冷亮，肖艳姣，王珏，等，2026. 风暴雷达特征参量驱动的 TFM+CNN 龙卷识别算法[J]. 气象，52(11)：1346-1357. Leng L，Xiao Y J，Wang J，et al，2026. The TFM+CNN based tornado identification algorithm driven by radar feature parameters of storms [J]. Meteor Mon，52(11)：1346-1357(in Chinese).

风暴雷达特征参量驱动的 TFM+CNN 龙卷识别算法*

冷 亮[1,2,3,4] 肖艳姣[1,2,3,4] 王 珏[1,2,3,4] 王志斌[1,2,3,4]

1 全国暴雨研究中心，武汉 430205
2 中国气象局流域强降水重点开放实验室，武汉 430205
3 暴雨监测预警湖北省重点实验室，武汉 430205
4 中国气象局武汉暴雨研究所，武汉 430205

提 要： 针对常用参量阈值法识别龙卷风准确率不高的问题，提出一种基于风暴多个雷达特征量驱动的机器学习龙卷风识别方法。该方法首先基于我国 2003—2023 年的龙卷个例雷达基数据，计算并建立了风暴的雷达特征量数据集，包括 924 个有龙卷风暴、1174 个有显著 3DVS 的无龙卷风暴及 5658 个无显著 3DVS 的无龙卷风暴，然后利用 XGBoost 对特征量进行重要性排序，结合主成分分析选取前 14 个特征量作为模型的输入。采用 TFM 的编码器，并利用 CNN 提高输入向量维度和在模型尾端加入多个全连接层，构建了龙卷风识别的 TFM+CNN 深度学习模型。使用该模型进行三个二分类识别试验，分别针对有龙卷和无显著 3DVS 的无龙卷风暴，有龙卷和有显著 3DVS 的无龙卷风暴及有龙卷和无龙卷风暴，对比了 XGBoost 模型预报。结果表明：TFM+CNN 模型在三个二分类试验中的 CSI 分别为 0.84、0.71、0.73，POD 分别为 0.94、0.83、0.83，FAR 分别为 0.11、0.16、0.14；和 XGBoost 模型相比，TFM+CNN 模型除了在试验 1 和试验 3 的 FAR 分别高 0.05、0.03 以外，其余指标均相对更优；两种模型都在试验 2 的表现最差，表明有龙卷和有显著 3DVS 的无龙卷风暴最难区分。此外，TFM+CNN 模型在 ROC 下的 AUC 更大，表明其泛化能力更强。可见 TFM+CNN 模型有较强的龙卷识别能力。

关键词： 龙卷识别，TFM，CNN，XGBoost，风暴雷达特征参量

中图分类号： P413，TP301 **文献标志码：** A **DOI：** 10.7519/j.issn.1000-0526.2025.111801

The TFM+CNN Based Tornado Identification Algorithm Driven by Radar Feature Parameters of Storms

LENG Liang[1,2,3,4] XIAO Yanjiao[1,2,3,4] WANG Jue[1,2,3,4] WANG Zhibin[1,2,3,4]

1 Heavy Rainfall Research Center of China, Wuhan 430205
2 CMA Basin Heavy Rainfall Key Laboratory, Wuhan 430205
3 Hubei Key Laboratory for Heavy Rain Monitoring and Warning Research, Wuhan 430205
4 Institute of Heavy Rain, CMA, Wuhan 430205

Abstract: In response to the issue of insufficient accuracy in identifying tornadoes using conventional parameter threshold methods, in this study we propose a machine learning-based tornado identification method driven by multiple radar-derived storm features. With this method we first calculate and construct a storm feature dataset based on radar base data from tornado cases during 2003－2023 in China. The

* 湖北省自然科学基金项目(2022CFD012)、中国气象局重点创新团队(CMA2022ZD07)、湖北省重点研发计划项目(2023BCB119)和全国暴雨研究开放基金(BYKJ2024Z04)共同资助
2025 年 7 月 4 日收稿； 2026 年 7 月 31 日收修定稿
第一作者：冷亮，主要从事强对流人工智能识别技术研究. E-mail：lengliang@whihr.com.cn
通讯作者：肖艳姣，主要从事雷达资料处理与应用技术研究. E-mail：yanjiao.xiao@163.com

dataset contains 924 tornadic storms, 1174 non-tornadic storms with significant three-dimensional vortex signature (3DVS), and 5658 non-tornadic storms without significant 3DVS. Then, feature importance ranking is performed with the XGBoost hyper-parameter. Combined with principal component analysis, the top 14 features are selected as model inputs. We construct a TFM+CNN deep learning model for tornado identification by employing a TFM-based encoder, using CNN to increase the input vector dimensionality, and adding multiple fully connected layers at the end of the model. Moreover, with this model, we conduct three binary classification experiments, i.e., tornado storms and non-tornadic storms without significant 3DVS, tornado storms and non-tornadic storms with significant 3DVS, and tornado storms and non-tornadic storms. Comparative experiments of the TFM+CNN model with the XGBoost model are also performed. The results show that the TFM+CNN model achieves CSI values of 0.84, 0.71 and 0.73, POD values of 0.94, 0.83 and 0.83, and FAR values of 0.11, 0.16 and 0.14 in the three binary classification experiments, respectively. Compared to the XGBoost model, the TFM+CNN model performs better in most metrics, except for 0.05 and 0.03 higher FAR in experiments 1 and 3, respectively. The two models both perform worst in experiment 2, indicating that it is most challenging to distinguish the tornadic storms and the non-tornadic storms with significant 3DVS. Furthermore, the TFM+CNN model has a larger AUC under ROC, demonstrating its stronger generalization ability. Therefore, the TFM+CNN model is superior in identifying tornadoes.

Key words: tornado identification, TFM, CNN, XGBoost, radar-derived storm feature parameter

引　言

龙卷是从雷暴云底伸展到地面的漏斗状的剧烈旋转气流,通常发生在极不稳定的天气条件下,是破坏力极大的小尺度(几十米到几百米)灾害天气,其最强地表阵风风速可达 125～150 m·s^{-1}(郑永光等,2021),给人民的生命和财产造成巨大损失,如 2016 年 6 月 23 日盐城 EF4 级龙卷(张小玲等,2016;郑永光等,2016)、2021 年 5 月 14 日武汉蔡甸 EF3 级龙卷(李彩玲等,2022;王艺娟等,2023)造成多人死亡和受伤。龙卷因其空间尺度极小、生命史极短、形成迅速且随机性强、移动路径复杂多变等特点,给监测和预报带来了巨大困难。根据不同的形成机制,龙卷通常分为中气旋龙卷(也称为超级单体龙卷)和非中气旋龙卷(也称为非超级单体龙卷);根据不同的对流组织形式,可分为超级单体龙卷、线状对流龙卷和其他类型龙卷(无组织类型);根据是否由热带气旋(飓风或台风等)引发,可分为热带气旋龙卷和非热带气旋龙卷(郑永光等,2025a)。

