计算太阳能——基于印度东部气象参数的太阳能发电预测集成学习方法

DEBOJYOTI CHAKRABORTY 和 JAYEETA MONDAL，数据科学与工程系，BITS-Pilani，印度拉贾斯坦邦 333031；HRISHAV BAKUL BARUA，机器人与自主系统，TCS Research，印度加尔各答 700156；ANKUR BHATTACHARJEE，电气与电子工程系，BITS-

印度海得拉巴，500078，皮拉尼校区

通讯作者：电子邮箱 - a.bhattacharjee@hyderabad.bits-pilani.ac.in

摘要

太阳能应用面临的挑战在于其间歇性，以及对气象参数的依赖，例如太阳辐射、环境温度、降雨、风速等，还有灰尘积聚等其他物理参数。因此，估算特定地理位置的太阳能光伏发电量至关重要。机器学习（ML）模型日益受到重视，并广泛用于预测太阳能发电厂性能。本文首次采用多种集成机器学习模型（EML），如装袋法、提升法、堆叠和投票法，评估气象参数对太阳能光伏发电的影响。通过印度东部地区一座10千瓦峰值太阳能光伏电站的实地数据集，验证了所选机器学习算法的性能。此外，本文还设计了完整的测试平台框架，用于数据挖掘并选择合适的学习模型。该框架还支持对数据集进行特征选择和降维，以降低学习模型的空间和时间复杂度。结果表明，堆叠和投票法集成机器学习模型的预测准确率约为96%，表现优异。本文提出的方法具有通用性，也可用于预测大型太阳能光伏电站的性能。

关键词：太阳能光伏、集成学习、气象数据、功率预测

本项目的官方源代码和实现可在我们的GitHub 项目页面上找到

符号说明		
缩写	全称	国际单位制单位
RH_Avg	平均相对湿度	RH
RA_mm	降雨累积量	毫米
平均气温	平均气温	$^{\circ}C$
平均风速	平均风速	m/s
平均风向	平均风向	度
DP_Avg	平均露点	C
GR_Avg	平均全球辐射	千瓦时/平方米
DiffR_Avg	平均散射辐射	千瓦时/平方米
AP_QNH_Avg	平均QNH大气压	毫巴
平均场压	QFE 平均大气压	毫巴
DirR_Avg	平均直接辐射	千瓦时/平方米
HSR_Avg	平均水平太阳辐射	千瓦时/平方米

作者地址：Debojyoti Chakraborty 和 Jayeeta Mondal，数据科学与工程系，BITS-Pilani，印度拉贾斯坦邦 333031，debojyotichakrabarti@gmail.com，jayeetamondal5555@gmail.com；Hrishav Bakul Barua，机器人与自主系统，TCS Researc h，印度加尔各答 700156，hbarua@acm.org；Ankur Bhattacharjee，电气与电子工程系，BITS-Pilani，海得拉巴校区，印度海得拉巴，500078，a.bhattacharjee@hyderabad.bits-pilani.ac.in。

ISR_Avg	倾斜面太阳辐射平均值	千瓦时/平方米
SA_Avg	太阳方位角平均值	度
DTR_Avg	直接理论辐射平均值	千瓦时/平方米
GR_Avg	总能量平均值	焦耳
DiffE_Avg	散射能量平均值	千瓦时/平方米
DirE_Avg	直接能量平均值	千瓦时/平方米
ML	机器学习	-
EML	集成机器学习	-
PV	光伏	-
GW	吉瓦	-
KW	千瓦	-
自回归积分滑动平均	自回归积分滑动平均	-
ANN	人工神经网络	-
KNN	K近邻	-
PCA	主成分分析	-
NMP	数值气象预报	-
SVM	支持向量机	-
GEFS	全球集合预报系统	-
SRRA	太阳辐射资源评估站点	-
DST	科学技术部	-
CEGESS	绿色能源与传感器系统卓越中心	-
KDE	核密度估计	-
RMSE	均方根误差	-
GOSS	基于梯度的单边采样	-
EFB	互斥特征捆绑	-
GBDT	梯度提升决策树	-
GBM	梯度提升机	-
R^2	R平方分数	-
ROI	投资回报率	-
MNRE	可再生能源部	-
R^d	d 维实数集	-
R	实数集	-
X	输入数据点集	-
Y	X 个输入数据点对应的真实数据点集	-
E	期望损失	-
n	输入数据点总数	-
D	用于最小化误差的分布	-
B	用于更新 AdaBoost 中权重向量的参数	-
ψ	梯度提升机中联合分布上的期望损失	-
L	K 个回归器的学习集	-
Ω	XGBoost 型号复杂度的惩罚函数	-

1 引言

为减轻使用化石燃料能源对环境造成的影响，并满足人类社会、工业等日益增长的电力需求，可再生能源正成为颇具前景的解决方案。由于太阳能和风能发电能力持续提升，美国可再生能源发电量的年度占比将从2021年的20%升至2023年的24% [1]。截至2021年9月，印度的可再生能源装机容量为101.53吉瓦（GW），占总装机容量的38%。目前的目标是到2030年实现约450吉瓦的可再生能源装机容量，其中约280吉瓦（超过60%）预计来自太阳能[2] 本身。鉴于上述统计数据以及近期太阳能发电项目预计将增加，准确、高效地进行太阳能发电预测[3] 显然需求迫切。本文针对一座安装在印度东部地区、千瓦级的太阳能光伏电站开展了实际案例研究，并对其性能进行了预测和验证。这对于太阳能光伏新建项目具有重要意义，涉及：1）气象和环境影响；2）满足某一区域的太阳能发电预期；3）选择合适的预测算法；以及4）估算太阳能投资者的投资回报率（ROI）。为了确保太阳能渗透率较高的发电系统平稳运行，采用合适的太阳能发电预测方案至关重要。本文着重利用印度东部地区的数据开展区域太阳能发电预测。尽管预测方法种类繁多，鉴于集成机器学习（EML）预测算法在构建准确系统方面取得了大规模成功，本文将重点研究此类算法。由于气象数据通常具有时间序列特征，因此可以在一定程度上采用基于学习的方法[4]。

已有大量关于基于深度学习的太阳能发电预测系统的研究 [5], [6], [7], [8],[9]。一些早期研究采用了主成分分析（PCA）等常用数据分析工具，构建适用于太阳辐照度预测多变量分析的数据集 [10], [11], [12]。一些早期研究展示了利用基于领域知识的深度学习方法 [13], 、短期预测模型 [14], 、基于皮尔逊相关性的预测[15], ，以及通过调整特定地点的光伏（PV）输出预测[16] 来预测太阳辐射。[17], [18], [19], 中的作者采用了多种预测回归模型和集成技术，以提高太阳能发电预测的准确度。本文从自身贡献的角度，对该领域的相关研究进行了简短而全面的综述。通过对早期研究的深入考察，我们得出结论：数据挖掘技术在显著提高任何基于机器学习的预测系统的准确度方面发挥着重要作用。Voyant等人 [50] 提出了一种基于预测区间方法的稳健函数模型，其效率和效果优于采用自助法的经典预测工具。不过，这类模型专用于周期性时间序列，并不局限于平稳时间序列。Ayodele等人 [51] 将K均值与支持向量回归（SVR）结合，用于全球太阳辐射预测。作者利用K均值将季节性数据划分为不同类别，并为雨季和旱季确定质心。研究使用伊巴丹六年（2010—2015年）的气象数据，对模型的性能进行了验证，以预测未来年份（2016—2017年）的每日全球太阳辐射。预测值还与三种成熟模型（即人工神经网络、Angstrom–Prescott模型和自回归滑动平均模型）得出的预测值进行了比较。Peder等人和Coimbra等人 [20] 针对最长36小时预测时段的太阳能发电小时值开展了研究，对自回归积分滑动平均模型（ARIMA）、人工神经网络（ANN）和K近邻（KNN）模型进行了基准测试。研究使用光伏板的历史输出水平，即未将外生数值天气预报（NWP）数据作为输入。Davò等人 [21] 采用主成分分析技术进行特征选择和降维，并应用人工神经网络和模拟集合（AnEn）预测太阳辐照度。作者指出，云量、湿度水平、是否有雾等因素造成的每日气象条件细微变化，都可能影响太阳能发电。Shi等人 [22] 提出了一种气象分类算法，并采用支持向量机（SVM）按15分钟间隔预测次日光伏功率输出。气象数据被分为四类：晴空、多云天气、有雾天气和雨天。[23] 中的作者研究了基于GEFS（即全球集合预报系统）计算的数值天气预报（NWP）；该系统可预测网格中各点的气象变量。他们研究了不同机器学习技术在以下情境中的应用：

使用基于NOAA/ESRL 全球集合预报系统的数值天气预报模型开展太阳能预测，该系统针对网格中的不同节点计算数值天气预报。在我们的研究中，我们进一步扩展了用于气象数据训练和测试的机器学习算法种类，以便更深入地了解变化情况和预测分析结果。表中系统地呈现了现有类似研究与本文的比较。1。

表1. 与我们研究相似的现有研究对比（考察最近5年）。

研究	使用的方法	预测参数	集成学习方法	现场数据测试	端到端数据分析与机器学习框架（广泛气象特征完成的工程与分析
Fan等（2018年） [24]	支持向量机和XGBoost	每日全球太阳辐射	✗	✗	✗
Srivastava等人（2018年）[25]	前馈、反向传播、深度学习，以及模型平均网络	提前6天预测全球太阳辐射	✗	✓	✗
Mpfumali等（2019年）[26]	随机梯度提升和支持向量回归	日前逐小时全球水平面辐照度	✗	✓	✗
Coban 等人（2019）[27]	传统机器学习模型、概率朴素模型和基于树的回归	全球水平辐照量	✗	✗	✗
Chen 等人（2020）[28]	静态和动态集成	功率生成的	✓	✓	✗
Amarasinghe 等人（2020年）[29]	堆叠和加权平均使用深度置信网络的集成方法网络（DBN）、随机森林（RF）和SVM	发电量	✓	✓	✗
Mutavhatsindi 等人（2020年）[30]	静态和动态集成	发电量	✓	✓	✗
Guermoui 等人（2020年）[31]	通用集成学习方法(GELA)、基于聚类的集成学习方法(CELA)、基于分解的集成学习方法（DELA），基于分解聚类的学习方法（DCLE），进化集成学习方法（EELA）	全球水平辐照量	✓	✗	✗
Zhou 等（2021）[32]	人工神经网络、核、树、集成方法的综述，基于分解聚类的机器学习算法	全球太阳辐射	✓	✗	✗
Jebli 等（2021）[33]	循环神经网络（RNN）、长短期记忆网络（LSTM）和门控循环单元（GRU）	太阳能	✗	✓	✗
拟议工作	经典机器学习算法和集成技术（装袋法，提升法、堆叠和投票法）	太阳能发电	✓	✓	✓

基于上述讨论，本文的主要贡献体现在以下三个方面。1. 对从太阳辐射资源评估（SRRA）中心和屋顶太阳能光伏电站收集的现场数据进行了预处理（数据清洗、降维和特征选择）并加以分析。

2. 处理后的数据集已输入各种机器学习模型，用于预测气象参数（直射辐射、总辐射、倾斜辐射和散射辐射、环境温度、太阳方位角、风速、风向和相对湿度）影响下的太阳能光伏发电。在此阶段，研究认为，与其他经典回归模型相比，集成机器学习模型在预测准确率方面更优。本研究选用的集成机器学习模型包括装袋法、提升法、堆叠和投票法。

3. 研究使用一座10千瓦$_p$屋顶太阳能电站的数据集验证了这些算法的性能。预测结果表明，堆叠和投票法集成机器学习算法的预测准确率约为96%，优于本研究考虑的其他集成机器学习算法。因此，所推荐的集成机器学习算法对于预测位于特定地理位置的大型太阳能电站的性能也非常有用。此外，可以推断，所提出的测试平台还可用于任何新建太阳能项目，只要适当确定合适的训练参数，即可预测其性能。

所提出测试平台框架的官方工作代码（以及预训练学习模型和新颖的气象数据集）已发布在我们的GitHub项目网页。

本文组织如下。第2节介绍了我们提出的研究与分析方法。第3节详细说明数据集的收集与设置、预处理、分析，以及多种气象参数对太阳能发电的影响。第4.1节探讨了论文中使用的各种学习技术，随后在第4.2节讨论实验。第4.3 节展示结果分析，最后在第5节总结我们的分析、研究和实验。

