【电子与信息科学／Electronics and Information Science】 ·人工智能与数字经济专题·

基于残差的分布式光伏发电功率组合预测方法

吴明朗[1]，庞振江[1]，洪海敏[2]，占兆武[1]，靳飞[2]，唐远洋[1]，叶璇[3]

1）深圳市国电科技通信有限公司，广东深圳 518109；2）深圳智芯微电子科技有限公司，广东深圳 518048；3）人工智能与数字经济广东省实验室(深圳)，广东深圳 518107

摘　要：分布式光伏发电功率预测在保障电网运行安全和就近消纳方面发挥着重要作用，为提升分布式光伏发电功率预测精度，提出一种基于多元气象的特征提取方法和基于残差连接的多模型融合的光伏发电功率预测模型．在特征提取时，引入统计、交叉、周期信息、近似熵和光伏板温度等特征提取方法，实现对时间、气象和发电功率的深层特征提取，丰富模型的输入．在模型构建时，建立基于残差连接的多层模型融合方法，首先提出基于k最近邻(k-nearest neighbor，kNN)的softmax回归预测模型，其次设计3层模型整体结构，并通过残差连接和多层堆叠的方式融合多个预测模型，持续提升光伏发电功率预测精度．基于电力公司真实数据，采用本研究方法与随机森林(random forest，RF)、TabNet和极端梯度提升(extreme gradient boosting，XGBoost)等模型，对光伏发电功率进行预测．结果表明，所提模型在均方根误差、平均绝对误差、均方误差和平均绝对百分比误差等方面可分别降低0.109 7、0.059 1、0.050 7和0.036 8，拟合优度可提升0.080 4．基于多元气象的特征提取方法和基于残差连接的多模型融合的光伏发电功率预测模型能有效提升分布式光伏发电功率预测的精度和稳定性．

关键词：人工智能；太阳能；特征提取；残差连接；随机森林；TabNet；极端梯度提升；功率预测

中图分类号：TP311；TM615　　**文献标志码：**A　　**DOI：**10.3724/SP.J.1249.2024.03293

Skip-based combined prediction method for distributed photovoltaic power generation

WU Minglang[1], PANG Zhenjiang[1], HONG Haimin[2], ZHAN Zhaowu[1], JIN Fei[2], TANG Yuanyang[1], and YE Xuan[3]

1) China Gridcom Co. Ltd., Shenzhen 518109, Guangdong Province, P. R. China

2) Shenzhen Smart-Chip Microelectronics Technology Co. Ltd., Shenzhen 518048, Guangdong Province, P. R. China

3) Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518107, Guangdong Province, P. R. China

Abstract: Distributed photovoltaics (PV) power generation forecasting plays an important role in ensuring the safety of power grid operation and nearby consumption. In order to enhance the accuracy of distributed PV power generation forecasting, we propose a meteorological feature extraction method and futher design a PV power generation forecasting model based on skip-connect models fusion. In the feature extraction, we use statistical analysis, features cross-correlation, periodicity information, approximate entropy, and the temperature of PV panels to achieve deep feature extraction of time, weather, and power generation data, enriching the model inputs. In model construction, we propose a multi-layer model fusion method based on residual connections. Firstly, we introduce a k-nearest neighbor (kNN)-based softmax regression prediction model. Secondly, we design a three-layer model structure with multiple prediction models fused through residual connections and multi-layer stacking, continuously

Received: 2023-05-28; **Accepted:** 2023-09-24; **Online** (CNKI)**:** 2024-03-09

Foundation: Natural Science Foundation of Guangdong Province (2023A1515011667); Shenzhen Basic Research Foundation (JCYJ20220818100205012, JCYJ20210324093609026)

Corresponding author: Senior engineer ZHAN Zhaowu (zhanzhaowu@sgchip.sgcc.com.cn)

Citation: WU Minglang, PANG Zhenjiang, HONG Haimin, et al. Skip-based combined prediction method for distributed photovoltaic power generation [J]. Journal of Shenzhen University Science and Engineering, 2024, 41(3): 293-302. (in Chinese)

impg the prediction accuracy of PV power generation forecasting. Based on the real data of electric power companies, we compare the proposed method with others, such as random forest (RF), TabNet and extreme gradient boosting (XGBoost) for photovoltaic power generation prediction. The results show that the proposed model can reduce the root mean square error, mean absolute error, mean squared error, and mean absolute percentage error by 0.109 7, 0.059 1, 0.050 7, and 0.036 8 respectively, and improve the goodness of fit by 0.080 4. The feature extraction method based on multi-meteorological factors and the photovoltaic power generation prediction model based on residual connections for multi-model fusion effectively improve the accuracy and stability of distributed PV power generation forecasting.

Key words: artificial intelegient; solar energy; feature extraction; skip-connect; random forest; TabNet; extreme gradient boosting; power prediction

随着中国能源低碳化转型步伐的加快与能源体系改革的推进，新能源发电技术逐渐成为主流发电技术，其中光伏发电是新能源的重要组成部分，分布式光伏发电已成为中国新能源发电的重要方式.

光伏发电技术机理复杂，影响发电功率的因素[1]有很多，其中分布式光伏发电的发电功率不仅受太阳辐照强度和温度的影响[2]，还受云量、湿度、风速、压强和光伏板温度等因素的影响，且每个影响因素的影响程度随气象的时间和条件等动态变化[3]. 分布式光伏发电的发电功率受各方面因素的影响，这些影响因素的不确定性和强随机性，导致发电功率不稳定. 对发电功率进行精准预测能有效解决分布式光伏发电的不稳定性，这不仅对电网安全和稳定运行起着至关重要的作用，同时对新能源的消纳和负荷调控等发挥着重要作用.

随着气象学和人工智能的深入发展，目前对光伏发电功率的预测大多是采用机器学习和深度学习的方法[3-4]并结合数值天气预报(numerical weather prediction, NWP). 分布式光伏发电的功率预测从建模逻辑角度出发，可分为两类：一是机理驱动方法，由气象信息和光伏系统参数，依据物理原理建立预测模型[5]；二是数据驱动方法，通过分析历史出力[6-7]和NWP信息等数据间的关系来建立预测模型[8-10]. 王彪等[11-12]提出光伏发电功率的物理计算模型，基于太阳辐射、气象因子和光伏板自身参数通过物理模型计算发电功率，但因分布式光伏发电大多缺少光伏板自身参数，无法使用该模型. 同时，物理模型在预测问题上，完全取决于气象预报数据，对气象变化不敏感，导致预测精度较低. 李光明等[13-17]采用了一般的机器学习算法进行发电功率预测，如多元回归、支持向量机(support vector machines, SVM)和轻量的梯度提升机(light gradient boosting machine, LightGBM)等算法，可快速地基于不同数据构建模型，通过融合多种特征处理的方法来预测发电功率. 但是，采用机器学习方法预测发电功率时，因无法有效学习到气象和历史功率数据的深层特征，导致最终的预测精度较低. 文献[18-20]采用时序的深度学习方法，很好地学习到了功率曲线周期信息和功率在时间上的依赖关系，但因无法有效学习气象数据对发电功率之间的影响关系，导致预测精度较低且稳定性较差. 文献[21-22]提出基于天气类型的预测模型，根据不同的天气类型分别构建模型，这虽然能提升部分天气类型的预测精度，但整体预测精度仍相对较低.