美国是世界上受龙卷影响最严重的国家,平均每年发生约 1200 个龙卷(Taszarek et al,2020),也是最早开始监测和预报龙卷的国家。龙卷监测和临近预警主要依靠高时空分辨率的天气雷达资料,低层中气旋、龙卷涡旋特征(TVS)、龙卷碎屑特征(TDS)是龙卷监测的主要依据,钩状回波、有界弱回波区(BWER)、差分反射率因子(Z_{DR})柱、Z_{DR}弧与差分传播相移率的分离、下降的反射率因子核(DRC)等特征也都属于判断超级单体或龙卷的参考依据(郑永光等,2025b; 龚佃利等,2024)。基于对这些特征的认知,一些有助于龙卷识别的特征自动识别算法被研发,如中气旋探测算法(Zrnić et al,1985; Stumpf et al,1998)、TVS 识别算法(Wieler,1986)、龙卷探测算法(TDA)(Wakimoto and Wilson,1989; Mitchell et al,1998)和 TDS 识别算法(Ryzhkov et al,2005;Palmer et al,2011)等。由于上述算法多基于径向速度数据识别同距离圈上相邻方位速度差大于预设阈值的三维区域,单一参量和固定阈值的使用以及径向速度本身存在的噪声干扰和速度模糊等质量问题,使得复杂多变的龙卷风的识别率有限,龙卷预警存在较高的虚警率(郑永光等,2021),如非超级单体龙卷因不存在显著的三维涡旋特征(3DVS)容易被漏报,而非龙卷超级单体常因存在显著的 3DVS 而被虚报龙卷。为此,基于多参量的龙卷识别及其强度估计的模糊逻辑方法和机器学习方法(Smith et al,2015;2020;Gibbs and Bowers,2019;王志斌等,2022;Sessa and Trapp,2023)被发展起来。中国气象局武汉暴雨研究所(2020)在改进了中

气旋识别算法基础上，统计分析了2003—2018年中国33个超级单体龙卷个例，归纳了龙卷风暴的中气旋旋转速度和最大方位涡度切变的边界函数，使用模糊逻辑方法研发了超级单体龙卷识别算法，于2021年集成到SWAN 3.0中(郑永光等，2025b)。Snyder and Ryzhkov(2015)基于双偏振雷达观测将反射率因子、Z_{DR}、相关系数和基于径向速度计算的方位涡度切变等参量作为模糊逻辑的输入，研发了TDS识别算法。模糊逻辑方法需要人工专家基于有限个例的统计分析结果设计模糊规则库、输入参量及其隶属度函数，更适合于规则相对清晰、维度较低、且需要透明决策过程的应用场景，而从已有的认知中，产生龙卷的风暴及其环境复杂多变，机器学习方法在龙卷风这类高动态、多变量、小样本的极端天气识别和预报方面更具优势。Wang and Yu(2015)利用神经元自适应系统，提出一种基于神经模糊网络的龙卷检测算法。曾强宇等(2023)实现了基于随机森林的组网雷达龙卷检测算法。Sandmæl et al(2023)基于雷达反射率因子、径向速度、速度谱宽、双偏振量以及特征识别产品等多参量数据，使用随机森林技术发展了新的龙卷识别算法。Xie et al(2024)使用雷达最低3个仰角的反射率因子、径向速度和谱宽数据，发展了卷积神经网络(CNN)与Transformer(TFM)技术相结合的深度学习龙卷识别模型MTI-Net，CNN使用空间注意力单元模块提取局部细节，TFM通过注意力机制增强全局连接，该模型的龙卷识别产品准确率高于其他算法并已集成于短时临近预报系统SWAN 3.0(周康辉等，2025)。

本文基于2003—2023年龙卷个例的雷达体扫基数据，在识别风暴和提取诸多风暴属性雷达特征量的基础上，针对龙卷使用梯度提升决策树(XGBoost)(Chen，2016)对雷达特征量进行重要性排序，并使用XGBoost和TFM＋CNN(Vaswani et al，2017)两种模型建立了龙卷风识别模型，以期提高龙卷风暴识别准确率。

1 数据及处理方法

1.1 数 据

收集了2003—2023年的全国龙卷和疑似龙卷的发生日期、地点和时间的大致信息以及相应的多普勒天气雷达基数据。使用新一代天气雷达业务软件系统ROSE2.1，通过人工查看反射率因子、径向速度、中气旋和TVS等产品，对收集的龙卷信息进行清洗，剔除明显由下击暴流、飑线和阵风锋等引起的地面直线型大风个例；对于龙卷个例，确定龙卷发生的时间段和在雷达极坐标中的方位和斜距。本文所用雷达基数据选自上述龙卷个例中有中气旋或TVS的CINRAD/SA、CINRAD/SB和CINRAD/SAD型号的雷达，并且剔除了最低2个仰角采用双PRF模式采样径向速度的雷达基数据，共有101个龙卷日的244个龙卷风暴(1个龙卷风暴可能先后产生多个龙卷)的雷达基数据。未按形成机制、对流组织形式或是否由热带气旋引起对龙卷进行分类。雷达基数据采用VCP21体扫模式(6 min扫描9个固定仰角)或VCP11体扫模式(5 min扫描14个固定仰角)采样，雷达波束宽度约为1°，反射率因子库长为1 km或250 m，径向速度库长为250 m。

1.2 数据处理

1.2.1 风暴雷达特征量计算

对雷达基数据进行了杂波抑制(吴涛等，2013)和径向速度退模糊(肖艳姣等，2012；2016)处理后，使用雷暴单体识别追踪算法(SCIT)进行风暴识别追踪；使用二维局部线性最小二乘导数(LLSD)方法估计方位涡度切变，基于方位涡度切变数据，使用类似SCIT算法进行多阈值(分7级，缺省阈值分别为70、60、50、40、30、20、10，单位为$10^{-4}\,s^{-1}$)3DVS识别，并把3DVS与风暴关联，自动提取风暴单体的诸多属性特征量(肖艳姣等，2021)。本文选用以仰角分层的二维风暴分量特征量共28个，其中以二维涡旋特征(2DVS)参量为主，基于龙卷风暴雷达特征量数据，使用XGBoost模型进行雷达特征量重要性排序，结合主成分分析选取前14个特征量作为TMF＋CNN模型的输入进行训练(表1)，其中部分特征量的计算公式如下：

$$x_1 = 10 \times \frac{V_{j,k} - V_{j-1,k}}{2R_k \sin\left(\frac{\Delta\theta_{j,j-1}}{2}\right)} \tag{1}$$

$$x_2 = \frac{V_{\max} - V_{\min}}{2} \tag{2}$$

$$x_7 = \sum_{k=-M}^{M} \sum_{j=-N}^{N} \left[(V_{j,k} - V_{j-1,k}) \times W_B \times 2R_k \sin\left(\frac{\Delta\theta_{j,j-1}}{2}\right) \right] \tag{3}$$

$$x_{14}=\sum_{k=-M}^{N}\sum_{j=-N}^{M}\left[W_{\mathrm{B}}\times 2R_k\sin\left(\frac{\Delta\theta_{j,j-1}}{2}\right)\right] \quad (4)$$