2 我们研究与分析的拟议框架：概述与设计

本节概述了我们为研究、分析和实验提出的测试平台框架。流程始于从 SRRA 站点的源头收集和整理数据。本文使用的是在区域发电站收集的 SRRA 气象数据，其中存在因传感器系统停机或传感器故障等造成的缺失值和异常值等问题。因此，必须谨慎设计处理流程，以尽量减轻所收集数据中的上述质量问题对模型预测的影响。该框架的框图（含数据/信息流向箭头）如图1所示。

(1) 数据采集：第3.1 节讨论了本研究所用实地数据集的采集。使用实地数据有助于验证相关论断，并可在实施新建太阳能项目时加以复现。(2) 数据分析与预处理：数据采集完成后，需要进行清理，以去除数据集中的误差，相关内容已在第3.2.1节讨论。探索性数据分析揭示了气象数据特征对目标变量预测的影响。第3.2.2节展示了分析与处理所采取的各个步骤。(3) 集成学习模型的探索、实验与结果分析：本节分为以下部分：I. 第4.1 节从理论层面探讨了各种集成学习模型。II. 接着，第4.2 节展示了用于训练和测试集成机器学习算法的实验设置与流程。III. 最后，第4.3 节基于评估指标和性能，对在SRRA数据集上测试的各种算法进行比较，并提出最适用于东部地区气象数据的算法。

3 实际数据采集、预处理与数据分析

本节讨论在印度东部地区收集的气象数据集，以及用于数据分析和预处理的各种方法。

3.1 数据采集

本研究使用的太阳能发电数据集由连续的气象参数以及每组观测数据所对应的发电量组成。数据收集自太阳辐射资源

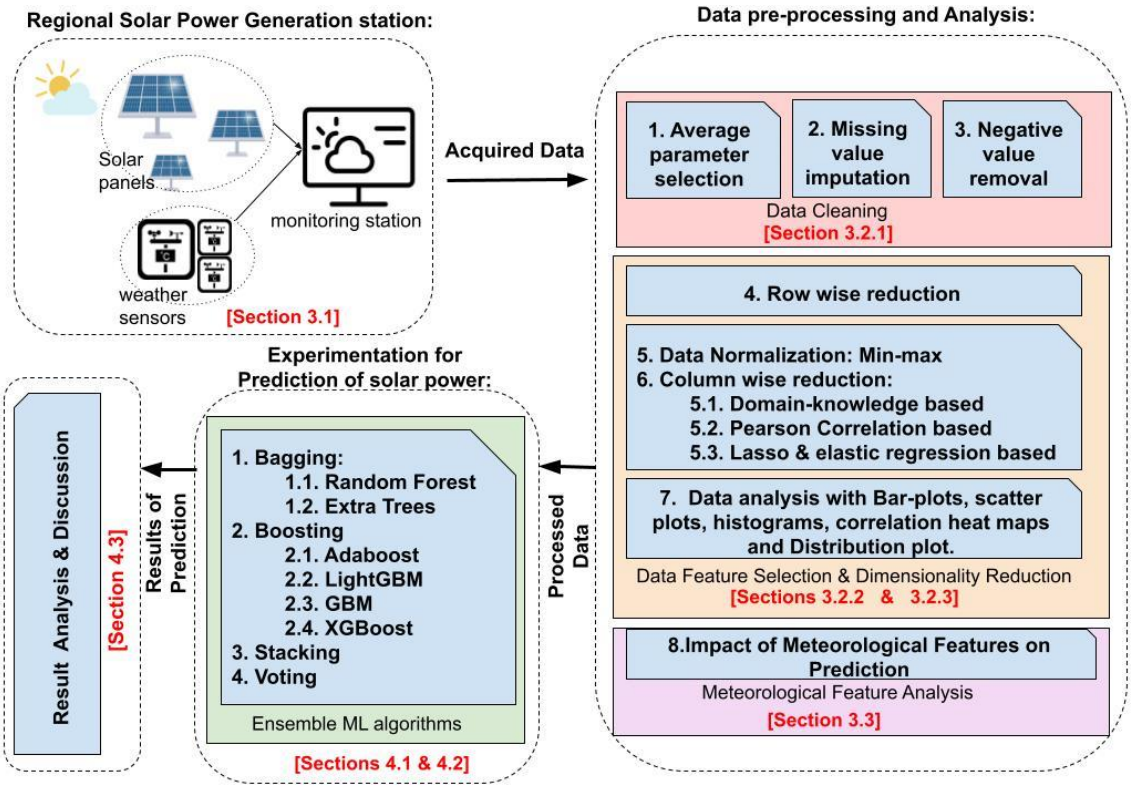


图 1. 所提出的测试平台框架全貌，展示了针对 SRRA 区域气象数据（此处为印度东部的数据）开展数据收集/整理、数据挖掘/分析和监督学习方法研究、分析与实验的工作流程，用于太阳能发电预测。

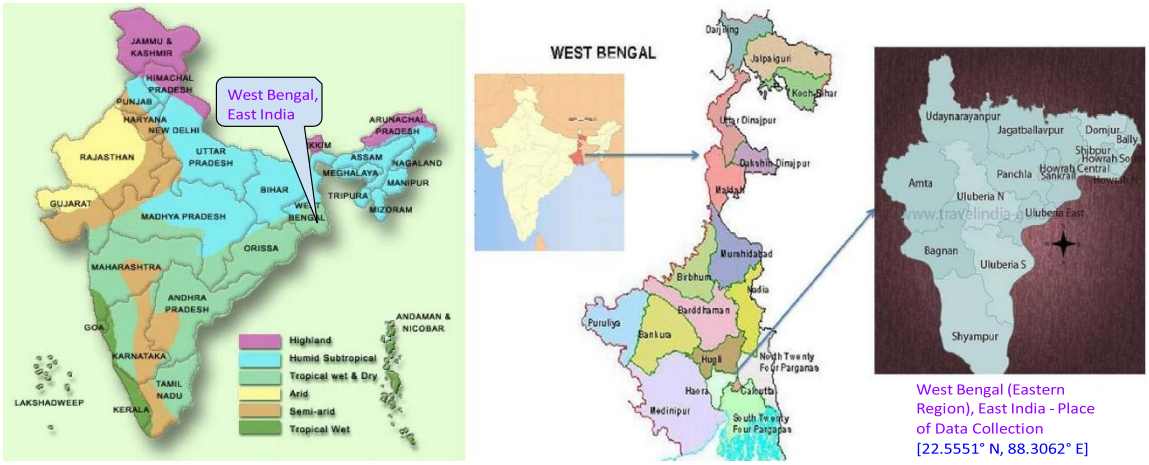


图 2. 数据收集地点（位于印度东部西孟加拉邦（WB），地图所示坐标：22.5551° N，88.3062° E）及一般气象条件（图片采自 Google 图片）。

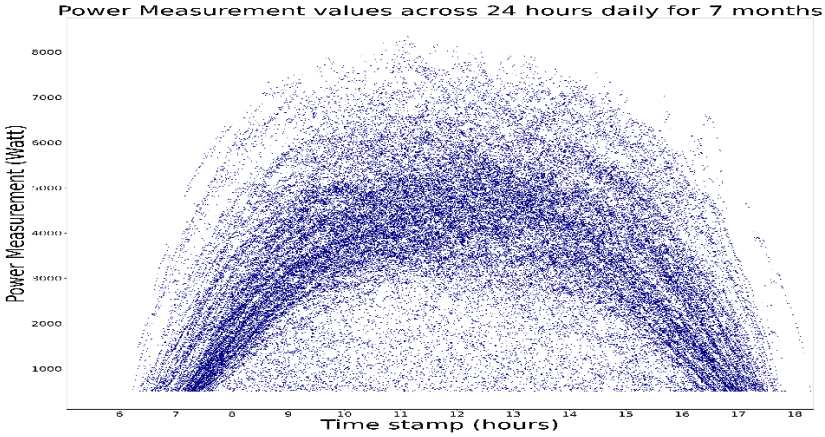


图 3. 7 个月期间 24 小时太阳能发电测量的可视化结果。

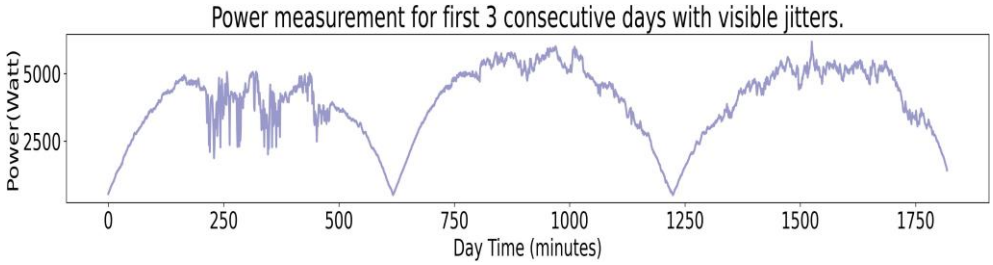


图 4. 前 3 个连续日间的功率测量，展示太阳能发电中明显的抖动。

评估站点（SRRA）[1]于2014年建成，太阳能光伏电站位于印度工程科学与技术学院（IIEST）希布尔校区的屋顶上，该地点位于西孟加拉邦（WB），印度东部，如图2所示。它是DST-IIEST太阳能光伏（PV）中心的一部分[2]，该中心是由科学技术部（DST）[3]，（印度政府）与绿色能源与传感器系统卓越中心（CEGESS）[4]，（印度工程科学与技术学院（IIEST），希布尔）共同发起的项目。图5 a。该屋顶太阳能光伏电站的装机容量为10千瓦$_p$，测量功率输出，SRRA中心则实时采集气象参数。采集的数据按批次存储在集中监测站，时间跨度为7个月（2019年8月至2020年2月）。图3 展示了精心挑选的四个典型日期的发电功率（单位：瓦特）样本曲线，这些曲线按1分钟间隔记录，用于呈现曲线的典型变化。从图中可以观察到，白天6:00至18:00之间，曲线呈桌面状，而代表最大功率值的曲线顶部出现在正午左右，即约11:00至13:00。此外，图4 中可以观察到一些随机抖动

[1] http://dst-iiestsolarhub.org.in/about_SRRA.php [2] http://dst-iiestsolarhub.org.in/about_DST_IIEST_solar_hub.php[3] https://dst.gov.in/[4] https://old www.iiests.ac.in/index.php/home-g