可见，现有的分布式发电功率预测方法主要存在以下问题：① 分布式光伏发电受众多因素的影响且机理复杂，一般的特征提取方法不足以支撑稳定和精准的发电功率预测；② 在提取特征时，缺少对时序特征多变量相互作用和周期特征的提取；③ 基于物理模型进行分布式光伏发电功率预测，需实时获取如光伏板温度、辐照度等物理量，时效性和准确性要求较大，而物理模型在应用时难以支撑；④ 基于数据驱动的方法，一般在学习过程中会偏重于某一方面，不足以解决分布式光伏发电功率预测的问题. 为此，本研究提出面向分布式光伏发电功率预测的特征提取方法，全面且有效地提取气象数据和历史发电功率的深层特征，解决了特征少和模型学习不充分的问题. 采用基于k最近邻(k-nearest neighbor, kNN)的softmax回归预测模型和基于残差连接(skip-connect)的多模型融合方法，在层与层之间采用残差连接的方式和堆叠(stacking)集成方式，逐级提升模型预测的精度和稳定性.

1 特征提取

特征提取会直接影响模型训练的收敛程度和分

布式光伏发电功率预测模型的预测精度．本研究针对分布式光伏发电功率预测提出基于多元信息的征提取方法，包括统计、交叉、周期和信息熵等特征提取和构造方法．

1.1　基于多元信息构造基础特征

针对气象数据和历史光伏发电功率的多元数据，采用统计、多变量交叉和预测的方法构造多维特征．

1）统计特征．影响分布式光伏发电功率的气象因素或历史发电功率都是时间序列型变量，基于变量的历史序列值，使用统计的方法构造出新的统计特征，包括最大值、均值、峰态系数和偏态系数．比如，对辐射强度变量在t时刻的前m个历史值组成的序列$\{I_{t-m-1}, I_{t-m}, \cdots, I_{t-1}\}$，计算其统计特征．

2）交叉特征．分布式光伏发电功率的影响因素众多，各因素之间可能相互作用，因此，通过构造不同变量之间的交叉特征，形成不同的特征组合，有助于模型对变量之间的关系进行学习．t时刻两个变量交叉后的特征为

$$c_t = x_t^{(1)} \times x_t^{(2)} \tag{1}$$

其中，$x_t^{(1)}$和$x_t^{(2)}$分别为t时刻的变量1和变量2；c_t为t时刻交叉特征提取后的特征变量．

3）趋势特征．在对某时刻的发电功率进行预测时，影响因素的趋势是非常重要的信息，因为，构建变量的趋势特征能帮助模型学习到有效信息，提升预测模型的精度和泛化能力．t时刻两个变量的趋势特征为

$$v_t = h\left(x_t^{(1)}, x_t^{(2)}\right) = \frac{x_t^{(1)} - x_{t-w}^{(1)}}{x_t^{(2)} - x_{t-w}^{(2)} + c} \tag{2}$$

其中，$x_{t-w}^{(1)}$和$x_{t-w}^{(2)}$分别为$t-w$时刻的变量1和变量2；v_t为两变量在t时刻的趋势特征；c为一个较大的常数，用于防止分母过小导致v_t过大，一般分布式光伏发电机装机容量在50 kW以下，本研究取$c=50$．

1.2　基于时间构造周期特征方法

分布式光伏发电受气象条件的影响，因此，光伏发电的变化具有强周期性特征，即在时间上具有明显的周期性，这种周期信息是预测模型的重要特征信息．因此，本研究针对功率发电所在的时间戳，提取时间戳的月份(m)、周数(w)、天数(d)，以及时刻(h)信息．

利用sin和cos函数构造出如式(3)的以月、周、日和时为基础的周期特征，将原始离散的时间特征转化转换为周期序列．

$$\begin{cases} z_{\sin} = \sin\left(2\pi \dfrac{z}{z_{\max}}\right) \\ z_{\cos} = \cos\left(2\pi \dfrac{z}{z_{\max}}\right) \end{cases} \tag{3}$$

其中，z为输入的时间特征；$z_{\max}$为所在时间特征对应的最大值；月份、周数、天数和时刻对应的最大值分别为12、53、366和24；$z_{\sin}$和$z_{\cos}$分别为时间特征提取后的正弦和余弦特征变量．

1.3　基于近似熵提取波动性特征

在分布式光伏发电功率预测中，从时间序列的角度出发，未来的发电功率与历史的发电功率存在某种关系，当前时间点前后的发电功率具有一定的因果关系[11]．因此，对历史发电功率序列进行信息提取，并输入到预测模型中，可提升预测模型性能．

熵可用于衡量一个系统的混乱程度或随机程度，熵值越小系统越确定，熵值越大系统越混乱．序列的熵是序列信息的一种表示，因此，历史的发电功率序列可使用熵对其进行信息提取．本研究使用近似熵对序列信息进行特征提取(图1)，详细步骤为：

输入： 长度为N的$S_N \leftarrow \{s_1, s_2, \cdots, s_N\}$序列，距离度量阈值$r_0$，子序列$z(i)$的长度为$m$

输出： 近似熵的值E_A

for $m \leftarrow m$ to $m+1$ do

$z(i) \leftarrow \{s_i, s_{i+1}, s_{i+2}, \cdots, s_{i+m}\}$

$Z_{N-m+1} \leftarrow \{z(1), z(2), \cdots, z(i), \cdots, z(N-m+1)\}$

$\boldsymbol{D}_{(N-m+1)\times(N-m+1)} \leftarrow d_m\left(z(i), z(j)\right)$

$C_i^{(m)}(r_0) \leftarrow \dfrac{\mathrm{num}\left(d_m\left(z(i), z(j)\right) < r_0\right)}{N-m+1}$

$\Phi_m(r_0) \leftarrow \dfrac{\sum_{i=1}^{N-m+1} \ln\left(C_i^{(m)}(r_0)\right)}{N-m+1}$

end for

$E_A \leftarrow \Phi_{m=m}(r_0) - \Phi_{m=m+1}(r_0)$

图1　近似熵计算伪代码

Fig. 1　Pseudocode of approximate entropy calculation.