式中:j、k 分别为方位和斜距序号,V 为径向速度(单位:$\mathrm{m\cdot s^{-1}}$),R_k 为斜距(单位:km),W_{B} 为径向速度距离库库长(单位:km),$\Delta\theta$ 为相邻径向的方位角差(单位:弧度)。

逾量转动动能的计算公式如下:

$$x_{21}=\frac{\rho\pi\Delta h r^2\times(x_2-rS_{\min})^2}{4} \quad (5)$$

式中:ρ 为空气密度(单位:$\mathrm{kg\cdot m^{-3}}$),取国际通用的标准大气密度;Δh 为厚度,由于 2DVS 在平面上,不存在厚度,故取 $\Delta h=1$;r 为最低阈值($10\times10^{-4}\ \mathrm{s^{-1}}$)的二维切变特征半径(单位:m),$S_{\min}$ 为维持中等强度中气旋的最小涡度切变阈值(可调,缺省值取 $50\times10^{-4}\ \mathrm{s^{-1}}$)。二维切变特征的逾量转动动能的单位为 $\mathrm{kg\cdot m\cdot s^{-2}}$,即 $\mathrm{J\cdot m^{-1}}$。当 $x_2-rS_{\min}<0$ 时,认为 $x_{21}=0$。x_{21} 的值越大,其超出维持最小中气旋的旋转动能越大,也就越有可能形成龙卷。

在极坐标下,取一个 $(2N+1)\times(2M+1)$ 大小的二维局地窗口,设窗口内各斜距库-方位的径向速度为 $V_{j,k}$,以窗口中心为基准,沿径向中心以远,k 为正,沿切向顺时针方向,j 为正,则窗口中心的方位涡度切变 x_{26} 的计算公式如下:

表 1　二维风暴分量雷达特征量信息

Table 1　Radar feature information of two-dimensional storm component

特征量名称	参量说明	单位	在 XGBoost 中的重要性比例/%
x_1	2DVS 内最大相邻方位速度差(NADV)处的差分方位切变	$10^{-4}\ \mathrm{s^{-1}}$	8.49
x_2	2DVS 的旋转速度,即最大、最小速度差的一半	$\mathrm{m\cdot s^{-1}}$	8.23
x_3	2DVS 内最大相邻方位速度差所在的方位	°	7.13
x_4	2DVS 内的最大速度	$\mathrm{m\cdot s^{-1}}$	6.30
x_5	2DVS 内的最小速度	$\mathrm{m\cdot s^{-1}}$	5.92
x_6	2DVS 内最大相邻方位速度差所在的斜距	km	4.82
x_7	2DVS 内各库的 NADV 与面积的乘积之和	$\mathrm{km^2\cdot m\cdot s^{-1}}$	4.33
x_8	2DVS 内最大速度所在方位	°	3.92
x_9	2DVS 内最大速度所在斜距	km	3.68
x_{10}	2DVS 内的平均反射率因子值	dBz	3.59
x_{11}	2DVS 内最小速度所在方位	°	3.19
x_{12}	以相邻方位速度差为权重的 2DVS 的经度位置	°	2.95
x_{13}	以相邻方位速度差为权重的 2DVS 的纬度位置	°	2.89
x_{14}	2DVS 的面积	$\mathrm{km^2}$	2.88
x_{15}	2DVS 的质心高度	km	2.69
x_{16}	2DVS 的质心方位	°	2.64
x_{17}	2DVS 的最大 NADV 等级	$10^{-4}\ \mathrm{s^{-1}}$	2.63
x_{18}	2DVS 内的最大相邻方位速度差	$\mathrm{m\cdot s^{-1}}$	2.63
x_{19}	2DVS 内的最大反射率因子值	dBz	2.59
x_{20}	2DVS 内的最小反射率因子值	dBz	2.55
x_{21}	2DVS 的逾量转动动能	$\mathrm{J\cdot m^{-1}}$	2.35
x_{22}	2DVS 的质心斜距	km	2.15
x_{23}	2DVS 的最大、最小速度对应的差分方位切变	$10^{-4}\ \mathrm{s^{-1}}$	2.06
x_{24}	2DVS 内的最大速度谱宽	$\mathrm{m\cdot s^{-1}}$	1.99
x_{25}	2DVS 内最小速度所在斜距	km	1.98
x_{26}	2DVS 的最大方位涡度切变	$10^{-4}\ \mathrm{s^{-1}}$	1.97
x_{27}	与低仰角垂直关联的 2DVS 之间的质心水平距离	km	1.73
x_{28}	2DVS 内最大最小速度所在处的水平距离	km	1.71

$$x_{26}=\frac{\sum_{k=-M}^{M}\sum_{j=-N}^{N}S_{j,k}V_{j,k}w_{j,k}}{\sum_{k=-M}^{M}\sum_{j=-N}^{N}S_{j,k}^2w_{j,k}} \quad (6)$$

式中:$S_{j,k}=(R+kW_{\mathrm{B}})\times j\Delta\varphi$,$R$ 为局地二维窗口中心离雷达的斜距(单位:km),$\Delta\varphi$ 为波束宽度,用弧度表示。$w_{j,k}$ 为权重,本文取 $w_{j,k}=1$。如果速度场中有气旋性旋转,则 $x_{26}>0$。

由于使用 LLSD 方法计算方位涡度切变时使用的是极坐标径向速度数据,其方位分辨率随着离开

雷达的距离 R 而变化，故所取局地窗口大小也随着 R 变化。设中气旋核的大小是某个常数 D（缺省 1.25 km），那么 N、M 的取值分别为：

$$N = (\text{int})\left(\frac{\frac{D}{W_B}-1}{2}\right) \tag{7}$$

$$M = (\text{int})\left(\frac{\frac{D}{R\Delta\varphi}-1}{2}\right) \tag{8}$$

式中：(int)表示四舍五入取整。如果 $M<1$，则令 $M=1$；如果 $M>13$，则令 $M=13$。

1.2.2 特征量归一化处理

由于各个雷达特征量值的动态变化范围很大，有些是量级上的差别，因此需要对雷达特征量值进行归一化处理，最终使其范围控制在(1,2)。

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} + 1 \tag{9}$$

式中：$\hat{x}$、x、$x_{\min}$、$x_{\max}$ 分别为某个特征量的归一化值、当前值、最小值和最大值。

1.2.3 特征量缺失值处理

3DVS 是由多个垂直关联的 2DVS 构成的，2DVS 个数除了与风暴本身特征有关外，还会受到波束阻挡、距离折叠、风暴离雷达的距离等因素的影响，因此不同的 3DVS 包含的 2DVS 个数、底层所在仰角等都会有差异，从而导致设置固定层数的 3DVS 的部分参量出现缺失。如果 1 个 3DVS 数据中出现 1 个 2DVS 的部分特征量为缺失值，则剔除该 3DVS 数据，这种情况比较少；如果 1 个 3DVS 数据中的每个 2DVS 的所有特征量都无缺失值，只是 2DVS 层数少于设置的数据层数，则使用一个小的数值（如 -1.0×10^9）对无数据层进行填充，这种情况很常见。这种数据缺失值处理方法，无论是对类似决策树模型还是对 TFM 模型都比较合理。树模型中选择走哪个分支是根据计算树的节点信息增益大小来决定的，树模型总是选择增益大的那一支，填充小值的分支因计算增益小而被弃选。在 TFM 模型中，填充的小值可以用系统中的缺失值进行替换，如 Pytorch 系统中的 torch.inf（PyTorch Team，2025），而系统中的缺失值会被忽略（Mask）。