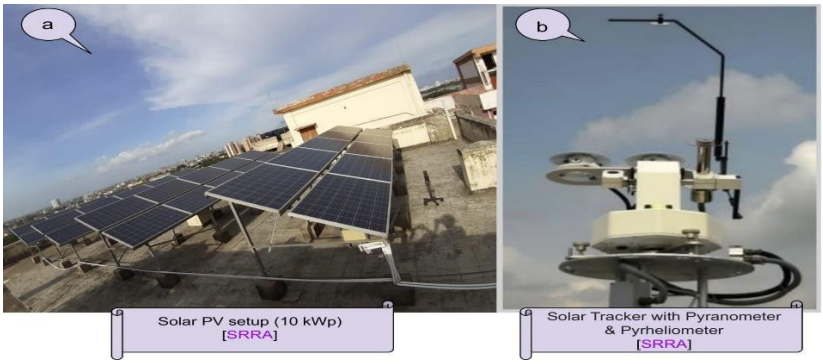


图 5. SRRA（印度东部西孟加拉邦）的标准测量系统及用于数据收集/整理的实际实验设置图像。

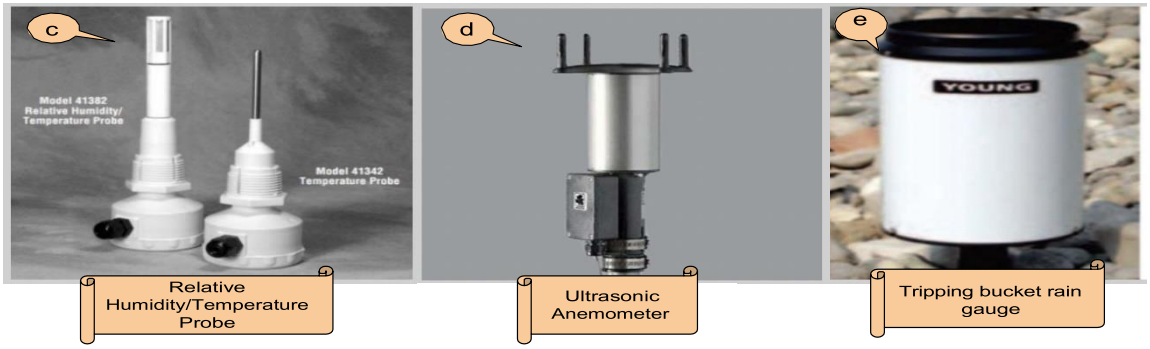


图 6. SRRA 中用于数据收集/整理和预处理阶段的气象测量仪器图像。

，这可能是云的形成导致太阳能发电受阻的结果。原始数据集包含286187个数据点和40个属性，其中包括目标属性功率。表2 介绍了目标数据集中的各项属性。

表 2. 所用气象特征说明。

参数	所用仪器	规格	制造商 和型号
RH_Avg	相对湿度探头 [图6 c]　相对湿度：(41382) 测量范围：0-100 %相对湿度 20 摄氏度时的准确度：±2 %相对湿度， 稳定性：每年优于 ±1 %相对湿度 响应时间：10 秒（不带滤波器） 传感器类型：Rotronic Hygromer TM 输出信号：V 选项：0-1 直流电压， L 选项：4-20 毫安		R.M. YOUNG 公司 特拉弗斯城， 密歇根州 49686 USA
降雨累积量	翻斗式雨量计 [图 6 e]	分辨率：每次翻斗 0.1 毫米 准确度：最高可达 25 毫米/小时，误差为 2% 最高可达 50 毫米/小时，误差为 3% 输出：磁簧开关（常开） 触点额定值：24VAC/DC 500毫安 工作温度：-20 至 +50摄氏度（加热型） 加热器恒温器设定值：10摄氏度 ±3摄氏度	R.M. YOUNG 公司 特拉弗斯城， 密歇根州 49686 USA
平均气温	温度探头 [图6 c]	温度：（41382, 41342） 校准测量范围： -50 至 50 摄氏度（后缀 C） -50 至 150 °F（后缀 F） 响应时间：10 秒（不含滤波器） 0 摄氏度时的准确度：±0.3 摄氏度** ±0.1 摄氏度（可选），并提供美国国家标准与技术研究院可溯源校准 传感器类型：铂电阻温度探测器 输出信号：V 选项：0-1 直流电压，L 选项：4-20 毫安，4 线制 电阻温度探测器（仅限 41342）	R.M. YOUNG 公司 特拉弗斯城， 密歇根州 49686 USA

平均风速	超声波风速仪 [图6 d]	测量范围：0.01 – 75 米/秒 准确度：对于 >5毫秒的测量值，误差为 ±2% 分辨率：0.01米/秒	Biral，布里斯托尔 工业与研究 联合有限公司
平均风向	超声波风速仪 [图6 d]	测量范围：0-360° 分辨率：1° 准确度：±10	Biral，布里斯托尔 工业与研究 联合有限公司
DP_Avg	湿度计	测量范围：0 -100 % 相对湿度- 30 -100 摄氏度 分辨率：0.01% 相对湿度 0.01 摄氏度 准确度：±2.0 % 相对湿度，25 摄氏度时 / 20 ... 80 % 相对湿度 其他测量区域：±2,5 % 相对湿度 测量速率：2,5 / 秒 显示屏：双液晶显示屏，4.5 位（背光） 响应时间：相对湿度 约10秒	PCE Instruments UK Limited
GR_Avg	总辐射表 [图5 b]	灵敏度：约 9 µV/Wm-2。 阻抗：约 650 欧姆。 温度依赖性：在环境温度范围内为 ±1%- 20 至 +40摄氏度。线性度：从 0 到 2800 Wm 时为 ±0.5% $^{-2}$. 响应时间：1 秒（1/e 信号）。 余弦响应：在天顶角 0-70° 范围内，以 ±1% 归一化； ±3% 70-80° 天顶角。	Eppley实验室，罗德岛屿，美国。PSP
DiffR_Avg	天空扫描辐射计 [图7 g]	测量原理：多波段滤光辐射计 紫外、可见光和近红外辐射计 探测器：紫外增强型硅光电二极管 波长：315、400、500、675、870、940 和 1020 nm 波长精度：2 nm	Kipp & Zonen B.V Delftechpark 36, 2628 XH 代尔夫特 P.O. 邮政信箱 507，2600 AM 代尔夫特 荷兰 荷兰
AP_QNH_Avg	数字气压计	压力范围：500 至 1100 百帕 工作温度：-40 至 +60摄氏度 数字精度*0.2 百帕（25摄氏度）0.3 百帕（-40 至 +60摄氏度） 模拟精度** 模拟压力范围的 0.05% 模拟温度依赖性 模拟压力范围的 0.0017%/摄氏度（参考温度为 25摄氏度） 更新速率：最高 1.8 赫兹 串行输出：全双工 RS-232，9600 波特 轮询或连续 ASCII 文本、NMEA 半双工 RS-485（仅限 61302L） 模拟输出：0 至 5000 mV、0 至 2500 mV（61302V） 4 至 20 毫安（61302L） 分辨率：串行 0.01 百帕；模拟扫描值的 0.025%	R.M. YOUNG 公司 特拉弗斯城， 密歇根州 49686 USA
AP_QFE_Avg	数字气压计	压力范围：500 至 1100 百帕 工作温度：-40 至 +60摄氏度 数字精度*0.2 百帕 (25摄氏度)0.3 百帕 (-40 至 +60摄氏度) 模拟精度** 模拟压力范围的 0.05% 模拟温度依赖性 模拟压力范围的 0.0017%/ 摄氏度（参考温度 25摄氏度） 更新速率：最高 1.8 赫兹 串行输出：全双工 RS-232，9600 波特 轮询或连续 ASCII 文本，NMEA 半双工 RS-485（仅限 61302L） 模拟输出：0 至 5000 毫伏，0 至 2500 毫伏（61302V） 4 至 20 毫安（61302L） 分辨率：串行 0.01 百帕；模拟扫描的 0.025%	.M. YOUNG 公司 特拉弗斯城， 密歇根州 49686 USA
DirR_Avg	多光谱光度计 [图7 f]	分辨率：0.05 度 最大速度：10 度/秒 旋转范围：280o 执行器：步进电机（DPM60SH86 - 2008AF）	Holmarc 光电机电有限公司， 印度喀拉拉邦
HSR_Avg	宽带长波辐射表 [图 7 i 和 j]	光谱范围（总体）4.4 至 50 µm = 4400 至 50000 nm 响应时间（95%）< 18 秒 窗口加热偏移< 15 瓦/平方米 零点偏移 B< 5 瓦/平方米 灵敏度温度依赖性（-10 oC 至 +40 oC）< 3 % 工作温度范围 -40 至 +80 摄氏度 视场角 150 ° 非线性< 1 % 模拟输出（-A / -V 版本）4 至 20 毫安 / 0 至 1 伏 数字输出2 线制 RS-485	OTT HydroMet B.V. 代尔夫特 - 荷兰 荷兰
ISR_Avg	宽带长波辐射表 [图 7 i 和 j]	光谱范围（总范围）4.4 至 50 µm = 4400 至 50000 nm 响应时间（95%）< 18 秒 窗口加热偏移< 15 瓦/平方米 零点偏移 B< 5 瓦/平方米	OTT HydroMet B.V. 代尔夫特 - 该 荷兰

		灵敏度温度依赖性（-10 oC 至 +40 oC）< 3 % 工作温度范围 -40 至 +80 摄氏度 视场角 150 ° 非线性 < 1 % 模拟输出（-A / -V 版本）4 至 20 毫安 / 0 至 1 V 数字输出 2 线制 RS-485	
SA_Avg	方位罗盘	分辨率：1Â° 准确度：0.5Â°	Brunton International 有限责任公司，里弗顿， 怀俄明州，美国。
DTR_Avg	紫外辐射计 [图7 g]	光谱范围（整体）280 至 400 nm 灵敏度300 - 500 µV/W/m2 温度响应< 每摄氏度-0.3% 响应时间（95%）< 1 秒 非线性（0 至 100 W/m2）< 1 % 最大UVA/UVB辐照度400 W/m2 工作温度范围-40 至 +80 摄氏度 方向响应（在 100 W/m2 光束下最高可达 70°）< 5 % 视场角180 °	OTT HydroMet 美国公司 斯特林 - 美国
GR_Avg	总辐射表 [图5 b]	灵敏度：约 9 µV/Wm-2。 阻抗：约 650 欧姆。 温度依赖性：在环境温度范围内为 ±1%- 20 至 +40摄氏度。线性度：±0.5%，范围为 0 至 2800 Wm-2。 响应时间：1 秒（1/e 信号）。 余弦响应：在 0-70° 天顶角范围内归一化为 ±1%； ±3%（天顶角 70-80°）。	Eppley实验室，罗德 岛屿，美国。PSP
DiffE_Avg	天空扫描辐射计 [图7 g]	测量原理：多波段滤光辐射计、紫外、可见光 和近红外辐射计 探测器：紫外增强型硅光电二极管 波长：315、400、500、675、870、940 和 1020 nm 波长精度：2 nm	Kipp & Zonen B.V Delftechpark 36, 2628 XH 代尔夫特 邮政信箱 邮政信箱507，2600 AM 代尔夫特 荷兰 荷兰
DirE_Avg	多光谱光度计 [图7 f]	测量原理：多波段滤光辐射计 紫外、可见光 和近红外辐射计 探测器：紫外增强型硅光电二极管 波长：315、400、500、675、870、940 和 1020 nm 波长精度：2 nm	Kipp & Zonen B.V Delftechpark 36, 2628 XH 代尔夫特邮政信箱 邮政信箱 507，2600 AM 代尔夫特，荷兰 荷兰

3.2 数据分析与预处理

本研究所使用的未经处理的传感器现场数据集直接从源头采集而来，存在与传感器日志错误相关的问题。数据记录中的不确定性可能源于仪器选择、仪器状态、仪器校准、环境、观测与读数以及测试规划。因此，在将数据输入机器学习模型进行训练之前，有必要进行数据清洗和预处理。本节介绍了所实施的各种数据挖掘技术的详细信息，这些技术大体上归类于数据清洗，同时还开展了数据分布分析，以便从数据集中提取有意义的模式。

3.2.1 通过缺失值填补进行数据清洗：当日志中包含缺失数据时（由于系统停机等各种问题导致采集阶段发生系统故障时便会出现这种情况），直接将其丢弃虽然简单，却可能导致数据量减少并删除重要的数据点。因此，我们采用了一种称为缺失值填补的技术 [34]，使用均值填补、回归填补、热卡填补、冷卡填补或中位数填补来替换缺失值。根据数据的性质，我们采用了均值填补。

3.2.2 特征选择与降维：训练特征数量增加，会使预测建模任务更具挑战性，这种现象称为维数灾难。降维是指减少数据集中特征数量的技术。数据缩减/降维通常沿两个方向进行，即按行进行数据样本缩减，以及按列进行数据变量缩减。由于印度的日照时长约为 8 至 10 小时，且大量目标数据为零，从而产生偏差。通过按行数据采样可以消除这种偏差；具体做法是删除目标变量（发电量）低于阈值（500 瓦特）的行。这样一来，数据点数量已从 286187 个降至

92684 个，包含 42 个气象特征。按列数据变量降维的三种主要方法是：i) 利用领域知识直接选择感兴趣的变量；ii) 使用统计特征选择方法，选出用于进一步分析的重要变量；iii) 利用皮尔逊相关性找出高度相关的特征，并从中删除一个。按列进行数据降维时，仅保留气象特征的平均值读数，舍弃瞬时值、最大值和最小值数据读数。某些传感器数据，如“太阳直射辐射表错误”和“Geotrac3K 状态”，也予以舍弃，因为它们不会影响此应用场景。因此，列维度已从 40 个特征降至 18 个。随后，采用一种名为套索回归的 L1 正则化方法，找出最重要的特征。套索方法通过收缩回归系数对模型参数进行正则化，使其中一些系数降为零。收缩后进行特征选择，选取所有非零值用于模型。该方法有助于最大限度地减少统计模型中常见的预测误差。为了提高特征选择的可信度，训练时采用带有 L1 和 L2 惩罚项的弹性网络回归；在某些问题上，其表现优于仅采用其中一种惩罚项的模型。图 8 展示了使用上述两种方法绘制的重要性图。表 3 列出了两种正则化方法的系数。