步骤1　设置长度为N的历史发电功率序列$S_N = \{s_1, s_2, \cdots, s_N\}$，子序列的长度为$m$，并定义距离度量的阈值$r_0$．

步骤2　将S_N按长度为m的子序列进行重构，得到$N-m+1$个子序列$\{z(1), z(2), \cdots, z(N-m+1)\}$，其中，第$i$个子序列$z(i) = \{s_i, s_{i+1}, s_{i+2}, \cdots, s_{i+m}\}$，

$i = 1, 2, \cdots, N - m + 1$.

步骤3 计算任意两个重构子序列$z(i)$和$z(j)$之间的距离$d_m\left(z(i), z\left(j\right)\right)$，得到距离矩阵$\boldsymbol{D}_{(N-m+1)\times(N-m+1)}$. 其中，$j = 1, 2, \cdots, N - m + 1$，包括$i = j$的距离，距离的计算方式可以自定义，本研究使用两个序列对应位置元素的最大差值.

步骤4 统计满足任意两个重构子序列之间的距离小于阈值r_0条件的子序列个数，其与重构后的子序列数之间的比值为

$$C_i^{(m)}\left(r_0\right) = \frac{\mathrm{num}\left(d_m\left(z(i), z\left(j\right)\right) < r_0\right)}{N - m + 1} \tag{4}$$

其中，num为满足条件$d_m\left(z(i), z\left(j\right)\right) < r_0$的子序列个数的统计函数.

步骤5 计算长度为m的子序列的平均相似率

$$\Phi_m\left(r_0\right) = \frac{\sum_{i=1}^{N-m+1} \ln\left(C_i^{(m)}\left(r_0\right)\right)}{N - m + 1} \tag{5}$$

步骤6 重复步骤1至步骤5，计算当子序列$z(i)$长度为$m + 1$时的平均相似率$\Phi_{m+1}\left(r_0\right)$.

步骤7 输出历史发电功率序列的近似熵，即

$$E_{\mathrm{A}} = \Phi_m\left(r_0\right) - \Phi_{m+1}\left(r_0\right) \tag{6}$$

1.4 基于温度构造的光伏板温度特征

根据光伏发电机理，影响发电功率的因素不仅包括气象、光伏板面积和转化率等，光伏板自身温度也是影响发电功率的重要因素[11]. 本研究通过物理模型来构造光伏板的温度，采用气象数据集中的温度T_{NWP}近似替代环境温度T_{amb}，则可获得光伏板运行温度T_{C}的估计值[11]为

$$\hat{T}_{\mathrm{C}} = T_{\mathrm{amb}} + \frac{I_t}{I_{\mathrm{NOCT}}}\left(T_{\mathrm{C,\,NOCT}} - T_{\mathrm{amb}}\right) \tag{7}$$

其中，$T_{\mathrm{C,\,NOCT}}$为标称环境下的板温，一般取20 ℃；I_{NOCT}为标称环境下的辐照度，一般取800 W/m^2；I_t为t时刻的辐照度，本研究采用气象预报中的辐照度.

2 基于残差连接和多模型融合的发电功率预测方法

采用残差连接和多模型融合的方法构造分布式光伏发电功率预测模型，模型的整体框架如图2. 在第1层基于原始数据集$D_{\mathrm{raw}}\langle \boldsymbol{X}, \boldsymbol{Y}\rangle$，进行相似日的分布式光伏发电功率回归预测，同时提取原始数据集的特征得到其特征数据集$D'\langle \boldsymbol{X}', \boldsymbol{Y}\rangle$. 其中，$\boldsymbol{X}$为原始的输入变量；$\boldsymbol{X}'$为特征提取后的特征变量；$\boldsymbol{Y}$为目标变量(发电功率). 在第2层，将相似日预测的输出和$D'\langle \boldsymbol{X}', \boldsymbol{Y}\rangle$作为输入构建多模型预测模块，TabNet[10]、极端梯度提升(eXtreme gradient boosting，XGBoost)[9]和随机森林(random forest，RF)[7]分别输出预测值. 在第3层，基于第2层多模型预测模块的输出结果和原始数据的残差连接构建输出层LightGBM模型，最终将LightGBM模型预测结果作为分布式光伏发电功率的预测值.

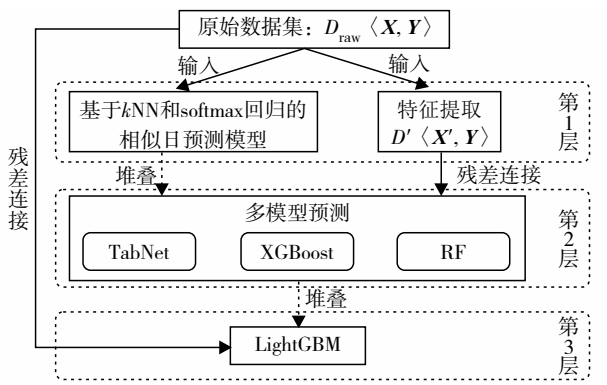

图2 残差连接和多模型融合的预测模型框架

Fig. 2 The framework of the skip-connect models fusion.

2.1 基于kNN和softmax回归的相似日预测模型

相似日理论是指根据历史数据，寻找与预测对象具有相似特征的时间段，并利用这些日期的历史数据来预测对象的未来状态，常应用于时间序列预测领域，可以有效提高预测精度. 因此，基于相似日理论和光伏发电功率预测原理，若分布式光伏发电在不同时间点的各个影响因素的值相同，则发电功率必然相同. 例如，同一光伏发电设备发电功率的影响因素包括温度、辐照度、湿度、地理环境、光伏板面积和季节等，若它们在两个时间点的值相近，则这两个时间点的发电功率必然非常接近. 因此，若能在预测阶段找到与预测时间点相似的历史时间点，则可将历史的相似时间点的发电功率作为预测值. 为保证模型的稳定性，本研究基于注意力机制的思想，取top-k的相似时间点，采用softmax函数计算发电功率预测值为

$$y = f\left(\boldsymbol{q}, \boldsymbol{P}, V\right) = V\mathrm{softmax}\left(\frac{1}{d_{\text{top-}k}\left(\boldsymbol{q}, \boldsymbol{P}\right)}\right) \tag{8}$$

其中，$\boldsymbol{q}$为预测时间点向量；$\boldsymbol{P}$和V分别为历史时间点向量和该时间点对应的发电功率值；$d_{\text{top-}k}$为取距离度量值按升序排序后的前k个值函数.