通过上述数据处理后，本文建立了三类风暴雷达特征量数据集共 7756 个样本，其中第一类为有龙卷风暴雷达特征量数据集，共有 924 个样本，取龙卷发生期间或发生之前半小时内且最低 2 个仰角的最大相邻方位速度差≥20 $m\cdot s^{-1}$ 的 3DVS，同一个风暴在不同体扫时次的 3DVS 个数进行累加计算；第二类为有显著 3DVS 的无龙卷风暴雷达特征量数据集，共有 1174 个样本；第三类为无显著 3DVS 的无龙卷风暴雷达特征量数据集，共有 5658 个样本。每个样本都包含了 28 个特征量，通过筛选提供给模型使用的有 14 个特征量。

在上述三类数据的基础上，建立了有龙卷风暴和有显著 3DVS 的无龙卷风暴、有龙卷风暴和无显著 3DVS 的无龙卷风暴、有龙卷风暴和有/无显著 3DVS 的无龙卷风暴三个数据集；把这三个数据集按照 7∶3 的比例拆分为训练数据集和测试数据集；有龙卷风暴雷达特征量数据中有 636 个样本用于训练，273 个用于测试；有显著 3DVS 的无龙卷风暴雷达特征量数据中有 822 个样本用于训练，352 个用于测试；无显著 3DVS 的无龙卷风暴雷达特征量数据中有 3908 个样本用于训练，1675 个样本用于测试。另外，预留了 15 个龙卷和 75 个非龙卷风暴样本进行回代试验，包括 2020 年 6 月 27 日发生在湖北荆门的龙卷风暴和 2021 年 5 月 14 日发生在湖北武汉的龙卷风暴个例。

2 模型算法和方案

本文选择了 XGBoost 和 TFM＋CNN 两种模型进行对比试验。

2.1 XGBoost 模型算法

XGBoost 是高效的梯度提升决策树（GBDT）算法（Chen，2016），是集成学习中提升方法的一种改进版本。XGBoost 本质上仍然是一个基于决策树的集成模型，但其对传统 GBDT（仅用一阶梯度）做了很多优化处理：(1)利用二阶信息加速收敛；(2)通过多个系数控制树结构的复杂度，避免过拟合；(3)基于分裂增益最大化选择特征和分割点，支持近似算法处理大规模数据等优化处理。通过上述优化处理，XGBoost 在速度、精度、鲁棒性和可扩展性方面表现更优。

XGBoost 模型算法主要包含三个部分：(1)目标函数优化，XGBoost 的目标函数由损失函数和正则化项两部分组成；(2)决策树构建，将目标函数转换为关于叶子节点权重的形式，使用 CART（分类与回归树）作为基学习器，每次迭代生成新树拟合残差；(3)稀疏数据处理，通过缺失值的自动处理、稀疏

特征的压缩存储、稀疏感知的分裂算法、正规化与稀疏性控制等措施,实现高效准确学习。XGBoost 的稀疏数据处理能力是其核心优势之一,用户无需复杂预处理即可直接训练。

2.2　TFM+CNN 算法

TFM 通过自注意力、多头注意力、位置编码、残差连接及层归一化等手段解决循环神经网络和长短期记忆网络不能很好地解决长距离的依赖及并行化处理等问题。

2.2.1　模型输入处理

模型最初由所有仰角含时间序列的向量(取表 1 中排序前 14 的特征量)组成,通过大量的试验分析发现,使用所有仰角的向量数据并不能提高模型预测精度,主要原因是仰角层越高,缺失值越多,模型学习到的信息越来越少,因此,最终选取低层 6 个仰角的特征量作为模型的输入,特征量排列方法如表 2。

表 2　特征量排列方法

Table 2　Feature ordering method

仰角	特征量
1	$x_1,x_2,\cdots,x_{14}$
2	$x_1,x_2,\cdots,x_{14}$
…	$x_1,x_2,\cdots,x_{14}$
6	$x_1,x_2,\cdots,x_{14}$

2.2.2　模型结构

完整的 TFM 架构由编码和解码模块两部分组成,本文所要完成的任务只是识别龙卷,因此只选取编码器模块(图 1),其主要作用是提取特征信息后给解码器使用,但也可以利用这些信息进行分类识别。对于 TFM 模型的编码模块而言,在输入的 6 个仰角序列的每一个向量中只包含 14 个维度的特征量。因维度太少所含的信息量有限,如果使用多头自注意机制进行拆分时,每个头包含的信息更有限,所以使用一维 CNN 把向量维度从 14 个提升到 120 个,在信息通过 TFM 的编码模块后再次利用一维 CNN 把向量维度从 120 个提升到 384 个,并在模型尾端加入多个全连接层。由于提升向量维度时不包含位置信息,因此还要使用正弦/余弦函数生成位置信息。其中多头注意力层包含了键向量($\boldsymbol{K}$)、