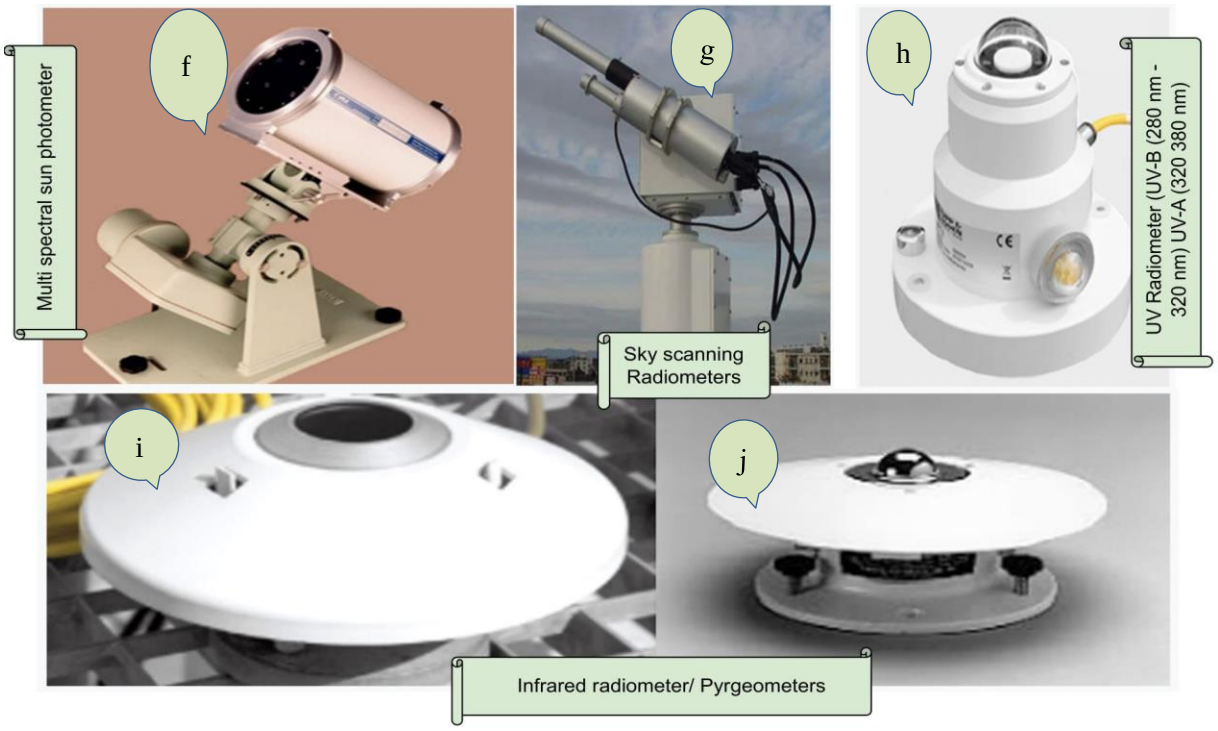
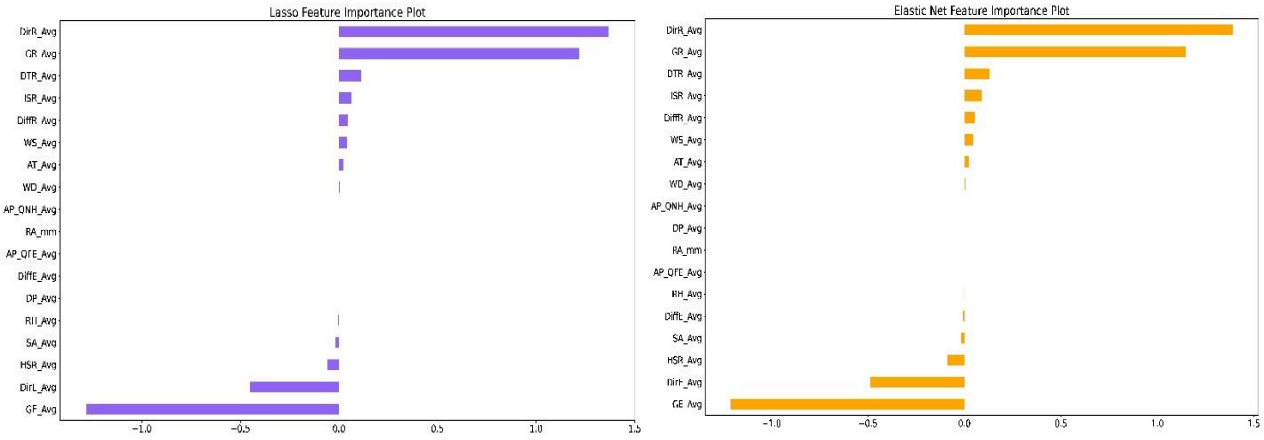


图 7. SRRA 提供的其他高级测量设备图像，这些设备用于我们的数据收集/整理流程。



(a) Lasso 特征重要性　　(b) 弹性网络特征重要性

图 8. 特征选择的特征重要性图。

因此，常见的最不重要特征‘AP_QFE_Avg’、‘RA_mm’、‘DP_Avg’和‘AP_QNH_Avg’被剔除。尽管理论上降雨是云量的一种间接衡量指标，进而会影响发电，但在本例中发现降雨累积量的特征重要性较低。为交叉验证这一判断，我们检查了 RA_mm 的方差，发现其极低，仅为 3.36 x 10^-5。因此，降雨累积量传感器可能存在数据收集不准确的情况。由于这会影响预测的准确度，因此将其舍弃。进一步分析了各属性之间的皮尔逊相关性 [35]，方法是将皮尔逊系数热图可视化，如图 9(a) 所示。根据图中结果，使用 0.95 的系数阈值筛选出高度相关的属性。一些高度相关的属性对包括：HSR_Avg 和 ISR_Avg、GR_Avg 和 GE_Avg、DiffR_Avg 和 DiffE_Avg、DirR_Avg 和 DirE_Avg、DTR_Avg 和 DirR_Avg 等。根据图 9(a) 中的观测结果，为构建更优的机器学习算法输入特征空间，这些属性对已从数据集中移除。最终的特征集包括‘RH_Avg’、‘AT_Avg’、‘WS_Avg’、‘WD_Avg’、‘GR_Avg’、‘DiffR_Avg’、‘DirR_Avg’、‘ISR_Avg’和‘SA_Avg’。降维后，这些输入特征的相关系数相对较低，如图 9(b) 所示。因此，维度从 18 个可训练参数减少至 9 个。

表 3. 特征选择的特征重要性表。

序号	特征	Lasso 重要性	弹性网络重要性
1	GE_Avg	-1.281	-1.212
2	DirE_Avg	-0.449	-0.491
3	RH_Avg	-0.002	-0.002
4	SA_Avg	-0.017	-0.018
5	平均风向	0.005	0.005
6	DP_Avg	0.000	0.000
7	平均场压	0.000	0.000
8	HSR_Avg	-0.060	-0.0885
9	DTR_Avg	0.111	0.130
10	降雨累积量	0.000	0.000
11	DiffE_Avg	0.000	-0.007
12	平均气温	0.021	0.023
13	AP_QNH_Avg	0.000	0.000
14	平均风速	0.043	0.043
15	ISR_Avg	0.064	0.089
16	DiffR_Avg	0.046	0.054
17	DirR_Avg	1.369	1.391
18	GR_Avg	1.219	1.147

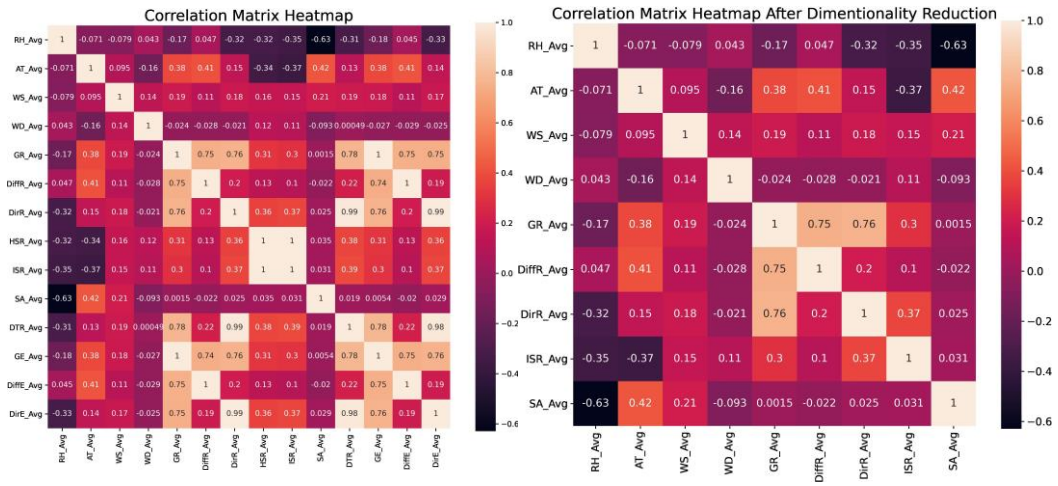
一些高度相关的属性对包括：HSR_Avg 和 ISR_Avg、GR_Avg 和 GE_Avg、DiffR_Avg 和 DiffE_Avg、DirR_Avg 和 DirE_Avg、DTR_Avg 和 DirR_Avg 等。根据图 9(a) 中的观测结果，为构建更优的机器学习算法输入特征空间，这些属性对已从数据集中移除。最终特征集包括'RH_Avg'、'AT_Avg'、'WS_Avg'、'WD_Avg'、'GR_Avg'、'DiffR_Avg'、'DirR_Avg'、'ISR_Avg'和'SA_Avg'。降维后，这些输入特征的相关系数相对较低，如图 9(b) 所示。因此，维度从 18 个可训练参数减少至 9 个。

3.2.3 核密度估计（KDE）分析：本小节中，核密度估计是使用核函数估计未知概率密度函数的过程。核密度估计是一种用于估计连续随机变量的概率密度函数（PDF）的非参数方法。参见图10 可观察到以下情况：

i) '功率'、'RH_Avg'、'WS_Avg'、'WD_Avg'、'DiffR_Avg'、'DirR_Avg'和'GR_Avg'均呈单峰分布。ii) ISR_Avg、SA_Avg呈双峰分布。这表明双峰分布中可能存在两种不同的组，因此回归模型必须能够分别处理这些分布。

3.3 气象参数对太阳能光伏发电的影响

太阳能电站的性能取决于气象参数，因此其发电具有间歇性。在本研究中，考虑了直接影响光伏发电的主要环境参数。为验证我们对特征及其在预测中重要性的分析，结果显示，使用某些环境参数通过[36] 计算得到的发电量接近实际发电量。各种