2.1.1　距离度量

本研究选用L2(欧式)距离进行距离的度量，如式(9)．由于L2距离对异常值比较敏感，在高维空间中，可令相近的向量更易区分．

$$d(\boldsymbol{I},\boldsymbol{O})=\sqrt{\sum_{m=1}^{M}\left(I_m-O_m\right)^2} \tag{9}$$

其中，$\boldsymbol{I}$和$\boldsymbol{O}$分别为某个时间点的输入和输出向量；I_m和O_m分别为$\boldsymbol{I}$和$\boldsymbol{O}$中的第m个元素；由式(9)可见，相似日理论的关键在于时间点的输入向量，该向量将不同时间点之间的差异表示越精确，基于相似日预测的精度会越高．

2.1.2　基于kNN实现相似日softmax回归预测

基于相似日理论和kNN[6]算法的原理，使用kNN实现相似日的预测．由于历史时间点的数据量较大，需要特殊的数据结构来实现快速的近邻搜索．KD(k-dimension)树是一种可高效处理高维空间信息的数据结构，也是kNN算法中常用的一种数据结构，可以快速实现高维度数据的近邻查询．因此，基于KD树和欧式距离的度量方法可快速实现相似日的top-k查询．

采用softmax函数得到的top-k时间点的距离计算k个时间点的回归系数，并和对应时间的发电功率值加权，即可实现基于相似日回归的分布式光伏发电功率预测．

2.2　基于残差连接融合的分布式光伏发电功率预测模型

为提升模型精度，本研究基于残差连接的思想，通过融合RF[7]、XGBoost[9]和TabNet[10]这3个差异性较大的模型来预测分布式光伏发电功率．由于不同模型原理各异，即使针对同一数据集训练，得到的预测结果也各不相同，若能较好地利用这种差异，不仅可以提升最终的预测效果，也能增强预测的稳定性，令模型泛化能力更强．

2.2.1　TabNet模型

TabNet是基于注意力机制的表格数据学习模型，不仅具有深度网络的优势，还具有传统机器学习的可解释能力．它通过多层的特征处理(feature transformer)模块来实现特征工程，避免了人工特征工程，有效利用了深度网络的优势来提取深层特征．模型提取到的特征可供决策步使用，并基于注意力机制找到最相关的特征，实现了基于实例的重要特征选择，学习到了最突出的特征，这使模型不仅在表格数据上有高性能，还具有可解释性．

TabNet模型整体结构如图3[10]，编码器由多个特征处理模块和多个注意力(attentive transformer)模块堆叠而成．编码器的输入是处理好的数值型特征，输出则包括编码后的特征和供最终决策使用的数据．解码器中的决策步(step)类似策树中的判断结点，每个step都接收所有输入的特征，并使用上一个step的输出对数据特征加权，最终输出所有累加的step结果用于最终决策．

在TabNet模型结构中，特征处理模块先使用共享决策步(shared across decision steps)模块对输入的特征进行处理，再使用决策步相关(decision step dependent)模块进行进一步处理，如图4[10]．其中，FC为全连接(full connect)；BN为批处理归一化(batch normalization)；GLU为门控线性单元(gated linear unit)．

注意力模块则用于特征选择．图5[10]给出了注意力模块结构，通过全连接层、归一化层和稀疏化层(sparsemax)，输出各个特征的稀疏的重要性表示．

对于分布式光伏发电功率预测，TabNet模型可

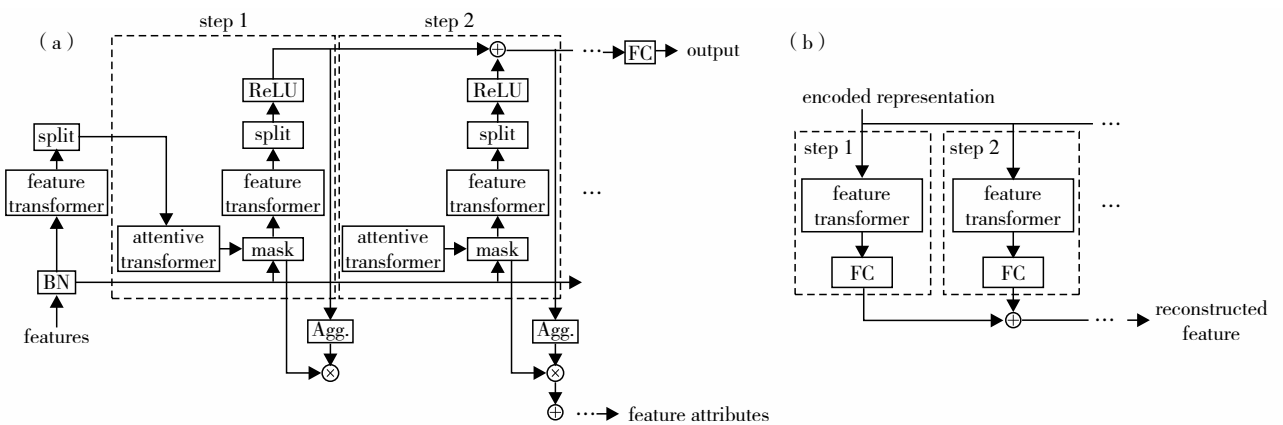

图3　TabNet模型整体结构[10]　(a)编码器；(b)解码器

Fig. 3　The structure of the TabNet model[10]. (a) Encoder and (b) decoder.

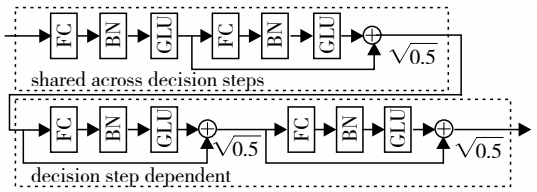

图4 特征处理模块的结构[10]

Fig. 4 Feature transformer block[10].

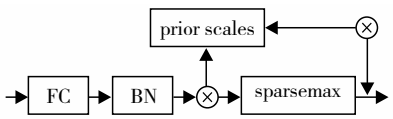

图5 注意力模块结构[10]

Fig. 5 Attentive transformer block[10].