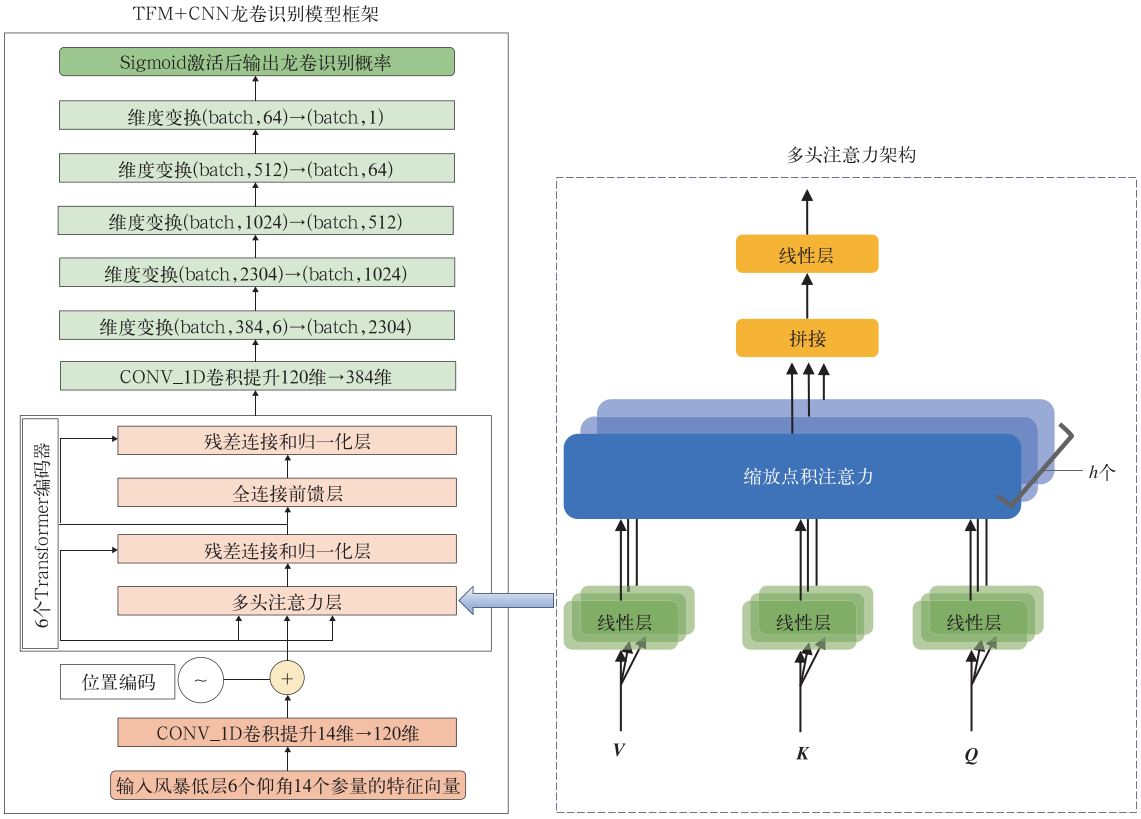


图 1　TFM+CNN 龙卷识别模型框架和多头注意力层架构图

Fig. 1　Diagram of TFM+CNN tornado identification model framework and multi-head attention layer architecture

查询向量($\boldsymbol{Q}$)和值向量($\boldsymbol{V}$)的计算,将 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 的特征维度 d 分割为 h 个头并行计算注意力后拼接,将每个头 $N\times d_k$(d_k 为一个单头的维度)输出拼接为 $N\times d$ 矩阵,再通过线性层映射回模型维度。

2.3 模型试验和分析

在模型试验前,通过主成分分析有龙卷和无龙卷风暴数据集,发现前 14 个特征量(表 1)累计贡献接近 100%(图 2),特别是前 3 个包含了约 73%的信息(图 3),说明在选取因子时有些因子相关性较大,因此只选取了前 14 个特征量(表 1)进行学习。

采用 3 个二分类模型进行龙卷风暴识别,第 1 个二分类为有龙卷和无显著 3DVS 的无龙卷风暴,第 2 个二分类为有龙卷和有显著 3DVS 的无龙卷风暴,第 3 个二分类为有龙卷和无龙卷风暴(包含有/无显著 3DVS 的风暴)。采用常规参量阈值法就很容易实现第 1 个二分类,但难以实现第 2 个二分类。在实际业务中,可以先用常规参量阈值法排除无显著 3DVS 的无龙卷风暴,再用机器学习进行第 2 个二分类处理。

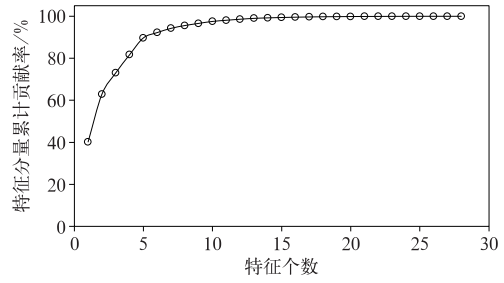


图 2 主成分特征分量累计贡献分布

Fig. 2 Cumulative contribution plot of principal component eigenvalues

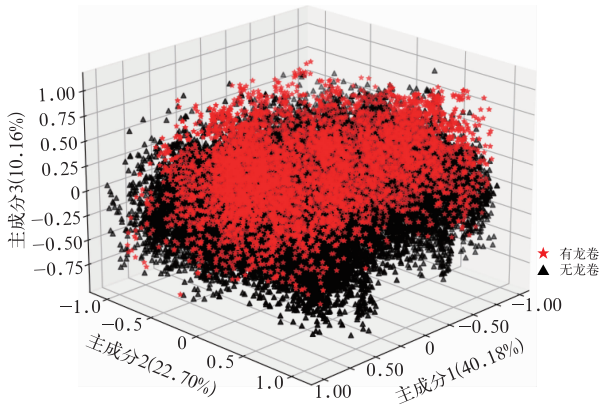


图 3 特征量主成分分析前 3 个分量三维散点分布

Fig. 3 Three-dimensional scatter plot of the first three principal components from feature principal component analysis

2.3.1 XGBoost 模型

通过网格寻优算法,选择的超参数如表 3 所示,其中,树的数量 n_estimators=120,树的深度 max_depth=15,正负样本的比例 scale_pos_weight=6,损失函数的最低值 $\gamma=0.15$,L2 正则化惩罚系数 reg_λ=3。

在有龙卷和无显著 3DVS 的无龙卷风暴数据集中,正负样本的比例为 1∶6,其中正样例为龙卷风暴的,在测试集上,识别准确率为 96.0%,受试者工作特征(ROC)曲线下的面积(AUC,用于衡量二分类问题机器学习算法的泛化能力)为 0.976(表 4,图 4a);有龙卷和有显著 3DVS 的无龙卷风暴数据集的正负样本比例接近 1∶1,在测试集上,识别准确率为 80.5%,ROC 曲线下的 AUC 为 0.872(表 4,图 4a);在有龙卷和无龙卷风暴数据集中,正负样本的比例为 1∶6,其中正样例为龙卷风暴的,在测试集上,识别准确率为 95.3%,ROC 曲线下的 AUC 为 0.975(表 4,图 4a)。

2.3.2 TFM+CNN 模型

在有龙卷和无显著 3DVS 的无龙卷风暴数据集中,正负样本的比例为 1∶6,其中正样本为龙卷风暴的,在测试集上,识别准确率为 97.8%,ROC 曲线下的 AUC 为 0.983(表 4,图 4b);有龙卷和有显著 3DVS 的无龙卷风暴数据集的正负样本比例接近 1∶1,在测试集上,识别准确率为 86.3%,ROC 曲

表 3 XGBoost 超参数

Table 3 XGBoost hyperparameter

超参数	值	超参数	值
scale_pos_weight	6	min_child_weight	1
learning_rate	0.01	max_delta_step	0
max_depth	15	subsample	0.85
n_estimators	120	reg_α	0
silent	true	reg_λ	3
objective	binary:logistic	seed	1440
γ	0.15	missing	none

表 4 模型测试集的评估指标和结果

Table 4 Evaluation metrics and results of the test sets

数据集	模型	准确率/%	AUC
有龙卷和无显著3DVS的无龙卷风暴	XGBoost	96.0	0.976
	TFM+CNN	97.8	0.983
有龙卷和有显著3DVS的无龙卷风暴	XGBoost	80.5	0.872
	TFM+CNN	86.3	0.903
有龙卷和无龙卷风暴	XGBoost	95.3	0.975
	TFM+CNN	96.0	0.979