(a) 特征降维前的相关性图　　(b) 降维后的相关性图

图 9. 用于降维的皮尔逊相关系数图。

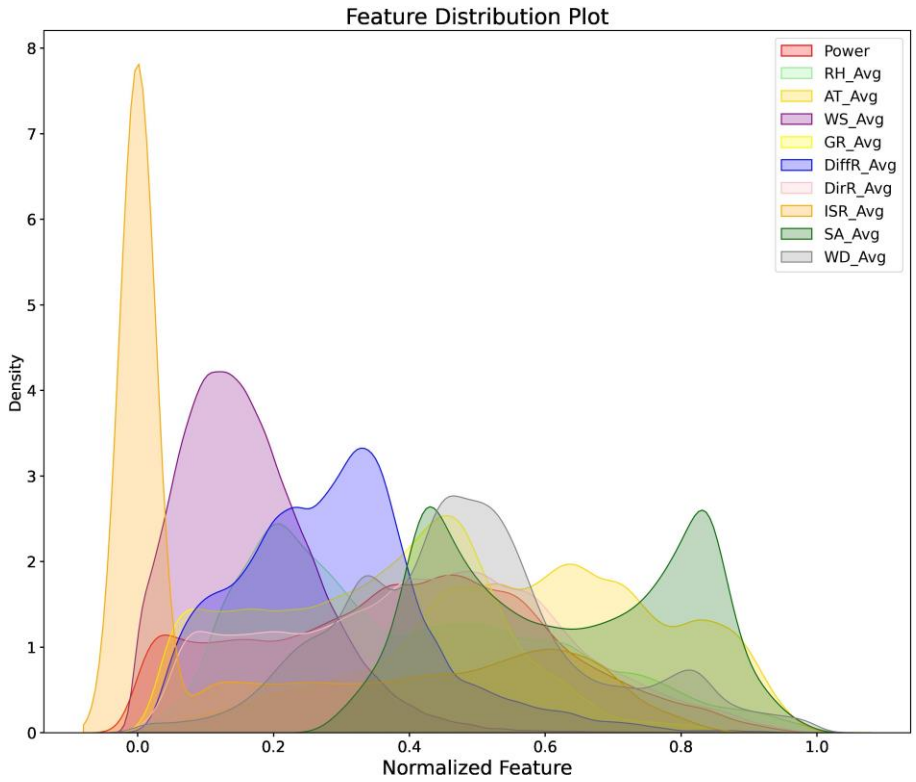


图 10. 数据分布图。

用于计算功率的方程式为方程 1- 4。所用常数的定义和值见表 4 和表 5。表 6中列出了使用 DR_Avg 和 WS_Avg 计算得到的发电量与实际功率的 20 个观测数据的随机样本范围。采用这种方法 [36] 相对于实际发电量，观测到的均方根误差（RMSE）为 799.25。因此，环境参数的选择会严重影响结果。但研究表明，DR_Avg 和 WS_Avg 并不是唯一影响结果的特征，加入少量其他参数可以改善我们的预测。

$$S_h = S_i \sin \propto \tag{1}$$

$$S_m = S_i \sin(\propto + \beta) \tag{2}$$

$$T_i = S_m/(26.9 + (6.2 * wi)) \tag{3}$$

$$P_c = P_r S_m\left(1 + (k1 \log(S_m)) + k2 \log\left({S_m}^2\right) + k3T_i\right) + k4T_i \log(S_m) + k5T_i \log\left({S_m}^2\right) + k6{T_i}^2 \tag{4}$$

通过查阅多篇关于太阳辐射和太阳能发电预测分析的文献，我们发现，太阳方位角和环境温度参数直接影响太阳辐射测量，并最终影响太阳能发电量。图 11 展示了三维图，呈现发电量如何随环境温度和太阳方位角测量值的平均归一化值而变化。这是一幅三维平面上的散点图，颜色从浅绿色渐变至红色。由图可见，大多数数据点集中在方位角和温度图的中间区域。最低功率出现在高方位角、低温度以及低方位角、低温度的情况下。两个参数读数都较高时，功率数值似乎达到最大。

表 4. 符号命名。

	参数描述	
1	α	仰角
2	β	倾斜角
3	δ	赤纬角
4	ϕ	SPV站点纬度
5	S_t	垂直于太阳入射的太阳辐射（DR_Avg）
6	S_m	倾斜表面上的太阳辐射
7	S_h	水平表面上的太阳辐射
8	P_c	组件计算功率
9	w_i	面板处的入射风速（WS_Avg）
10	T_i	随风速变化的光伏组件温度
11	$k1-k6$	光伏组件常数
12	P_r	填充率

表 5. 功率方程参数值。

参数	数值
Area	21.77 m^2
效率	0.19
P_r = 效率*面积	12.36
ϕ	26.2°
β	26 °
$\alpha = 90 - \phi + \beta$	89.8 ◦
k1	-0.06689
k2	-0.012844
k3	-0.002262
k4	0.0002276
k5	0.000159
k6	-0.000006

表 6. 实际功率与计算功率。

	时间戳	实际功率	计算功率
1	2019年1月5日 09:17:00	3355.04	2808.03
2	2019年1月5日 09:18:00	3346.00	2766.82
3	2019年1月5日 09:19:00	3284.97	2768.39
4	2019年1月5日 09:20:00	3426.53	2800.90
5	2019年1月5日 09:21:00	3324.33	2835.44
6	2019年1月5日 09:22:00	3480.50	2847.67
7	2019年1月5日 09:23:00	3432.44	2894.21
8	2019年1月5日 09:24:00	3406.01	2847.34
9	2019年1月5日 09:25:00	3487.33	2869.64
10	2019年1月5日 09:26:00	3554.86	2918.95
11	2019年1月5日 09:27:00	3637.89	2975.86
12	2019年1月5日 09:28:00	3606.71	2979.75
13	2019年1月5日 09:29:00	3602.64	3034.08
14	2019年1月5日 09:30:00	3665.04	3038.76
15	2019年1月5日 09:31:00	3678.98	3074.13
16	2019年1月5日 09:32:00	3868.26	3165.09
17	2019年1月5日 09:33:00	3910.53	3216.70
18	2019年1月5日 09:34:00	3906.38	3210.01
19	2019年1月5日 09:35:00	3870.09	3250.37
20	2019年1月5日 09:36:00	3983.11	3261.65

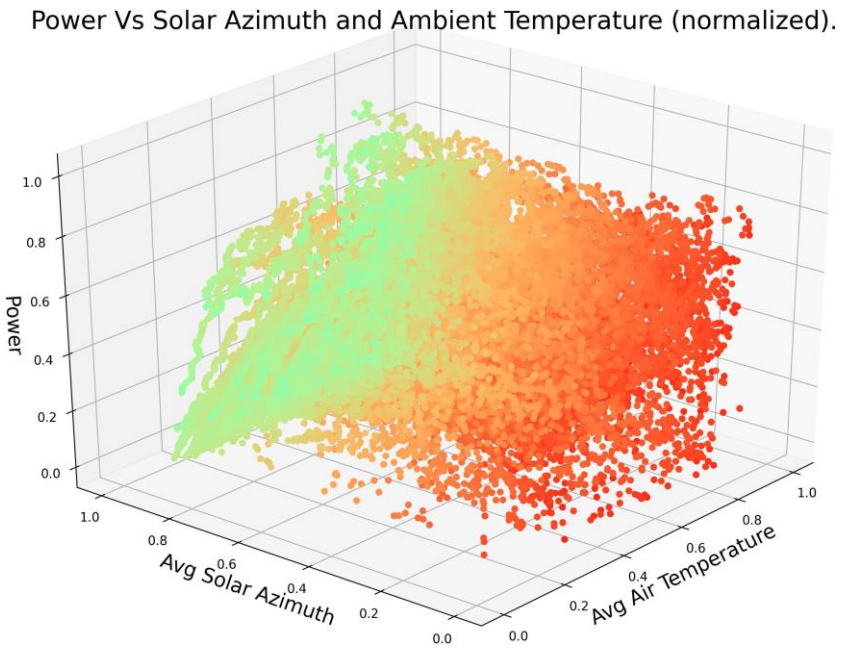


Fig. 11. 功率测量值随太阳方位角和环境温度测量值变化的可视化 .

4 集成学习模型探索、实验与结果分析

4.1 型号探索

正如 Buhlmann 等人所解释的，[37], 在估计如下形式的函数时：

$$g : \mathbb{R}^d \leftarrow \mathbb{R} \tag{5}$$

其中，d 为预测变量的维度，集成函数估计表示为：

$$\hat{g}_{ens}(\cdot) = \sum_{k=1}^{M} c_k \hat{g}_k(\cdot) \tag{6}$$

其中，M 是使用重新加权的输入数据集得到的函数估计数量。本文实现了以下集成学习方法：

4.1.1 装袋法：对于给定数量的基过程，基于装袋法的平滑操作可降低方差，进而降低均方误差。若 $X_i \in \mathrm{R}^d$ 是 d维预测变量，且 $Y_i \in \mathrm{R}$，其中 $i \in (1, n)$ 对应 n 个数据点，则函数估计器可表示为：

$$\hat{g}(\cdot) = h_n\big((X_1, Y_1), (X_2, Y_2), \ldots . (X_n, Y_n)\big)(\cdot) : \mathbb{R}^d \leftarrow \mathbb{R} \tag{7}$$

装袋估计器可表示如下：

$$\hat{g}_{Bag}(\cdot) = \mathbb{E}^*[\hat{g}*(\cdot)] \tag{8}$$

随机森林回归：Breiman等人提出的随机森林[38], 是一种基于树的集成模型，其中每棵树的结构都取决于一组随机变量。这些表示为 $\mathrm{h}\jmath(X, \Theta_j)$ 的树称为基学习器。对于一组特定的数据 $D = (x_1, y_1)、(x_N, y_N)$，以及 Θ_j 的一个特定实现 θ_j ，拟合后的树表示为 $\hat{\mathrm{h}}_j(x, \theta_j, D)$。考虑联合分布 $\mathrm{P_{XY}}(X, Y)$，其中 X 和 Y 分别是包含预测变量和响应变量的向量，则预测函数 $f(X)$ 由损失函数 $L(Y, f(X))$ 确定，该函数试图使损失的期望值 $\mathbb{E}_{\mathbb{XY}}(L(Y, f(X)))$ 最小化，因此回归函数可表示为 $f(X) = \mathbb{E}(Y|X = x)$。在随机森林的回归任务中，构成集成模型 f 的基学习器 $\mathrm{h}_1(x)$、...$\mathrm{h}_j(x)$ 取平均值，得到

$$f(x) = 1/J \sum_{j=1}^{J} h_j(x)$$

极端随机树回归： Geurts 等人。 [39] 提出了一种以自顶向下的方式构建未剪枝决策树集成的算法。其主要区别特征是，完全通过随机选择切分点来划分树的节点；并且，与使用自助法样本不同，该算法使用整个学习样本来生长树。极端随机树分裂算法会利用学习样本（用于构建型号的观测数据）多次进行分裂，最终生成集成模型。若干参数决定了极端随机树的性能，例如，K （每个节点随机选择的属性数量）可用于估计样本选择过程的强度，n_{min} （用于分裂节点的最小样本量）可用于估计平均输出噪声的能力，而M 则表示集成模型聚合的方差缩减能力

4.1.2 提升法： 在集成方法的提升法中，基过程 按顺序运行。提升法是一种非参数优化算法，通常被认为在回归、分类等任务中的表现优于装袋算法。基于表示为（X_i， Y_i）$\forall\ i \in (1,\ n)$ 的数据，其中n 是数据点的数量，提升算法估计函数g：$R_d \rightarrow R$，并最小化期望损失 $E[l\ \ (Y,\ g(X)], l\ (\cdot,\cdot): R \times R \rightarrow R^{+}$。这里，$Y$在回归和分类任务中分别可以是连续值和离散值

.

AdaBoost：Freund 等人 [40, 41] 提出了自适应提升算法，旨在通过持续生成分类器来降低集成的误差。该算法旨在找到一个假设，同时最小化训练样本 (x, y) 上分布 D 的误差。该算法通过在第 t_{th} 次迭代时对一组权重进行归一化，从而迭代训练学习器，使由此计算出的分布 P_t 馈送给弱基学习器。这些学习器最终生成一个在该分布下误差最小的假设 h_f。该假设可表示为：

$$h_f(x) = \begin{cases} 1, & if \ \sum_{t=1}^{T}(log\ 1/B_t)h_t(x) \geq 1/2\sum_{t=1}^{T} log\ 1/B_t \\ 0, & otherwise \end{cases}$$

其中，B 是用于更新权重向量的参数，而 h_f 是根据加权多数投票，将T个弱假设的输出组合而成的假设。

梯度提升机 (GBM)：Friedman 等人 [42] 提出了随机梯度提升算法，旨在找到一个函数 F_*(x)，将x 与y （分别为输入变量和输出变量）映射起来，使得给定联合分布 (x, y) 时的期望损失最小，其中期望损失由 $\Psi(y, F(x))$ 表示。最小化函数可写为：

$$F_*(x) = \underset{F(x)}{\arg\min} E_{(y,x)}\Psi(y, F(x)) \tag{9}$$

此类算法采用的主要概念是构建与上述损失函数的负梯度具有最大相关性的新弱基学习器。随后，这些新模型或基学习器依次叠加，以生成最终的集成。通过在集成上计算误差或损失来训练每个学习器。梯度提升机以配置灵活著称，因为可以从多种损失函数中选择一种用于训练集成。因此，该算法可以轻松定制，以适用于任何数据驱动任务 [43]。