学习到各个气象变量的深层特征，特别是在数据量大的情况下，效果更好．同时，TabNet模型还可根据实例之间的差异学习不同的特征，弥补了人工特征提取的不足，以及一般机器学习模型并非基于实例学习的不足、一般机器学习模型对大数据欠学习等缺点．用函数T表示TabNet模型，则TabNet模型预测结果为

$$Y_{\mathrm{T}} = T\left(\boldsymbol{X}_{\mathrm{T}}\right) \tag{10}$$

其中，$\boldsymbol{X}_{\mathrm{T}}$为模型的输入，即第1层的基于$k$NN和softmax回归的相似日预测模型预测输出结果和特征提取后的输出特征合并后的向量．

2.2.2 XGBoost模型和RF模型

XGBoost是一种基于boosting集成策略的学习模型，它对表格数据的处理表现非常好，在训练过程中能有效降低模型的偏差，提升模型的预测精度．XGBoost模型通过数据中特征的分裂和训练生成多棵回归树，每一轮的树基于上一轮的残差作为目标来学习和生成，因此模型可定义为

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), \qquad f_k \in \boldsymbol{F} \tag{11}$$

其中，x_i为第i个样本；f_k为第k棵树模型；K为回归树的总数；$\boldsymbol{F}$为包含所有回归树的函数空间．

由于XGBoost是加法模型，每一步迭代都是生成一棵新的回归树模型，当前迭代是基于上一次迭代的残差生成新的树模型．因此，可将式(11)按训练步进行拆分，即

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = f_1\left(x_i\right) = \hat{y}_i^{(0)} + f_1\left(x_i\right) \\ \hat{y}_i^{(2)} = f_1\left(x_i\right) + f_2\left(x_i\right) = \hat{y}_i^{(1)} + f_2\left(x_i\right) \\ \vdots \\ \hat{y}_i^{(K)} = \sum_{k=1}^{K} f_k\left(x_i\right) = \hat{y}_i^{(k-1)} + f_K\left(x_i\right) \end{cases} \tag{12}$$

由式(12)可见，k次迭后第i个样本的预测值$\hat{y}_i^{(k)}$等于前$k-1$次迭代后的预测值加上当前迭代的预测结果$f_k(x_i)$．$f_k(x_i)$是第k颗树，即第k次迭代需要生成的回归树．因此，式(12)是XGBoost模型在训练迭代过程中具体的数学形式．

XGBoost模型的目标函数主要包括损失函数部分和正则化部分．损失函数部分主要目的是使得模型拟合更好，而正则化部分目的是惩罚每棵树的复杂程度，令模型更简单，防止过拟合，使其泛化性更好．因此，XGBoost模型的目标函数为

$$O^{(K)} = \sum_{k=1}^{K} l\left(y_i, \hat{y}_i^{(k)}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{13}$$

其中，$l\left(y_i,\ \hat{y}_i^{(k)}\right)$为损失函数，$y_i$为第$i$个样本的真实值，XGBoost模型可自定义损失函数，但要求可以求二阶导数；$\Omega(f_k)$为正则化部分，用于对每棵树的复杂程度进行控制，

$$\Omega\left(f_k\right) = \gamma L + \lambda \frac{1}{2} \sum_{l=1}^{L} w_j^2 \tag{14}$$

其中，f_k为第k颗树；l为第k颗树(f_k)叶子节点序号；L为第k颗树(f_k)的叶子节点数；γ为惩罚系数；λ为正则化惩罚系数；w_j为叶子节点分数，$\frac{1}{2}\sum_{l=1}^{L} w_j^2$为子节点分数的L2正则化．

RF是基于引导聚合(bootstrap aggregating，Bagging)算法集成策略的学习模型，在训练过程中能有效降低模型方差，在多个随机子空间下学习不同的决策函数，降低模型偏差的同时，提升模型的泛化能力．RF模型的伪代码见图6．

输入： 数据集D和子模型个数K

输出： K个集成的树模型$\{T_k\}_{k=1}^{K}$

1，for $k \leftarrow 1$ to K do

2，　对数据集D进行又放回的行列采样得到D^*

3，　基于D^*通过CART树生成T_k

4，end for

5，回归预测：$y \leftarrow \frac{1}{K}\sum_{k=1}^{K} T_k(x)$

图6 RF回归算法伪代码

Fig. 6 Pseudocode of RF regression algorithm.

XGBoost模型和RF模型的集成策略和训练过程具有较大差异，可认为分别着重于减小偏差和减小方差．因此，两者的融合不仅可以提升预测精度，还可以提升预测的泛化能力．

在分布式光伏发电功率预测中，XGBoost模型

可有效提升预测的精度，而RF模型在保证精度的同时可以提升预测的稳定性．同时，对于较小型数据集，XGBoost模型和RF模型同样表现良好，弥补了TabNet模型在小数据集上的效果欠佳的缺点．分别用函数G和R表示XGBoost和RF模型，则对应的模型预测结果分别为

$$Y_{\mathrm{G}} = G(\boldsymbol{X}_{\mathrm{T}}) \tag{15}$$

$$Y_{\mathrm{R}} = R(\boldsymbol{X}_{\mathrm{T}}) \tag{16}$$

2.3　基于堆叠的融合预测模型

基于相似日的分布式发电功率预测和多模型的分布式光伏发电功率预测有各自的优势，能解决相应问题，因此，将它们进行有效融合可实现更高精度的分布式光伏发电功率预测．

基于相似日的分布式发电功率预测方法能很好地对目标变量的缺失值和异常值进行修正，特别是在缺失值较多、缺失值或异常值处理不合理时，能较好地进行二次修正，初步预测出分布式光伏发电功率．基于相似日的分布式光伏发电功率预测和特征工程，进行多模型的分布式光伏发电功率预测，可以学习特征工程之后的隐藏信息，进一步提升预测的精度．最后，将多模型的预测进行融合，并利用原始输入进一步对多模型预测结果进行微调，进而实现更高精度的分布式光伏发电功率预测．

本研究采用堆叠集成策略和残差连接的方式进行3层模型融合(图3)．首先，在第1层利用原始输入数据$\boldsymbol{X}$采用相似日预测模型预测发电功率，同时进行特征提取：

$$\boldsymbol{Y}_{\mathrm{S}} = S(\boldsymbol{X}) \tag{17}$$

$$\boldsymbol{X}' = F(\boldsymbol{X}) \tag{18}$$

其中，$\boldsymbol{X}$为原始数据集；函数S为相似日预测模型；函数F为特征提取过程；$\boldsymbol{X}'$为经过特征提取后的输出．通过式(16)和式(17)可实现对原始输入$\boldsymbol{X}$的特征扩展．

在第2层，采用向量拼接函数concat函数将两个向量$\boldsymbol{a}_1$和$\boldsymbol{a}_2$转为一个多维向量$\boldsymbol{a}_3$，即$\boldsymbol{a}_3 = \mathrm{concat}(\boldsymbol{a}_1, \boldsymbol{a}_2)$．例如，若$\boldsymbol{a}_1$和$\boldsymbol{a}_2$分别为大小为$b \times h \times d_1$和$b \times h \times d_2$的三维向量，则合并后$\boldsymbol{a}_3$则变成了大小为$b \times h \times (d_1 + d_2)$的三维向量．因此，将$\boldsymbol{X}'$和$\boldsymbol{Y}_{\mathrm{S}}$合并后可得

$$\boldsymbol{X}_{\mathrm{L2}} = \mathrm{concat}(\boldsymbol{X}', \boldsymbol{Y}_{\mathrm{S}}) \tag{19}$$