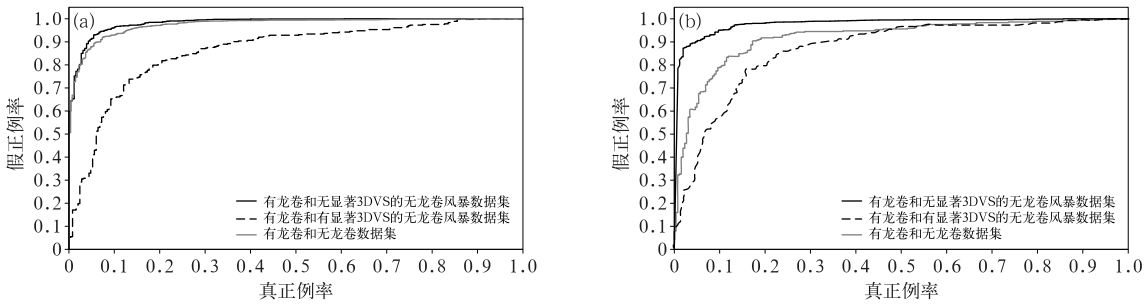


图 4　(a)XGBoost 和(b)TFM+CNN 的 ROC 曲线对比

Fig. 4　Comparison of ROC curves between (a) XGBoost and (b) TFM+CNN

线下的 AUC 为 0.903(表 4,图 4b);在有龙卷和无龙卷风暴数据集中,正负样本的比例为 1∶6,其中正样例为龙卷风暴的,在测试集上,识别准确率为 96.0%,ROC 曲线下的 AUC 为 0.979(表 4,图 4b)。

TFM+CNN 模型中的 TFM 部分参数通过试验确定(表 5),由 6 个子编码器和 4 个自注意力头组成模型参数的结果较优,优化器采用 Adam,损失函数由二元交叉熵损失函数构成,输入部分由于某些层会因为多种原因导致数据缺失,采用 Mask 手段替代这一层数据,模型在计算时会屏蔽掉这一层;另外输入层的数据维度提高到 240 个维度后,模型训练易出现过拟合情况;输入数据经过模型后,在模型的尾端增加了 5 个全连接层,增加模型深度,提高了识别准确率。模型训练 Epoch 数为 5000,每训练完一轮后对测试数据进行准确率、精确率、召回率、$F1$ 等指标的计算,其中模型识别龙卷发生概率的阈值根据数据集的正负样本比例确定,在有龙卷和无显著 3DVS 的无龙卷风暴数据集上的值为 86%,在有龙卷和有显著 3DVS 的无龙卷风暴数据集上的值为 56%,在有龙卷和无龙卷风暴数据集上的值为 88%。模型在上述三个数据类别进行训练时,在训练到 3500 个 Epoch 后损失函数值趋于稳定。

表 5　TFM+CNN 模型试验结果

Table 5　Experiment results of the TFM+CNN model

子编码器个数	自注意力头数/个	有龙卷和无显著 3DVS 的无龙卷风暴数据集在测试集上的准确率/%
2	4	96.6
4	4	96.8
6	4	97.8
2	6	96.3
4	6	97.1
6	6	97.4

TFM+CNN 模型在识别准确率和 AUC 方面都优于 XGBoost 模型,其中在有龙卷和无显著 3DVS 的无龙卷风暴数据集上识别准确率较 XGBoost 模型高 1.8%。

2.3.3　评　估

采用气象上常用三种评价指标,即击中率(POD)、虚警率(FAR)和临界成功指数(CSI)进行应用评估(Wang and Yu,2015),结果见表 6。从表中可见,TFM+CNN 模型除 FAR 在有龙卷和无显著 3DVS 的无龙卷风暴数据集、有龙卷和无龙卷风暴数据集上比 XGBoost 分别高 0.05 和 0.03 以外,其余指标均优于 XGBoost,其中 CSI 高 0.08～0.11,POD 高 0.12～0.16,在有龙卷和有显著 3DVS 的无龙卷风暴数据集上的 FAR 则偏低 0.03。