XGBoost：Chen 等人提出的 XGBoost [44], 是一种可扩展的树提升稀疏性感知算法。在这种提升算法中，预测模型以加法方式训练。在第 t_{th} 次迭代中，为预测第 i_{th} 个数据实例，将 f_t 加入以最小化目标函数：$L^{(t)} = \sum_{i=1}^{n} l(y_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$，$w$其中，$\Omega$ 用于惩罚 XGBoost 模型的复杂度，l 是可微凸损失函数，用于衡量预测值 $\hat{y}_i$ 与目标值 y_i 之间的差异，而 x_i 是输入。

LightGBM：LightGBM [45] 是一个基于梯度提升决策树（GBDT）的梯度提升框架，采用基于梯度的单边采样（GOSS）和互斥特征捆绑（EFB），旨在提高算法效率，同时减少内存使用量和训练时间。GBDT 通过决策树学习一个将输入空间 X_s 映射到梯度空间 G_s 的函数。对于每个特征 i，GBDT 在计算最大信息增益 $V_j(d^*_j)$时，确定d^*_j =arg $\max_d V_j(d)$，其中 V 表示在决策树的固定节点处，对数据点 d 上的特征 j 进行分裂所获得的方差增益。然而，在 GOSS 中，会在较小的数据实例子集上估算$V_j(d)$，从而大幅降低计算成本。

HGBoost：基于直方图的算法将未排序的原始数据值映射为离散值，以便为直方图建立索引，从而将浮点数据类型转换为字符形式的分箱数据。对于 XGBoost，原始数据值经过预排序，因此需要同时保存数据及其索引。对于相同的数据集，基于直方图的提升算法 [46] 的内存成本比 XGBoost 低 8 倍，且训练时的计算成本也更低。HGBoost回归估计器原生支持数据集中的缺失值。该方法借鉴了 LightGBM；LightGBM 最初提出了直方图算法的理念。

4.1.3 投票平均回归器集成：An等人 [47] 为回归集成提出了投票平均方案。对于回归函数 $f = (x_1; y_1)$、$(x_2; y_2)$、……$(x_t; y_t)$，有 n 个基学习器 $R_i(i = 1, \dots, n)$，以及 t 个训练样本。此类集成的简单加权平均为 $\hat{y}_j = \sum_{i=1}^{n} w_i x_{ij} = \sum_{i=1}^{n} w_i R_i(x_j)$、$\forall\, j \in (1, t)$，其中 wi 是第 i 个回归器 Ri 的权重，$\hat{y}$ 是集成输出，z_{ij} 是对应于第 j 个训练样本 x_i 、 $z_{ij} = R_i\,(x_j)$ 的第 i 个回归器输出。在此基础上，投票平均集成会将回归器输出划分为 m 个类别 $C_{jk} \forall\, k$ (1, m)，并通过多数投票确定获胜类别；例如，$k*$类为获胜类别，如

C_{jk*} = C_{j1} C_{jm} $k*$ j $\hat{y}_j$ = 投票法 (, … $Averaging\ C_{jk*}$ ），即对第 - 个样本对应的第 - 类取平均，表示为 ( )。

4.1.4 堆叠回归集成：Breinman等人 [48] 提出了线性组合预测器的概念，以利用交叉验证数据，并在非负约束下使用最小二乘法。对于集成中的K 个回归器，学习集可表示为$\mathbb{L} = (y_n, x_n)$， $n = 1,2, \dots N$)，其中x 和y 分别为输入向量和预测向量。通过将注意力限制在形式为$v(x) = \sum_k a_k\, v_k(x)$的组合上，可简化预测向量的计算问题；对于$\mathbb{L}$，一种方法是取$\alpha_k$ ，使表达式$\sum (y_n - \sum_k a_k\, v_k(x_n))^2$最小。$_n$

早期研究已通过实验表明，堆叠多个回归模型可以降低错误率。

4.2实验与评估

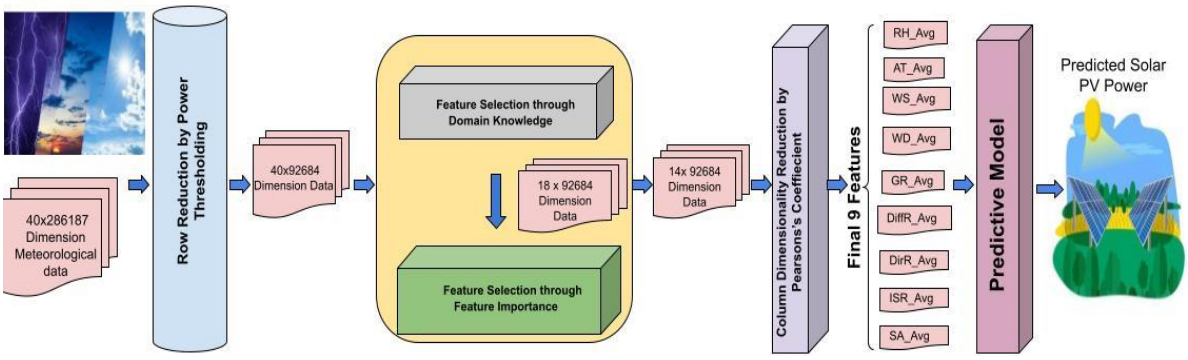


图12. 本文实现的所有学习模型均采用的通用预测框架。

本文本节介绍训练机器学习模型所涉及的各个步骤，以及用于比较的评估指标

t各种机器学习模型已得到阐述。预测框架如图所示：1 2.

- 训练-测试集划分：

下一步是 在数据预处理和分析之后，训练不同的回归模型，使其适用于该数据集。将清理和精简后的数据划分为训练集，其中 80 % 的数据用于训练预测模型。其余 20 % 用于测试模型。

- 评估指标：为了评估 模型的 性能，定义了一些评估指标，包括均方根误差（均方根误差）和 R平方分数（R^2）。预测框架如图12所示。

$$\text{RMSE} = \frac{\sqrt{\sum_{n=1}^{N}(x_i - \hat{x}_i)^2}}{N} \tag{10}$$

$$R^2 = \frac{SS_{regression}}{SS_{total}} \tag{11}$$

- 实验：

在所提出的预测测试平台中，SRRA数据集经过适当方法清理，并进行了有助于提升模型预测准确率的分析。这些选定特征用于训练所有传统机器学习模型和集成机器学习模型。为建立评估基准，我们训练了线性回归、弹性回归、贝叶斯岭回归和 K近邻等传统机器学习模型。我们对超参数进行了调优，以获得模型的最佳性能。相应的均方根误差图见图13(a)。所选传统模型能够预测发电量，并为下一组模型训练提供了合理的基准。

在[17], [49] 和[18] 中，作者采用集成机器学习技术，对光伏发电机的太阳辐射进行短期预测。在[49] 中，作者对用于风速/太阳辐照度预测的集成方法进行了综述和分类。他们针对NSRDB数据集的时间序列数据子集，比较了4种太阳辐照度预测集成方法，并以均方根误差和平均绝对缩放误差进行评估。尽管这些方法提出了采用集成机制，但并未对气象数据进行深入分析和处理。为验证经处理的特定区域测试数据集的效果，我们训练并测试了直方图梯度提升、XGBoost、梯度提升、轻量级梯度提升机、极端随机树、随机森林和 AdaBoost 等集成技术。超参数经过调优以获得最佳结果，详见表7。经训练的集成模型的均方根误差列于表8。此外，我们选取得分最高的4种模型（直方图梯度提升、轻量级梯度提升机、极端随机树和随机森林），并构建了用于投票法和堆叠回归器的处理流程。同样，我们也训练并测试了堆叠和投票集成算法。模型训练和测试所用时间列于表9。

所提出的测试平台使用 Python语言在一台标准笔记本电脑上开发，其规格如表 10 所示。Scikit-learn 机器学习库[5] 用于我们的实验。

[5] https://scikit-learn.org/stable/

表7. 超参数调优。

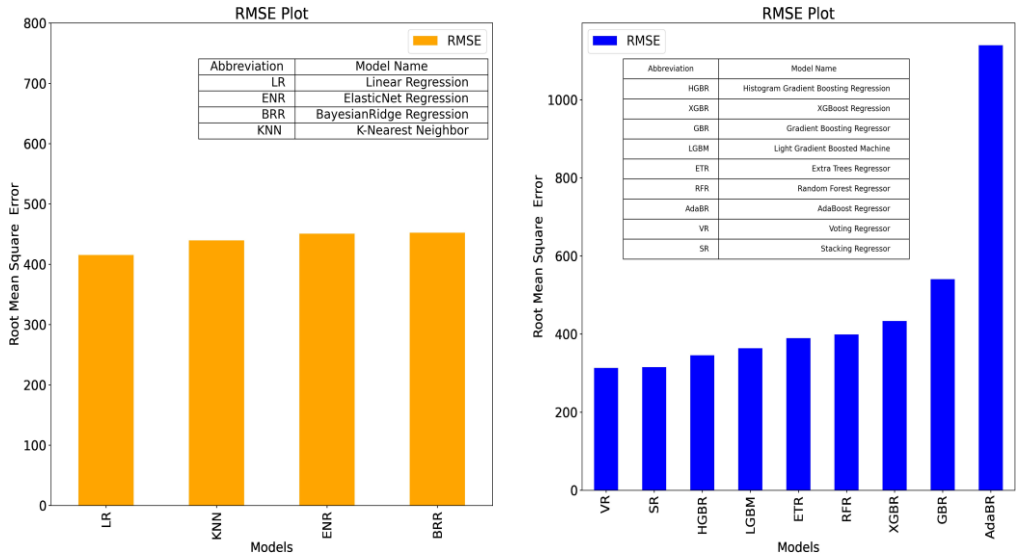
型号	机器学习模型的超参数
直方图 梯度 提升法回归	学习率=0.02，损失函数 = ‘最小_绝对 _偏差’ ，最大深度 = 40，最大迭代次数 = 750，最小叶节点样本数 = 2，随机状态 = 73
轻量级梯度提升机	目标函数 = ‘r 出口 ion’, 叶子数 = 5，学习率 = 0.05，估计器数量 = 720，最大分箱数 = 55，装袋比例 = 0.8，装袋频率 = 5，特征比例 = 0.2319，特征比例随机种子 = 9，装袋随机种子 = 9，叶节点最小数据量 = 6，叶节点最小 Hessian 和 = 11
极端随机树回归器	树的数量 = 1000，随机状态 = 73
随机森林回归器	自助法 = False，最大深度 = 60，最大特征数 = ‘sqrt’，最小样本分割数 = 4，树的数量 = 1700
XGBoost 回归	估计器数量 = 1200，学习率 = 0.05，随机状态 = 73
梯度提升回归器	max_depth = 10，max_features = ‘sqrt’，min_samples_split = 12，n_estimators = 1500
AdaBoost 回归器	学习率 = 0.03，损失函数 = ‘指数’，估计器数量 = 2300，随机状态 = 73

表8。均方根误差和 R^2 分数对比表。

序号	集成技术	均方根误差 ↓	R^2 分数 ↑
1	投票回归器	**313.07**	**0.96**
2	堆叠回归器	**314.90**	**0.96**
3	直方图梯度提升回归器	345.64	0.94
4	轻量级梯度提升机	363.64	0.94
5	极端随机树回归器	389.44	0.93
6	随机森林回归器	398.62	0.93
7	XGBoost 回归器	433.42	0.93
8	梯度提升回归器	540.34	0.91
9	AdaBoost 回归器	1139.38	0.526

表 9. 训练测试时间表。

集成技术	训练所用时间（秒）	每次预测所需平均时间（微秒）
投票回归器	192.66	115.11
堆叠回归器	981.60	114.5
随机森林回归器	517.63	567.50
极端随机树回归器	157.91	565.30
AdaBoost 回归器	229.97	226.00
梯度提升回归器	310.02	2.98
直方图梯度提升回归器	10.17	5.12
XGBoost 回归器	48.16	3.33
轻量级梯度提升机	1.26	2.15