再将$\boldsymbol{X}_{\mathrm{L2}}$作为第2层的输入，分别输入TabNet、XGBoost和RF模型，得到3个模型的预测结果分别为

$$\begin{cases} \boldsymbol{Y}_{\mathrm{T}} = T(\boldsymbol{X}_{\mathrm{L2}}) \\ \boldsymbol{Y}_{\mathrm{G}} = G(\boldsymbol{X}_{\mathrm{L2}}) \\ \boldsymbol{Y}_{\mathrm{R}} = R(\boldsymbol{X}_{\mathrm{L2}}) \end{cases} \tag{20}$$

在第3层，采用concat函数将第2层的预测值和原始输入数据$\boldsymbol{X}$进行合并，再输入到LightGBM模型，输出最终的预测结果为

$$\begin{cases} \boldsymbol{X}_{\mathrm{L3}} = \mathrm{concat}(\boldsymbol{X}, \boldsymbol{Y}_{\mathrm{T}}, \boldsymbol{Y}_{\mathrm{G}}, \boldsymbol{Y}_{\mathrm{R}}) \\ Y' = \mathrm{LightGBM}(\boldsymbol{X}_{\mathrm{L3}}) \end{cases} \tag{21}$$

3　案例分析

本实验使用某电力公司15个月内使用分布式光伏发电的用户数据，包括气象预报(numerical weather prediction，NWP)数据和各个分布式光伏发电用户的历史发电功率曲线数据，数据集的时间分辨率为1 h．数据集按时间维度划分训练集(2021-04-21—2022-05-31)和测试集(2022-06-01—2022-07-20)，并进行特征提取、模型训练、测试和效果对比分析．数据集已经上传至公开网盘https://pan.baidu.com/s/1Ssj_T6Zx3jglKqttmqRlyQ?pwd=ygfn，读者可自行下载并进行仿真和验证．

3.1　评价指标

对模型误差的评价主要使用均方误差(mean-square error，MSE)、均方根误差(root-mean-square error，RMSE)、平均绝对误差(mean absolute error，MAE)、平均绝对百分比误差(mean absolute percentage error，SMAPE)和拟合优度(R^2)5个指标进行综合评估，依次为

$$e_{\mathrm{MAE}} = \frac{1}{n}\sum_{s=1}^{n}\left|y_s - \hat{y}_s\right| \tag{22}$$

$$e_{\mathrm{MSE}} = \frac{1}{n}\sum_{s=1}^{n}\left(y_s - \hat{y}_s\right)^2 \tag{23}$$

$$e_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{s=1}^{n}\left(y_s - \hat{y}_s\right)^2} \tag{24}$$

$$e_{\mathrm{SMAPE}} = \frac{1}{n}\sum_{s=1}^{n}\frac{\left|y_s - \hat{y}_s\right|}{\frac{\left|y_s\right| + \left|\hat{y}_s\right|}{2} + 0.2} \times 100\% \tag{25}$$

$$R^2 = 1 - \frac{\sum_s\left(y_s - \hat{y}_s\right)^2}{\sum_s\left(\bar{y}_s - y_s\right)^2} \tag{26}$$

其中，下标s为样本序号；y_s为样本s的真实值；$\hat{y}_s$为样本s的预测值；$\bar{y}_s$为样本集的均值．e_{MAE}、e_{MSE}、e_{RMSE}和e_{SMAPE}的值越小模型的预测效果越好，而R^2

的值越接近1模型预测效果越好.

3.2 特征提取的影响分析

对输入的原始气象数据和历史功率曲线数据进行特征提取，生成新的特征变量数据集，并分别基于RF、XGBoost和TabNet模型进行特征提取前后对比(表1)．评价指标采用3.1节中介绍的e_{MAE}、e_{MSE}、e_{RMSE}、e_{SMAPE}和R^2．由表1可见，本研究提出的特征提取方法非常有效且能够较大幅度提升模型预测的精度.

表1 特征提取和模型效果对比结果

Table 1 Comparison results of feature extraction and model effects

评价指标	原始数据				特征数据			
	RF	XGBoost	TabNet	本研究模型	RF	XGBoost	TabNet	本研究模型
e_{MAE}	0.156 3	0.168 8	0.178 1	0.154 4	0.115 4	0.107 1	0.168 5	0.095 3
e_{MSE}	0.083 9	0.099 8	0.091 1	0.081 8	0.041 0	0.036 1	0.089 6	0.031 1
e_{RMSE}	0.289 6	0.315 9	0.301 8	0.286 0	0.183 2	0.190 1	0.299 3	0.176 3
e_{SMAPE}	0.111 2	0.119 1	0.157 4	0.105 6	0.083 6	0.079 9	0.101 9	0.068 8
R^2	0.867 0	0.841 8	0.855 6	0.870 3	0.932 4	0.942 7	0.877 9	0.950 7

注：灰底数据为该评价指标下的最优值.

3.3 模型架构有效性分析

本研究采用多模型融合的架构，为验证所提出模型结构的有效性，使用3.1节中的5个评价指标，在测试集基础上对不同的模型结构进行效果评估，结果见表2．其中，结构A为本研究所提模型结构；结构B为取消基于kNN和softmax回归的相似日预测模型；结构C为取消A结构中第3层中的残差连接，同时采用线性回归替代LightGBM模型．由表2可见，结构A在均方误差、均方根误差、平均绝对误差、平均绝对百分比误差和拟合优度指标上均优于其他模型结构，表明本研究提出的模型结构有效.

表2 不同模型结构的预测方法的预测性能

Table 2 Performance of predicted methords with different model structures

评价指标	结构A	结构B	结构C
e_{MAE}	0.095 3	0.122 5	0.110 6
e_{MSE}	0.031 1	0.044 5	0.041 5
e_{RMSE}	0.176 3	0.210 9	0.203 6
e_{SMAPE}	0.068 8	0.114 1	0.076 8
R^2	0.950 7	0.929 4	0.934 2

注：灰底数据为该评价指标下的最优值.