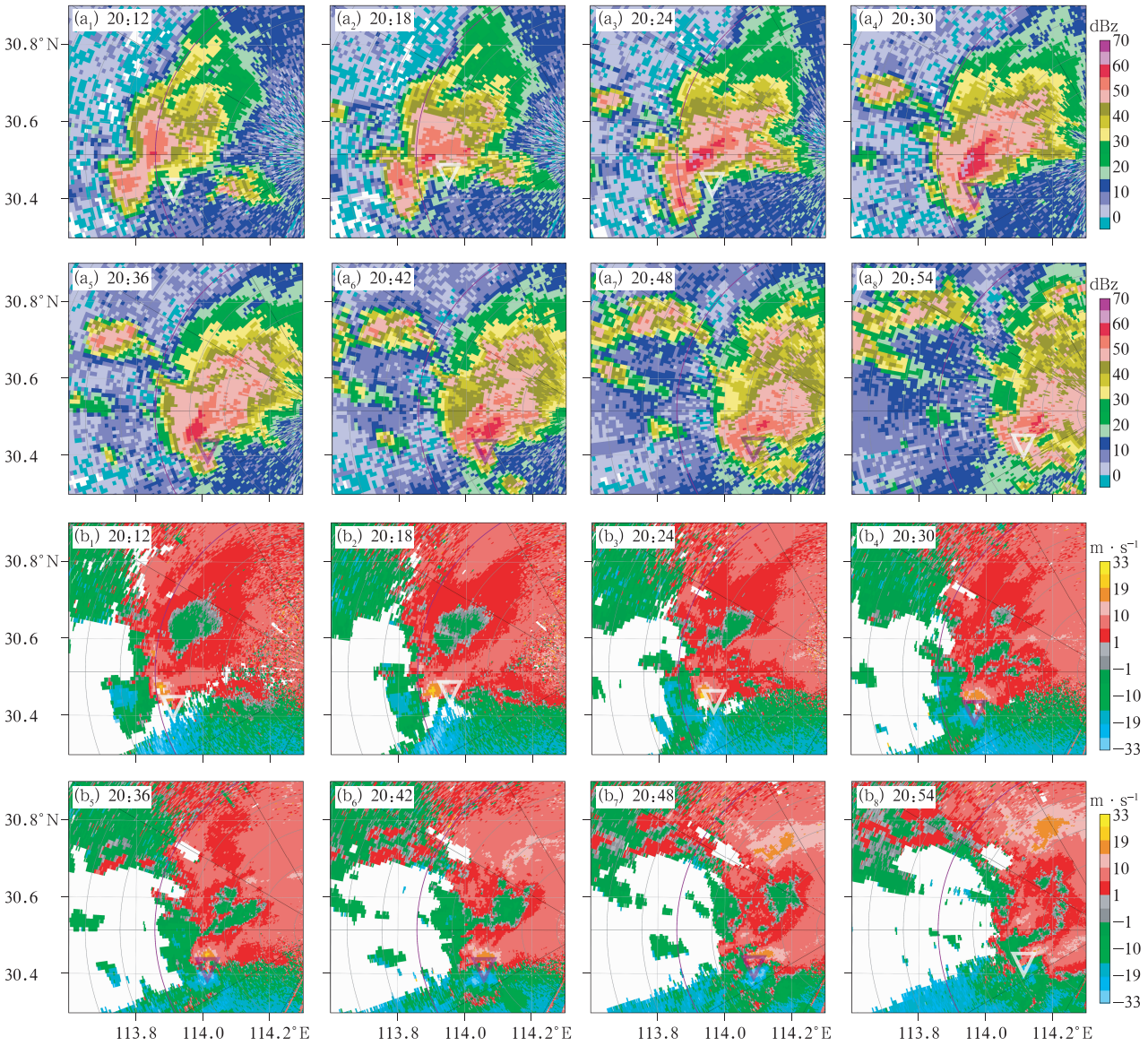
表 6　TFM+CNN 和 XGBoost 模型评估结果

Table 6　Evaluation indicators and results of TFM+CNN and XGBoost models

试验	模型	数据集	CSI	POD	FAR
试验 1	XGBoost	有龙卷和无显著 3DVS 的无龙卷风暴	0.74	0.78	0.06
	TFM+CNN		0.84	0.94	0.11
试验 2	XGBoost	有龙卷和有显著 3DVS 的无龙卷风暴	0.60	0.70	0.19
	TFM+CNN		0.71	0.83	0.16
试验 3	XGBoost	有龙卷和无龙卷风暴	0.65	0.71	0.11
	TFM+CNN		0.73	0.83	0.14

3 模型应用

个例 1:2021 年 5 月 14 日,在中低层西南暖湿气流作用下,武汉的汉口和蔡甸形成暖中心,并异常增温,傍晚冷空气入侵抬升触发导致一个风暴单体快速发展成为龙卷超级单体风暴,20:33(北京时,下同)在武汉蔡甸发生了 EF3 级龙卷风,维持约 21 min。利用 19:54—20:54(雷达基数据文件时间为体扫开始时间)的共 11 个体扫时次的武汉雷达数据,采用 TFM+CNN 模型识别龙卷,结果发现从 20:12—20:54 连续 8 个体扫识别到龙卷(图 5 和表 7),其他



注:白色倒三角表示识别有但实况无龙卷,紫色倒三角表示识别有且实况有龙卷。

图 5 2021 年 5 月 14 日 20:12—20:54 TFM+CNN 模型识别武汉龙卷结果(三角)及 1.5°仰角(a)反射率因子和(b)径向速度(填色)

Fig. 5 Identification result (triangle) of the Wuhan tornado by TFM+CNN model, along with the (a) reflectivity factor and (b) radial velocity (colored) at 1.5° elevation from 20:12 BT to 20:54 BT 14 May 2021

表 7　2021 年 5 月 14 日武汉龙卷 TFM+CNN 模型识别结果和 0.5°仰角 2DVS 特征参量值

Table 7　Identification result of the Wuhan tornado by TFM+CNN model and the 2DVS feature parameter values at 0.5° elevation on 14 May 2021

雷达体扫时间(BT)	位置	$x_1/(10^{-4}\ s^{-1})$	$x_2/(m\cdot s^{-1})$	$x_7/(km^2\cdot m\cdot s^{-1})$	x_{10}/dBz	识别结果	实况
20:12	30.403°N、113.892°E	452.5	22.9	858.4	1.9	有	无
20:18	30.429°N、113.926°E	589.5	30.1	1328.2	6.7	有	无
20:24	30.425°N、113.987°E	495.3	29.1	887.4	−5.8	有	无
20:30	30.417°N、113.958°E	507.5	28.1	356.7	19.2	有	有
20:36	30.414°N、114.006°E	826.0	29.6	759.8	33.3	有	有
20:42	30.421°N、114.043°E	831.4	28.4	406.0	35.1	有	有
20:48	30.420°N、114.079°E	849.1	25.9	430.0	51.2	有	有
20:54	30.424°N、114.109°E	906.3	22.2	475.0	46.5	有	无

3 个体扫时次未识别到龙卷。在龙卷发生时间段 20:33—20:54(对应雷达体扫时间 20:30—20:48)的识别成功率为 100%,在龙卷发生前有 3 个体扫时次识别有龙卷,首次识别有龙卷比龙卷开始发生时间提前 21 min,在龙卷结束但雷达体扫开始时次也识别到龙卷,这 4 个时次也应该为正确识别。

个例 2:2020 年 6 月 27—28 日,受梅雨锋影响,湖北荆门发生暴雨过程,27 日 19:00 左右荆门市掇刀区凤凰水库附近发生了 EF2 级龙卷,维持时间约 30 min,利用荆州雷达观测的 17:24—19:42 共 24 个体扫时次的数据,采用 TFM+CNN 模型识别龙卷,在龙卷发生时间段(19:00—19:30)有 5 个体扫时次识别有龙卷,1 个体扫时次(19:06)漏识别,识别成功率为 83.3%;首次识别有龙卷的时间为 19:00,相比龙卷开始发生时间没有提前时间量;在其他体扫时次,都未识别出龙卷(图略)。

4　结论与讨论

收集整理了 2003—2023 年的全国龙卷信息和相关的天气雷达体扫基数据,使用风暴识别追踪和 3DVS 识别算法,提取了风暴的 28 个雷达特征量,建立了 924 个有龙卷风暴、1174 个有显著 3DVS 和 5658 个无显著 3DVS 的无龙卷风暴特征量数据集。利用 XGBoost 对特征量进行重要性排序,利用主成分分析选取前 14 个特征量作为模型的输入。采用 TFM 的编码器,并利用 CNN 提高输入向量维度并在模型尾端加入多个全连接层,构建了龙卷风识别的 TFM+CNN 深度学习模型,进行 3 个二分类识别试验:(1)有龙卷和无显著 3DVS 的无龙卷风暴,(2)有龙卷和有显著 3DVS 的无龙卷风暴,(3)有和无龙卷风暴,并对比试验了 XGBoost 模型,得到如下结论。

(1)在模型参数选择方面,TFM+CNN 网络构建时,选择子网络数和自注意力头的个数,对深度学习网络训练效果有一定影响。针对本数据集建立的 TFM+CNN 模型选取 6 层子网络和 4 个自注意头较为合适;XGBoost 通过网格寻优后,当参数选择树的数量为 120、树的深度为 15 时效果最好。

(2)两种模型对比,TFM+CNN 模型在三个二分类试验中的 CSI 分别为 0.84、0.71、0.73,POD 分别为 0.94、0.83、0.83,FAR 分别为 0.11、0.16、0.14;和 XGBoost 模型相比,TFM+CNN 模型除了在试验 1 和试验 3 的 FAR 分别高 0.05、0.03 以外,其余指标均相对更优。此外,TFM+CNN 模型在 ROC 下的 AUC 更大,表明其泛化能力更强。

(3)在不同数据集上的表现,TFM+CNN 模型和 XGBoost 模型都在试验 2 的表现最差,表明有龙卷和有显著 3DVS 的无龙卷风暴最难区分,更容易导致虚警。

(4)模型应用方面,针对 2021 年 5 月 14 日武汉龙卷,TFM+CNN 模型的识别成功率为 100%,预警提前时间为 21 min;针对 2020 年 6 月 27—28 日荆门龙卷,识别成功率为 83.3%,无预警提前时间。