(a) 标准机器学习回归对比。　　(b) 集成技术均方根误差对比。

图 13. 模型与均方根误差图。

表 10. 所用的系统配置。

处理器	速度	内存	**OS**
Ryzen 7 4800 CPU	2.90 吉赫兹	16 吉字节	Windows 10 64 位

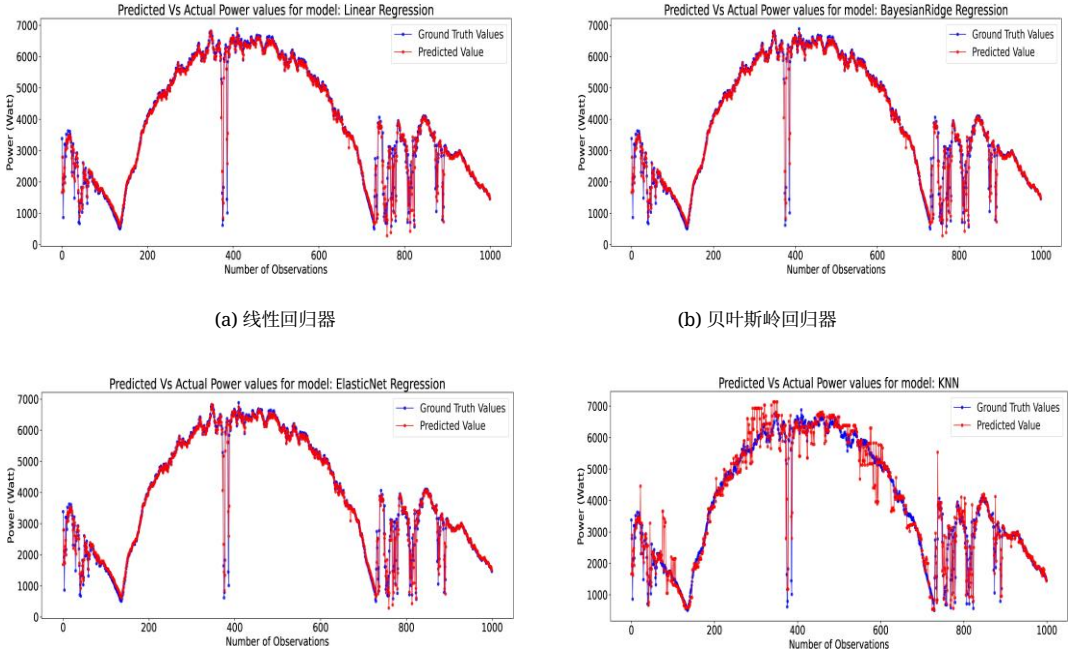
4.3 结果分析和未来讨论

拟议工　　作的各个方面、结果和成果已在本节中讨论并总结。

4.3.1　测试平台与集成技术分析。线性回归、弹性网络回归、贝叶斯岭回归和 K近邻等传统机器学习模型为预测准确率设定了良好的基准，见图 13(a)。从结果中可以看出，线性回归的表现最佳，均方根误差约为 415。本研究考察了易于实现且误差处于可接受范围内的集成机器学习算法。从图 13(b)中可以看出，除梯度提升、XGBoost 和 AdaBoost回归器外，大多数集成机器学习算法的表现都优于线性回归设定的基准。分析还发现，集成机器学习算法大多低估了功率输出，因此从电厂建设的角度来看，有助于获得预期发电量的最小值。根据图 13(b)，投票法和堆叠回归器的预测准确率最高，决定系数为 0.96，均方根误差分别为 313.07 和 314.90。根据所采用的所有误差指标，投票法和堆叠回归器更适用于印度东部地区。模型比较采用了 20 组准确率观测值的平均值。

为分析预测图，我们随机抽取了数据集中的一个时间区间，并将实际功率和预测功率随观测次数变化的情况表示为散点图（图 14 和 15 ）。以同一时间区间为例进行比较

所有传统机器学习算法和集成机器学习算法。这有助于我们直观地了解预测功率与实际功率之间偏差的特征。参见图 14(a)、14(b) 和 14(c) 可知，线性回归、贝叶斯岭回归和弹性网络回归在预测实际功率方面表现良好，但K近邻回归14(d) 的实际观测值与预测观测值之间存在较大差异。由图 15(a)、图 15(b) 以及图 13(b) 中的准确度可以清楚看出，投票法和堆叠算法的表现远优于传统机器学习算法，而图 15(h) 中的 AdaBoost 回归器在所有其他算法中表现最差。图 16 和 17 展示了多天内实际功率和预测功率随观测数量变化的结果。

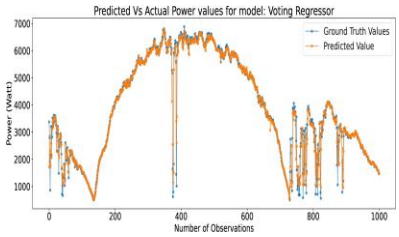


(a) 线性回归器　　(b) 贝叶斯岭回归器

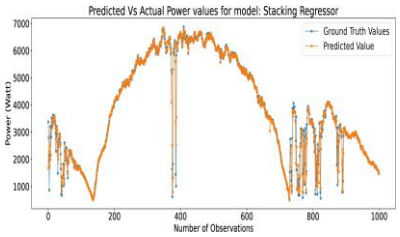
(c) 弹性网络回归器　　(d) K近邻回归器

图14. 机器学习模型在不同观测数量下实际能耗与预测能耗的随机样本图。

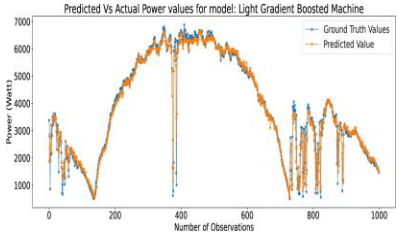
4.3.2 未来展望与研究方向。在本研究中，作者对大量集成学习方法进行了广泛调查和验证；这些方法被认为对动态天气条件下太阳能电站的优化非常有用。目前，研究团队正进一步关注其他机器学习算法（迁移学习、元学习、领域自适应、强化学习），并致力于开发新的算法和损失函数，以优化算法性能。当前研究成果还可用于估算区域性大型太阳能光伏电站的投资回报率（ROI）。



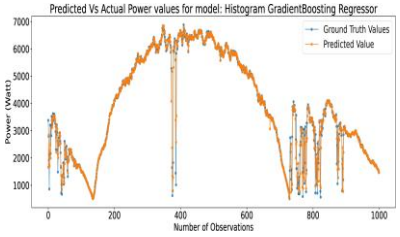
(a) 投票回归器



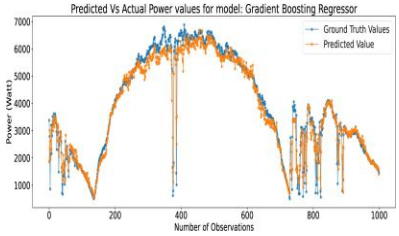
(b) 堆叠回归器

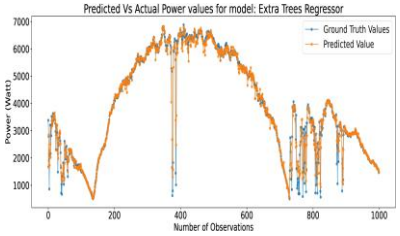


(c) 轻量级梯度提升机

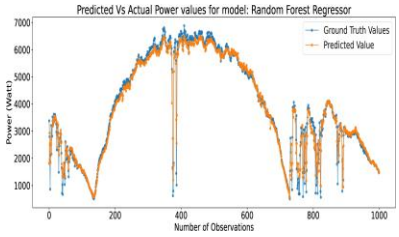


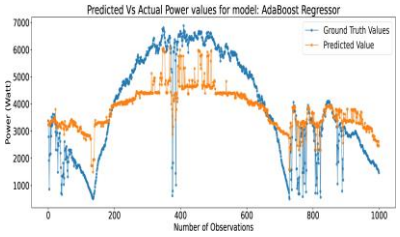
(d) 直方图梯度提升回归器



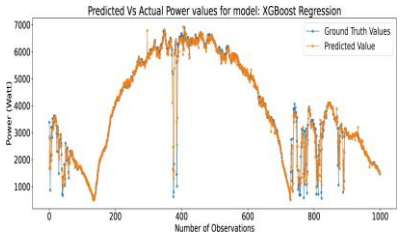


(e) 梯度提升回归器(f) 极端随机树回归器





(g) 随机森林回归器 (h) AdaBoost 回归器



(i) XGBoost 回归器

图15. 随机样本图：实际能耗与预测能耗随观测数量的变化。

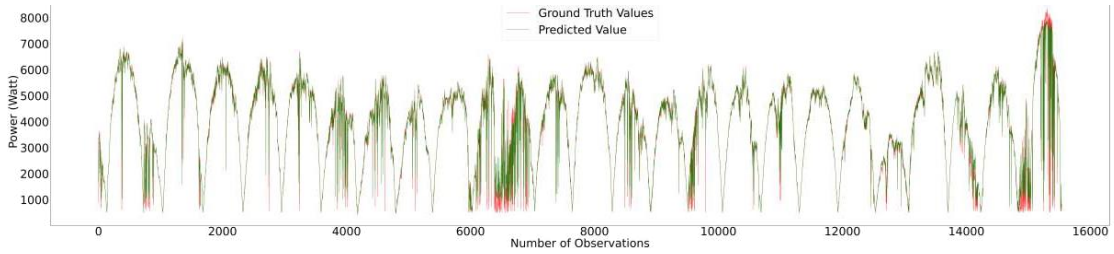


图 16. 投票回归器在数天内的实际功率和预测功率与观测数量的关系。

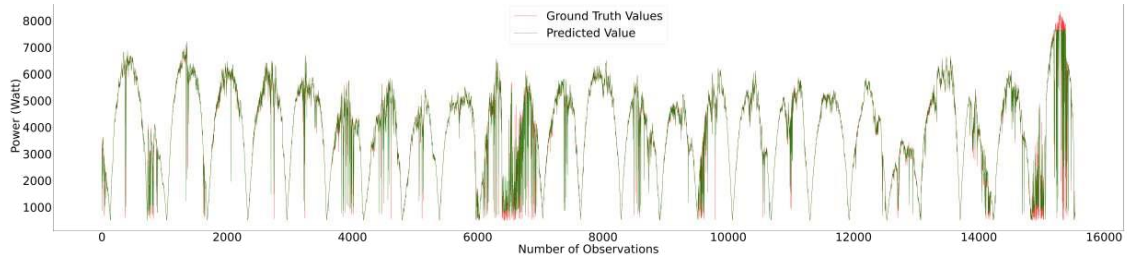


图 17. 堆叠回归器在数天内的实际功率和预测功率与观测数量的关系。

5结论

本文比较了多种集成机器学习算法，以预测气象数据影响下的太阳能发电。尽管已有不少研究报告采用传统推演方法、机器学习方法或基于深度学习的框架进行太阳能发电预测，但目前尚无针对区域太阳能发电数据的综合案例研究，提出从数据准备到机器学习模型性能比较的端到端解决方案。为确保在训练预测模型时考虑多种可能直接影响气象参数、进而影响目标区域太阳能发电的地理和地形因素，这一点至关重要。本文所展示的案例研究报告可供印度东部地区的新建太阳能项目参考，通过比较分析各种集成机器学习模型，选择最合适的预测模型。本文设计了一个学习模型测试平台，用于收集、整理和分析数据集，并根据气象参数选择合适的学习模型。该框架还支持对数据集进行特征选择和降维，以降低学习模型的空间和时间复杂度。本研究所得预测结果表明，投票法和堆叠算法的性能准确度最佳（均方根误差分别为313.07和314.9，决定系数分别为0.96和0.96）。所提出的基于集成机器学习的预测方法具有通用性，因此在气象条件相似的情况下，也可用于预测大型太阳能光伏电站的性能。后续研究可将本文工作扩展至已投运和新建太阳能电站的技术经济分析（如电站效率、投资回报率），并考虑采用各种高效深度学习算法。还可进一步研究如何优化机器学习算法性能并设计更先进的预测算法。

6 致谢

作者感谢 Hiranmay Saha 教授提供 SRRA Centre 的设施，以获取真实天气数据；同时感谢其提供安装在 CEGESS、IIEST 希布尔校区楼顶的太阳能光伏电站，该电站由印度政府可再生能源部（MNRE）资助。本研究工作在比拉尼比特斯海得拉巴校区完成。