3.4 模型性能分析

为分析所提模型在基础模型(RF、XGBoost和TabNet)上的提升效果，使用3.1节提出5个评价指标，对RF、XGBoost和TabNet模型和本研究提出的基于残差连接的多算法融合的发电功率预测模型的预测效果进行对比，基于原始数据集(特征提取前)和特征提取后的数据集，分别对预测效果进行对比，结果见表1．由表1可见，特征提取后4个模型的评价指标表现均优于原始数据集上的，而本研究所提模型无论是在原始数据集还是在特征提取后的数据集中表现均优于其他3个模型．其中，e_{MAE}值降低了0.059 1，e_{MSE}值降低了0.050 7，e_{RMSE}值降低了0.109 7，e_{SMAPE}值降低了0.036 8，R^2提升了0.080 4．因此，所提多层模型融合的发电功率预测模型能有效提升分布式光伏发电功率的精度，同时具有很好的稳定性.

为对比分析对所提预测模型在特征提取前后的预测结果，分别抽取测试集中2022-07-01 0:00至2022-07-05 23:00和2022-07-15 0:00至2022-07-19 23:00两个时段的预测结果进行可视化，结果如图8．由图8可见，真实的功率曲线(浅灰色)和预测的功率曲线(深灰色)能较好地重合．其中，图8(b)和(d)在2022-07-03、2022-07-04、2022-07-18和2022-07-19处的预测功率曲线(深灰色)和真实的发电功率曲线(浅灰色)更为接近，说明基于本研究提出的特征提取方法构建的模型预测效果更好，即使出现发电功率波动，所提特征提取方法和预测方法的预测效果依照更好，稳定性和精度也更高.

结 语

针对分布式光伏发电功率预测，提出面向分布式光伏的多种特征提取方法，包括基于多元信息的基础特征构造、基于近似熵的历史出力序列的波动特征提取、基于时间的周期特征构造、基于温度的光伏板温度特征构造等方法．提出基于残差连接和

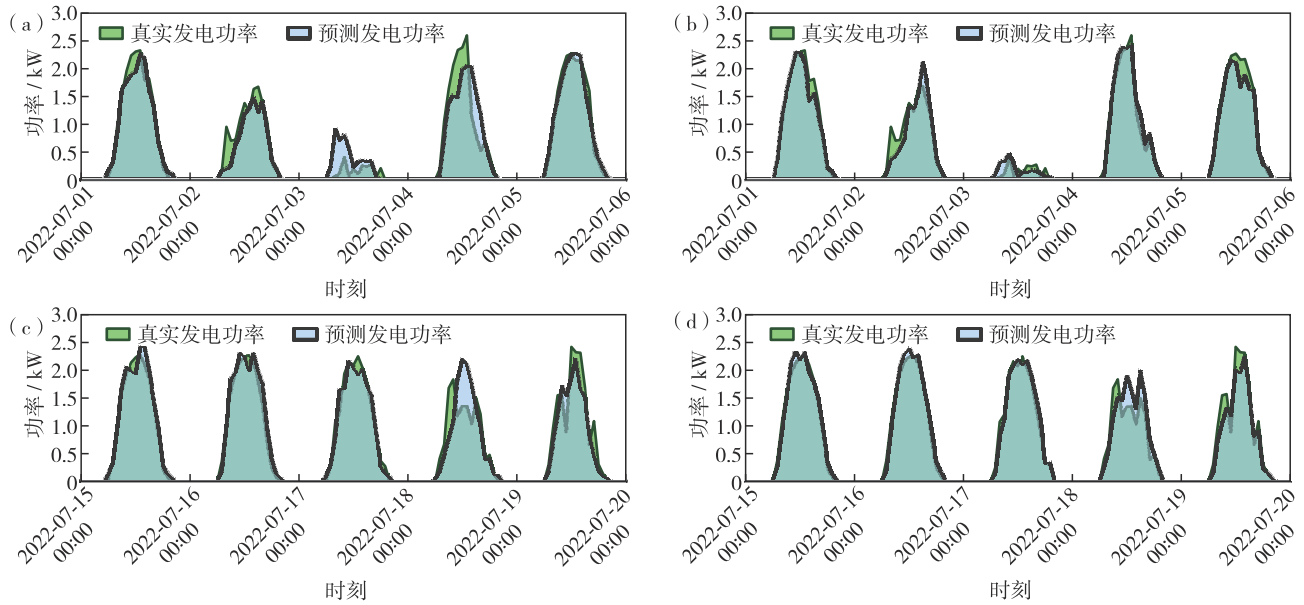

图 8　2022-07-01 至 2022-07-05(a)特征提取前和(b)特征提取后采用所提模型预测发电功率与真实发电功率的可视化结果对比；2022-07-15 至 2022-07-19(c)特征提取前和(d)特征提取后采用所提模型预测与真实发电功率的可视化结果对比

Fig. 8　**(Color online)** Comparison of visual results of predicted and real power generation using the proposed model (a) before feature extraction and (b) after feature extraction from 2022-07-01 to 2022-07-05. Comparison of visual results of predicted and real power generation using the proposed model (c) before feature extraction and (d) after feature extraction from 2022-07-15 to 2022-07-19.

多模型融合的发电功率预测模型，融合了基于kNN和 softmax 回归的相似日预测模型和不同差异的 TabNet、XGBoost 和 RF 模型，提升了分布式光伏发电功率预测的精度．通过对真实的某电力公司部分分布式光伏发电功率数据进行实验和分析，得到以下结论：

1）针对分布式光伏发电功率预测提出的特征提取方法能有效地大幅提升分布式光伏发电功率预测的精度，说明在分布式光伏发电功率预测中，特征提取方法非常重要，优秀的特征提取方法可提升发电功率预测的精度．

2）提出基于kNN 和 softmax 回归的相似日预测模型，以及基于多层模型融合的发电功率预测模型融合了多个模型的优势，不仅能有效提升分布式光伏发电预测精度，还能提升预测的稳定性．

3）特征提取方法和多层模型融合的发电功率预测模型结合后，进一步提升了模型精度，特别对于气象因素引起波动和突变的分布式光伏发电功率曲线，能准确的进行预测．

基金项目：广东省自然科学基金资助项目(2023A1515011667)；深圳市基础研究资助项目(JCYJ20220818100205012，JCYJ20210324093609026)

作者简介：吴明朗(429201375@qq.com)，深圳市国电科技通信有限公司助理工程师．研究方向：时空机器学习技术、新能源的功率预测技术．

引　　文：吴明朗，庞振江，洪海敏，等．基于残差的分布式光伏发电功率组合预测方法[J]．深圳大学学报理工版，2024，41(3)：293-302.

