, P. L., Pierfederici, F., Pontzen, A., Ptak, A., Refsdal, B., Servillat, M., 和 Streicher, O. Astropy：用于天文学的社区 Python 软件包。《天文学与天体物理学》，558: A33，2013年10月。

Berk, K. N. 回归分析（Ashish Sen 和 Muni Srivastava）。*SIAM Rev.*，34(1):157–158，1992。

Breiman, L. 装袋预测器。机器学习，24(2): 123–140，1996年。Breiman, L. 随机森林。机器学习，45:5–32，2001年1月。Camporeale, E. 空间天气中机器学习面临的挑战：临近预报与预报。《空间天气》，17 (8):1166–1207，2019年。Chen, T. 和 Guestrin, C. XGBoost：一种可扩展的树提升系统。载于 Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D. 和 Rastogi, R.（编），第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第785–794页，2016年。Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., Du-Fumier, B., Zhou, T., Wang, X., Zhu, H., Sun, Z. 和 Gombosi, T. I. 利用 SDO/HMI 图像时间序列和 SHARP 参数识别太阳耀斑前兆。《空间天气》，17(10):1404–1426，2019年。Cortes, C. 和 Vapnik, V. 支持向量网络。机器学习，20(3):273–297，1995年。Daglis, I., Baker, D., Kappenman, J., Panasyuk, M. 和 Daly, E. 空间天气对技术基础设施的影响。《空间天气》，2(2):S02004，2004年。Dietterich, T. G. 机器学习中的集成方法。载于 Kittler, J. 和 Roli, F.（编），多分类器系统：第一届国际研讨会论文集，收入《计算机科学讲义》第1857卷，第1–15页。Springer，2000年。ˇ Dumbovi´c, M., Calogovi´c, J., Martini´c, K., Vrˇ snak, B., Sudar, D., Temmer, M. 和 Veronig, A. 用于预报日冕物质抛射到达时间和速度的基于阻力模型（DBM）工具。*Frontiersin Astronomy andSpaceSciences*，8:58，2021年。

Goodfellow, I., Bengio, Y. 和 Courville, A. 深度学习。MIT出版社，2016年。Gopalswamy, N. 日冕物质抛射的历史与发展：太阳—地球关系中的关键因素。《地球科学快报》，3:8，2016年。Gopalswamy, N., Yashiro, S., Michalek, G., Stenborg, G., Vourlidas, A., Freeland, S. 和 Howard, R. SOHO/LASCO日冕物质抛射目录。《地球、月球与行星》，104(1-4):295–313，2009年。

Gruet, M. A.、Chandorkar, M.、Sicard, A. 和 Camporeale, E. 结合长短期记忆神经网络与高斯过程，对Dst指数进行提前数小时的预测。《空间天气》，16(11):1882–1896，2018年。

G¨ortler, J.、Kehlbeck, R. 和 Deussen, O. 高斯过程的可视化探索。*Distill*，2019年。

Hess, P. 和 Zhang, J. 对太阳活动周24中影响地球的日冕物质抛射的研究。《太阳物理学》，292(6):80，2017年。

Inceoglu, F.、Jeppesen, J. H.、Kongstad, P.、Hern´andez Mar- cano, N. J.、Jacobsen, R. H. 和 Karoff, C. 使用机器学习方法预测太阳耀斑是否会伴随日冕物质抛射和太阳高能粒子。《天体物理学杂志》，861(2):128，2018年。

Iong, D., Chen, Y., Toth, G., Zou, S., Pulkkinen, T. I., Ren, J., Camporeale, E., 和 Gombosi, T. I. I. 使用梯度提升机进行可解释 SYM-H预报的新发现。地球与空间科学开放档案，2022年。

Jiao, Z., Sun, H., Wang, X., Manchester, W., Gombosi, T., Hero, A., 和 Chen, Y. 使用机器学习模型预测太阳耀斑强度。《空间天气》，18(7): e02440，2020年。

Kawabata, Y., Iida, Y., Doi, T., Akiyama, S., Yashiro, S., 和 Shimizu, T. 活动区 S形结构与太阳耀斑和日冕物质抛射之间的统计关系。《天体物理学杂志》，869(2):99，2018年。

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. 使用基于梯度的学习进行物体识别。载于 Forsyth, D. A., Mundy, J. L., Ges`u, V. D., and Cipolla, R. (编), 《计算机视觉中的形状、轮廓与分组》，第 1681 卷，《计算机科学讲义》，第 319 页。Springer，1999年。Liu, C., Deng, N., Wang, J. T. L., and Wang, H. 利用SDO/HMI矢量磁场数据产品和随机森林算法预测太阳耀斑。《天体物理学杂志》，843(2):104，2017年。

Liu, H., Liu, C., Wang, J. T. L., and Wang, H. 使用长短期记忆网络预测太阳耀斑。《天体物理学杂志》，877(2):121，2019年。Liu, H., Liu, C., Wang, J. T. L., and Wang, H. 利用 SDO/HMI矢量磁场数据产品和循环神经网络预测日冕物质抛射。《天体物理学杂志》，890(1):12，2020年。Liu, H., Xu, Y., Wang, J., Jing, J., Liu, C., Wang, J. T. L., and Wang, H. 使用卷积神经网络从 GST/NIRIS 的斯托克斯谱推断矢量磁场。《天体物理学杂志》，894(1):70，2020年。