参考文献

[1] 美国能源信息署。https://www.eia.gov/outlooks/steo/，2022年。[2] 印度品牌权益基金会。可再生能源。https://www.ibef.org/industry/renewable-energy.aspx，2021年。[3] Jan Kleissl。太阳能预测与资源评估。学术出版社，2013年。[4] Hassan Ismail Fawaz、Germain Forestier、Jonathan Weber、Lhassane Idoumghar 和 Pierre-Alain Muller。用于时间序列分类的深度学习：综述。《数据挖掘与知识发现》，33(4):917–963，2019年，doi：https://doi.org/10.1007/s10618-019-00619-1。[5] Emilio Pérez、Javier Pérez、Jorge Segarra-Tamarit 和 Hector Beltran。利用基于卫星的估算，对光伏电站邻近区域的太阳辐照度进行日内预测的深度学习模型。《太阳能》，218:652–660，2021年，doi：10.1016/j.solener.2021.02.033。[6] Jae-Ju Song、Yoon-Su Jeong 和 Sang-Ho Lee。太阳能发电预测模型分析。《数字融合期刊》，12(3):243–248，2014年，doi：10.14400/JDC.2014.12.3.243。[7] Francisco J Rodríguez-Benítez、Clara Arbizu-Barrena、Javier Huertas-Tato、Ricardo Aler-Mur、Inés Galván-León 和 David Pozo-Vázquez。伊比利亚半岛短期太阳辐射预测系统。第1部分：模型描述与性能评估。《太阳能》，195:396–412，2020年，doi：https://doi.org/10.1016/j.solener.2019.11.028。[8] Javier Huertas-Tato、Ricardo Aler、Inés M Galván、Francisco J Rodríguez-Benítez、Clara Arbizu-Barrena 和 David Pozo-Vázquez。伊比利亚半岛短期太阳辐射预测系统。第2部分：基于机器学习的模型融合方法。《太阳能》，195:685–696，2020年，doi：https://doi.org/10.1016/j.solener.2019.11.091。[9] Fermín Rodríguez、Ainhoa Galarza、Juan C Vasquez 和 Josep M Guerrero。利用深度学习和气象参数预测光伏发电机的小时内输出功率区间，以实现智能电网控制。《能源》，239:122116，2022年，doi：https://doi.org/10.1016/j.energy.2021.122116。[10] Takeshi Watanabe、Hideaki Takenaka 和 Daisuke Nohara。利用静止卫星观测数据，对数值天气模型的地表太阳辐照度预报数据进行后处理校正的方法。《太阳能》，223:202–216，2021年，doi：https://doi.org/10.1016/j.solener.2021.05.055。[11] Hadrien Verbois、Robert Huva、Andrivo Rusydi 和 Wilfred Walsh。利用数值天气预报和统计学习预测热带地区的太阳辐照度。《太阳能》，162:265–277，2018年，doi：https://doi.org/10.1016/j.solener.2018.01.007。[12] Da Liu 和 Kun Sun。基于分类优化的随机森林太阳能发电预测。《能源》，187:115940，2019年，doi：https://doi.org/10.1016/j.energy.2019.115940。[13] Xing Luo、Dongxiao Zhang 和 Xu Zhu。结合领域知识的深度学习光伏发电预测。《能源》，225:120240，2021年，doi：https://doi.org/10.1016/j.energy.2021.120240。[14] Amir Rafati、Mahmood Joorabian、Elaheh Mashhour 和 Hamid Reza Shaker。基于数据驱动启发式方法的高维超短期太阳能预测。《能源》，219:119647，2021年，doi：https://doi.org/10.1016/j.energy.2020.119647。[15] Imane Jebli、Fatima-Zahra Belouadha、Mohammed Issam Kabbaj 和 Amine Tilioua。利用机器学习和皮尔逊相关性进行太阳能预测。《能源》，224:120109，2021年，doi：https://doi.org/10.1016/j.energy.2021.120109。[16] Herman Böök 和 Anders V Lindfors。基于数值天气预报的光伏发电预测的特定地点校正。《太阳能》，211:779–788，2020年，doi：https://doi.org/10.1016/j.solener.2020.10.024。[17] Waqas Khan、Shalika Walker 和 Wim Zeiler。利用基于深度学习的集成堆叠方法改进太阳能光伏发电预测。《能源》，240:122812，2022年，doi：https://doi.org/10.1016/j.energy.2021.122812。[18] Fermín Rodríguez、Fernando Martín、Luis Fontán 和 Ainhoa Galarza。集成机器学习和时空参数，预测超短期太阳辐射，以计算光伏发电机的输出功率。《能源》，229:120647，2021年，doi：https://doi.org/10.1016/j.energy.2021.120647。[19] Alexis Fouilloy、Cyril Voyant、Gilles Notton、Fabrice Motte、Christophe Paoli、Marie-Laure Nivet、Emmanuel Guillot 和 Jean-Laurent Duchaud。利用机器学习进行太阳辐照度预测：根据天气变化选择预测模型的方法。《能源》，165:620–629，2018年，doi：https://doi.org/10.1016/j.energy.2018.09.116。[20] Peder Bacher、Henrik Madsen 和 Henrik Aalborg Nielsen。在线短期太阳能发电预测。《太阳能》，83(10):1772–1783，2009年，doi：10.1016/j.solener.2009.05.016。[21] Federica Davò、Stefano Alessandrini、Simone Sperati、Luca Delle Monache、Davide Airoldi 和 Maria T Vespucci。区域风电和太阳辐照度预测的后处理技术与主成分分析。《太阳能》，134:327–338，2016年，doi：https://doi.org/10.1016/j.solener.2016.04.049。

[22] Jie Shi、Wei-Jen Lee、Yongqian Liu、Yongping Yang 和 Peng Wang。基于天气分类和支持向量机的光伏系统功率输出预测。IEEE工业应用汇刊，48(3):1064–1069，2012年，doi：10.1109/IAS.2011.6074294。[23] Ricardo Aler、Ricardo Martín、José M. Valls 和 Inés M. Galván。利用数值天气模型进行每日太阳能预测的机器学习技术研究。载于David Camacho、Lars Braubach、Salvatore Venticinque 和 Costin Badica（编），《智能分布式计算 VIII》，第269–278页，沙姆，2015年。施普林格国际出版公司。[24] Junliang Fan、Xiukang Wang、Lifeng Wu、Hanmi Zhou、Fucang Zhang、Xiang Yu、Xianghui Lu 和 Youzhen Xiang。在湿润亚热带气候下，利用温度和降水预测每日全球太阳辐射的支持向量机与极端梯度提升对比：中国案例研究。《能源转换与管理》，164:102–111，2018年，doi：https://doi.org/10.1016/j.enconman.2018.02.087。[25] Rachit Srivastava、AN Tiwari 和 VK Giri。利用各种人工神经网络模型预测印度的太阳辐射。载于2018年IEEE北方邦分会电气、电子与计算机工程国际会议（UPCON），第1–6页。IEEE，2018年。[26] Phathutshedzo Mpfumali、Caston Sigauke、Alphonce Bere 和 Sophie Mulaudzi。日前逐小时全球水平面辐照度预测——南非数据应用。《能源》，12(18):3569，2019年，doi：https://doi.org/10.3390/en12183569。[27] Veysel Çoban 和 Sezi Çevik Onar。基于机器学习的土耳其伊斯坦布尔太阳辐射预测。载于《智能与模糊系统国际会议》，第197–204页。施普林格，2019年。[28] Zezhou Chen 和 Irena Koprinska。太阳能发电预测的集成方法。载于2020年国际神经网络联合会议（IJCNN），第1–8页。IEEE，2020年。[29] PAGM Amarasinghe、NS Abeygunawardana、TN Jayasekara、EAJP Edirisinghe 和 SK Abeygunawardane。太阳能发电预测的集成模型——一种天气分类方法。AIMS能源，8(2):252–271，2020年，doi：10.3934/energy.2020.2.252。[30] Tendani Mutavhatsindi、Caston Sigauke 和 Rendani Mbuvha。利用机器学习模型预测南非逐小时全球水平面太阳辐照度。IEEE Access，8:198872–198885，2020年，doi：10.1109/ACCESS.2020.3034690。[31] Mawloud Guermoui、Farid Melgani、Kacem Gairaa 和 Mohamed Lamine Mekhalfi。太阳辐射预测混合模型的全面综述。《清洁生产杂志》，258:120357，2020年，doi：https://doi.org/10.1016/j.jclepro.2020.120357。[32] Yong Zhou、Yanfeng Liu、Dengjia Wang、Xiaojun Liu 和 Yingying Wang。从综合视角综述利用机器学习模型进行全球太阳辐射预测。《能源转换与管理》，235:113960，2021年，doi：https://doi.org/10.1016/j.enconman.2021.113960。[33] Imane Jebli、Fatima-Zahra Belouadha、Mohammed Issam Kabbaj 和 Amine Tilioua。基于深度学习的太阳能预测模型。Advances Sci，6:349–355，2021年，doi：10.25046/aj060140。[34] Shichao Zhang、Jilian Zhang、Xiaofeng Zhu、Yongsong Qin 和 Chengqi Zhang。基于数据聚类的缺失值填补。载于《计算科学汇刊 I》，第128–138页。施普林格，2008年。[35] Philip Sedgwick。皮尔逊相关系数。英国医学杂志，345，2012年，doi：10.1136/bmj.e4483。[36] Muhannad Alaraj、Astitva Kumar、Ibrahim Alsaidan、Mohammad Rizwan 和 Majid Jamil。基于气象参数预测沙特阿拉伯卡西姆地区太阳能光伏电站的能源产量。《IEEE Access》，9:83241–83251，2021年，doi：10.1109/ACCESS.2021.3087345。[37] Peter Bühlmann。装袋法、提升法与集成方法。载于《计算统计学手册》，第985–1022页。施普林格，2012年。[38] Leo Breiman。随机森林。《机器学习》，45(1):5–32，2001年，doi：https://doi.org/10.1023/A:1010933404324。[39] Pierre Geurts、Damien Ernst 和 Louis Wehenkel。极端随机树。《机器学习》，63(1):3–42，2006年，doi：https://doi.org/10.1007/s10994- 006-6226-1。[40] Yoav Freund 和 Robert E Schapire。在线学习的决策理论推广及其在提升法中的应用。《计算机与系统科学杂志》，55(1):119–139，1997年，doi：https://doi.org/10.1006/jcss.1997.1504。[41] Yoav Freund 和 Robert E. Schapire。提升法简明入门。载于《第十六届国际人工智能联合会议论文集》，第1401–1406页。摩根·考夫曼，1999年。[42] Jerome H Friedman。随机梯度提升。《计算统计与数据分析》，38(4):367–378，2002年。[43] Alexey Natekin 和 Alois Knoll。梯度提升机教程。《神经机器人学前沿》，7:21，2013年，doi：10.3389/fnbot.2013.00021。[44] Tianqi Chen 和 Carlos Guestrin。XGBoost：一种可扩展的树提升系统。载于《第二十二届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第785–794页，2016年。[45] Guolin Ke、Qi Meng、Thomas Finley、Taifeng Wang、Wei Chen、Weidong Ma、Qiwei Ye 和 Tie-Yan Liu。LightGBM：一种高效的梯度提升决策树。《神经信息处理系统进展》，30，2017年。[46] Wentao Cai、Ruihua Wei、Lihong Xu 和 Xiaotao Ding。一种利用梯度提升决策树对温室温度建模的方法。《农业信息处理》，9(3):343–354，2022年，doi：https://doi.org/10.1016/j.inpa.2021.08.004。[47] Kun An 和 Jiang Meng。用于回归器集成的投票平均组合方法。载于《智能计算国际会议》，第540–546页。施普林格，2010年。[48] L. Breiman。堆叠回归。《机器学习》，24:49–64，2004年。[49] Ye Ren、PN Suganthan 和 N Srikanth。风能和太阳能发电预测的集成方法——最新进展综述。《可再生与可持续能源评论》，50:82–91，2015年，doi：https://doi.org/10.1016/j.rser.2015.04.081。[50] Voyant, C.、Notton, G.、Duchaud, J.-L.、Almorox, J. 和 Yaseen, Z. M.（2020）。基于Box–Cox变换和周期自回归模型单变量表示的太阳辐照度预测区间。《可再生能源聚焦》，33，43–53。doi：10.1016/j.ref.2020.04.001。[51] Ayodele, T. R.、Ogunjuyigbe, A. S. O.、Amedu, A. 和 Munda, J. L.（2019）。利用混合K均值和支持向量回归算法预测全球太阳辐照度。《可再生能源聚焦》，29，78–93。doi：10.1016/j.ref.2019.03.003。