参考文献 / References：

[1] 丁明，王伟胜，王秀丽，等．大规模光伏发电对电力系统影响综述[J]．中国电机工程学报，2014，34(1)：1-14.
DING Ming, WANG Weisheng, WANG Xiuli, et al. A review on the effect of large-scale PV generation on power systems [J]. Proceedings of the CSEE, 2014, 34(1): 1-14. (in Chinese)

[2] 梁才浩，段献忠．分布式发电及其对电力系统的影响[J]．电力系统自动化，2001，25(12)：53-55.
LIANG Caihao, DUAN Xianzhong. Distributed generation and its impact on power system [J]. Automation of Electric Power Systems, 2001, 25(12): 53-55. (in Chinese)

[3] 赖昌伟，黎静华，陈博，等．光伏发电出力预测技术研究综述[J]．电工技术学报，2019，34(6)：1201-1217.
LAI Changwei, LI Jinghua, CHEN Bo, et al. Review of photovoltaic power output prediction technology [J]. Transactions of China Electrotechnical Society, 2019, 34(6): 1201-1217.

[4] 李丰君，王磊，赵健，等．基于天气融合和LSTM网络的分布式光伏短期功率预测方法[J]．中国电力，2022，55(11)：149-154.
LI Fengjun, WANG Lei, ZHAO Jian, et al. Research on distributed photovoltaic short-term power prediction method based on weather fusion and LSTM-net [J].

Electric Power, 2022, 55(11): 149-154. (in Chinese)

[5] BORISOV V, LEEMANN T, SEßLER K, et al. Deep neural networks and tabular data: a survey [J/OL]. IEEE Transactions on Neural Networks and Learning Systems. (2022-12-23) [2023-04-01]. https://doi.org/10.1109/TNNLS.2022.3229161.

[6] TAUNK K, DE S, VERMA S, et al. A brief review of nearest neighbor algorithm for learning and classification [C]// International Conference on Intelligent Computing and Control Systems. Piscataway, USA: IEEE, 2019: 1255-1260.

[7] BIAU G, DEVROYE L, LUGOSI G. Consistency of random forests and other averaging classifiers [J]. Journal of Machine Learning Research, 2008, 9: 2015-2033.

[8] 刘晓艳，王珏，姚铁锤，等. 基于时序数据处理的分布式光伏功率预测系统[J]. 数据与计算发展前沿，2021，3(4)：140-148.
LIU Xiaoyan, WANG Jue, YAO Tiechui, et al. A distributed photovoltaic power prediction system based on time series data processing [J]. Frontiers of Data & Computing, 2021, 3(4): 140-148. (in Chinese)

[9] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system [C]// In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: Association for Computing Machinery, 2016: 785-794.

[10] ARIK S Ö, PFISTER T. TabNet: attentive interpretable tabular learning [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(8): 6679-6687.

[11] 王彪，吕洋，陈中，等. 考虑信息时移的分布式光伏机理-数据混合驱动短期功率预测[J]. 电力系统自动化，2022，46(11)：67-74.
WANG Biao, LYU Yang, CHEN Zhong, et al. Hybrid mechanism-data-driven short-term power forecasting of distributed photovoltaic considering information time shift [J]. Automation of Electric Power Systems, 2022, 46(11): 67-74. (in Chinese)

[12] 赵波，薛美东，葛晓慧，等. 光伏发电系统输出功率计算方法研究[J]. 电网与清洁能源，2010，26(7)：19-24.
ZHAO Bo, XUE Meidong, GE Xiaohui, et al. Research on calculating methods of output power of the photovoltaic system [J]. Power System and Clean Energy, 2010, 26(7): 19-24. (in Chinese)

[13] 李光明，刘祖明，何京鸿，等. 基于多元线性回归模型的并网光伏发电系统发电量预测研究[J]. 现代电力，2011，28(2)：43-48.
LI Guangming, LIU Zuming, HE Jinghong, et al. Study on the generator forecasting of grid-connected PV power system based on multivariate linear regression model [J]. Modern Electric Power, 2011, 28(2): 43-48. (in Chinese)

[14] 傅美平，马红伟，毛建容. 基于相似日和最小二乘支持向量机的光伏发电短期预测[J]. 电力系统保护与控制，2012，40(16)：65-69.
FU Meiping, MA Hongwei, MAO Jianrong. Short-term photovoltaic power forecasting based on similar days and least square support vector machine [J]. Power System Protection and Control, 2012, 40(16): 65-69. (in Chinese)

[15] 茆美琴，龚文剑，张榴晨，等. 基于EEMD-SVM方法的光伏电站短期出力预测[J]. 中国电机工程学报，2013，33(34)：17-24.
MAO Meiqin, GONG Wenjian, ZHANG Liuchen, et al. Short-term photovoltaic generation forecasting based on EEMD-SVM combined method [J]. Proceedings of the CSEE, 2013, 33(34): 17-24. (in Chinese)

[16] 刘晓艳，王珏，姚铁锤，等. 基于卫星遥感的超短期分布式光伏功率预测[J]. 电工技术学报，2022，37(7)：1800-1809.
LIU Xiaoyan, WANG Jue, YAO Tiechui, et al. Ultra short-term distributed photovoltaic power prediction based on satellite remote sensing [J]. Transactions of China Electrotechnical Society, 2022, 37(7): 1800-1809. (in Chinese)

[17] SI Zhiyuan, YANG Ming, YU Yixiao, et al. A hybrid photovoltaic power prediction model based on multi-source data fusion and deep learning [C]// The 3rd Student Conference on Electrical Machines and Systems. Piscataway, USA: IEEE, 2020: 608-613.

[18] VANDEVENTER W, JAMEI E, THIRUNAVUKKARASU G S, et al. Short-term PV power forecasting using hybrid GASVM technique [J]. Renewable Energy, 2019, 140: 367-379.

[19] NICCOLAI A, DOLARA A, OGLIARI E. Hybrid PV power forecasting methods: a comparison of different approaches [J]. Energies, 2021, 14(2): 451.

[20] ABDEL-NASSER M, MAHMOUD K. Accurate photovoltaic power forecasting models using deep LSTM-RNN [J]. Neural Computing and Applications, 2019, 31(7): 2727-2740.

[21] 叶林，裴铭，路朋，等. 基于天气分型的短期光伏功率组合预测方法[J]. 电力系统自动化，2021，45(1)：44-54.
YE Lin, PEI Ming, LU Peng, et al. Combination forecasting method of short-term photovoltaic power based on weather classification [J]. Automation of Electric Power Systems, 2021, 45(1): 44-54. (in Chinese)

[22] 代倩，段善旭，蔡涛，等. 基于天气类型聚类识别的光伏系统短期无辐照度发电预测模型研究[J]. 中国电机工程学报，2011，31(34)：28-35.
DAI Qian, DUAN Shanxu, CAI Tao, et al. Short-term PV generation system forecasting model without irradiation based on weather type clustering [J]. Proceedings of the CSEE, 2011, 31(34): 28-35. (in Chinese)

【中文责编：英子；英文责编：木柯】