Liu, J.、Ye, Y.、Shen, C.、Wang, Y. 和 Erd´elyi, R. 一种使用机器学习算法预测日冕物质抛射到达时间的新工具：CAT-PUMA。《天体物理学杂志》，855(2):109，2018年。

Maloney, S. 日冕物质抛射在内日球层中的传播。博士论文，都柏林圣三一学院物理学院，2012年。

Maloney, S., Byrne, J., Gallagher, P. T., 和 McAteer, R. T. J. 三维空间中日冕物质抛射前沿的传播。载于第*38*届国际空间研究委员会科学大会，第38卷，第5页，2010年。Odstrcil, D., Riley, P., 和 Zhao, X. P. 1997年5月12日行星际日冕物质抛射事件的数值模拟。《地球物理研究杂志：空间物理》，109(A2)，2004年。Paouris, E. 和 Mavromichalaki, H. 第23太阳活动周期间利用SOHO和ACE组合数据研究由朝向地球的日冕物质抛射引发的行星际日冕物质抛射。《太阳物理学》，292(2):30，2017年。Pearson, K. 关于双亲情形下回归与遗传的说明。《伦敦皇家学会会报》，58(1):240–42，1895。Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., 和 Duchesnay, E. Scikit-learn：Python中的机器学习。《机器学习研究杂志》，12:2825–2830，2011年。Priest, E. R. 和 Forbes, T. G. 太阳耀斑的磁性质。《天文学与天体物理学评论》，10(4): 313–377，2002年。Raheem, A.-u., Cavus, H., Coban, G. C., Kinaci, A. C., Wang, H., 和 Wang, J. T. L. 通过确定源活动区，研究太阳黑子群与日冕物质抛射之间的因果关系。《皇家天文学会月报》，506 (2):1916–1926，2021年。

Richardson, I. G. 和 Cane, H. V. 第23太阳活动周（1996—2009年）期间近地行星际日冕物质抛射：目录及其性质概述。《太阳物理学》，264(1):189–237，2010年。

Riley, P., Mays, M. L., Andries, J., Amerstorfer, T., Biesecker, D., Delouille, V., Dumbovi´c, M., Feng, X., Henley, E., Linker, J. A., M¨ostl, C., Nu˜nez, M., Pizzo, V., Temmer, M., Tobiska, W. K., Verbeke, C., West, M. J. 和 Zhao, X. 日冕物质抛射到达时间预报：CCMC CME 评分榜分析。《空间天气》，16(9):1245–1260，2018年。

Schwenn, R., dal Lago, A., Huttunen, E., 和 Gonzalez, W. D. 日冕物质抛射与其在地球附近产生的影响之间的关联。《地球物理年鉴》, 23(3): 1033–1059, 2005年。Shen, C., Wang, Y., Pan, Z., Miao, B., Ye, P., 和 Wang, S. 全晕日冕物质抛射：到达地球。《地球物理研究杂志（空间物理学）》, 119 ( 7):5107–5116, 2014年。Showstack, R. 专家警告，空间天气带来的潜在风险可能增加。《*EOS* 汇刊》, 94(25): 222–223, 2013年。Sun, Z., Bobra, M. G., Wang, X., Wang, Y., Sun, H., Gom- bosi, T., Chen, Y., 和 Hero, A. 利用卷积神经网络和长短期记忆网络，根据两个太阳活动周的活动区数据预测太阳耀斑。《天体物理学杂志》, 931(2):163, 2022年。Tang, R., Zeng, X., Chen, Z., Liao, W., Wang, J., Luo, B., Chen, Y., Cui, Y., Zhou, M., Deng, X., Li, H., Yuan, K., Hong, S., 和 Wu, Z. 利用多种卷积神经网络变体和集成学习，按磁类型对太阳黑子群进行分类。《天体物理学杂志增刊》, 257(2):38, 2021年。Tiwari, A., Camporeale, E., Teunissen, J., Foldes, R., Napo- letano, G., 和 Del Moro, D. 日冕物质抛射到达时间预测：机器学习与集成方法。载于欧洲地球科学联盟大会论文摘要，《欧洲地球科学联盟大会论文摘要》，第 EGU21–7661 页，2021年。Vourlidas, A., Patsourakos, S., 和 Savani, N. P. 预测日冕物质抛射的地球效应特性：现状、未决问题与未来方向。《伦敦皇家学会哲学汇刊 *A* 辑》, 377(2148):20180096, 2019年。Wang, X., Chen, Y., Toth, G., Manchester, W. B., Gombosi, T. I., Hero, A. O., Jiao, Z., Sun, H., Jin, M., 和 Liu, Y. 利用机器学习预测太阳耀斑：研究太阳活动周依赖性。《天体物理学杂志》, 895(1):3, 2020年。

Wang, Y., Liu, J., Jiang, Y., 和 Erd´elyi, R. 利用卷积神经网络预测日冕物质抛射到达时间。《天体物理学杂志》, 881(1):15, 2019年。Yashiro, S. 和 Gopalswamy, N. 太阳耀斑与日冕物质抛射之间的统计关系。国际天文学联合会会议录, 4(S257): 233–243, 2008年。Yuan, J., Wang, K., Yu, T., 和 Fang, M. 基于高斯过程回归的高速线切割放电加工工艺可靠多目标优化。《国际机床与制造杂志》, 48(1):47–60, 2008年。

Zhao, X. 和 Dryer. 日冕物质抛射/激波到达时间预测的现状。《空间天气》，12(7):448–469，2014年。