本研究构建了 TFM+CNN 深度学习模型用于龙卷识别,其分类性能优于传统的机器学习方法如 XGBoost,主要得益于自注意力机制与 CNN 结合实现了“全局关联+局部深化”的特征提取。试验表明,有龙卷与有显著 3DVS 的无龙卷风暴最难区分,是虚警主因。模型在典型龙卷案例中表现出良好预警能力,但识别性能易受雷达径向速度质量、地物杂波、雷达波束阻挡、雷达扫描时间分辨率和龙卷的强度、尺寸、距离雷达远近等诸多因素影响。未来需进一步优化雷达资料质控、扩充多源数据、增强模型可解释性,并推动其在业务应用中的人机协同等,提高识别预警效果。

参考文献

龚佃利，朱君鉴，林曲凤，等，2024. 一次超级单体引发的秋季特大冰雹和龙卷雷达探测分析[J]. 气象，50(5)：561-576. Gong D L，Zhu J J，Lin Q F，et al，2024. Radar observation analysis of severe hailstorm and tornado caused by a supercell in autumn[J]. Meteor Mon，50(5)：561-576(in Chinese).

李彩玲，杨波，姚聃，等，2022. 2021年5月14日武汉龙卷现场调查及雷达探测特征分析[J]. 暴雨灾害，41(2)：142-150. Li C L，Yang B，Yao D，et al，2022. Damage survey and radar characteristics of the 14th May 2021 Wuhan tornado in Hubei Province[J]. Torr Rain Dis，41(2)：142-150(in Chinese).

王艺娟，姚聃，梁旭东，等，2023. 2021年苏州与武汉两地龙卷的环境条件与雷达特征对比分析[J]. 暴雨灾害，42(4)：415-426. Wang Y J，Yao D，Liang X D，et al，2023. Comparative analysis of environmental conditions and radar characteristics for two tornadoes occurred in Suzhou and Wuhan in 2021[J]. Torr Rain Dis，42(4)：415-426(in Chinese).

王志斌，肖艳姣，王珏，等，2022. 基于卷积神经网和SVM雷电监测预警[J]. 自然灾害学报，31(1)：219-225. Wang Z B，Xiao Y J，Wang J，et al，2022. Lightning nowcasting and warning based on convolutional neural network and SVM[J]. J Nat Dis，31(1)：219-225(in Chinese).

吴涛，万玉发，沃伟锋，等，2013. SWAN系统中雷达反射率因子质量控制算法及其应用[J]. 气象科技，41(5)：809-817. Wu T，Wan Y F，Wo W F，et al，2013. Design and application of radar reflectivity quality control algorithm in SWAN[J]. Meteor Sci Technol，41(5)：809-817(in Chinese).

肖艳姣，万玉发，王珏，等，2012. 一种自动多普勒雷达速度退模糊算法研究[J]. 高原气象，31(4)：1119-1128. Xiao Y J，Wan Y F，Wang J，et al，2012. Study of an automated Doppler radar velocity dealiasing algorithm[J]. Plateau Meteor，31(4)：1119-1128(in Chinese).

肖艳姣，万玉发，王志斌，2016. 多普勒天气雷达双PRF径向速度资料质量控制[J]. 高原气象，35(4)：1112-1122. Xiao Y J，Wan Y F，Wang Z B，2016. Quality control of dual PRF velocity data for Doppler weather radars[J]. Plateau Meteor，35(4)：1112-1122(in Chinese).

肖艳姣，王珏，王志斌，等，2021. 基于S波段新一代天气雷达观测的下击暴流临近预报方法[J]. 气象，47(8)：919-931. Xiao Y J，Wang J，Wang Z B，et al，2021. A downburst nowcasting method based on observations of S-band new generation weather radar[J]. Meteor Mon，47(8)：919-931(in Chinese).

曾强宇，卿智鹏，陈亚军，等，2023. 基于随机森林的组网雷达龙卷检测算法[J]. 热带气象学报，39(6)：825-837. Zeng Q Y，Qing Z P，Chen Y J，et al，2023. Tornado detection algorithm based on random forest in radar network[J]. J Trop Meteor，39(6)：825-837(in Chinese).

张小玲，杨波，朱文剑，等，2016. 2016年6月23日江苏阜宁EF4级龙卷天气分析[J]. 气象，42(11)：1304-1314. Zhang X L，Yang B，Zhu W J，et al，2016. Analysis of the EF4 tornado in Funing County，Jiangsu Province on 23 June 2016[J]. Meteor Mon，42(11)：1304-1314(in Chinese).

郑永光，刘菲凡，张恒进，2021. 中国龙卷研究进展[J]. 气象，47(11)：1319-1335. Zheng Y G，Liu F F，Zhang H J，2021. Advances in tornado research in China[J]. Meteor Mon，47(11)：1319-1335(in Chinese).

郑永光，曹艳察，周晓敏，2025a. 龙卷发生环境与机理及灾情调查研究进展[J]. 气象学报，83(4)：990-1006. Zheng Y G，Cao Y C，Zhou X M，2025a. Advances in research on environments，formation mechanisms，and damage survey of tornadoes[J]. Acta Meteor Sin，83(4)：990-1006(in Chinese).

郑永光，杨波，周康辉，等，2025b. 龙卷监测预报预警技术进展[J]. 海洋气象学报，45(2)：1-13. Zheng Y G，Yang B，Zhou K H，et al，2025b. Advances in monitoring，forecasting and early warning technology of tornado[J]. J Marine Meteor，45(2)：1-13(in Chinese).

郑永光，朱文剑，姚聃，等，2016. 风速等级标准与2016年6月23日阜宁龙卷强度估计[J]. 气象，42(11)：1289-1303. Zheng Y G，Zhu W J，Yao D，et al，2016. Wind speed scales and rating of the intensity of the 23 June 2016 tornado in Funing County，Jiangsu Province[J]. Meteor Mon，42(11)：1289-1303(in Chinese).

中国气象局武汉暴雨研究所，2020-11-13. 一种龙卷涡旋特征自动识别算法：中国，111929687A[P]. Institute of Heavy Rain，CMA Wuhan，2020-11-13. Automatic identification algorithm for tornado vortex characteristics：CN，111929687A[P](in Chinese).

周康辉，郑永光，杨波，等，2025. 强对流天气客观临近预报技术进展和展望[J]. 气象学报，83(3)：798-812. Zhou K H，Zheng Y G，Yang B，et al，2025. Objective nowcasting of severe convective weather：technological progress and outlook[J]. Acta Meteor Sin，83(3)：798-812(in Chinese).

Chen T Q，2016. XGBoost documentation[EB/OL]. https：//xgboost.readthedocs.io/en/stable/.

Gibbs J G，Bowers B R，2019. Techniques and thresholds of significance for using WSR-88D velocity data to anticipate significant tornadoes[J]. J Operational Meteor，7(9)：117-137.

Mitchell E D W，Vasiloff S V，Stumpf G J，et al，1998. The national severe storms laboratory tornado detection algorithm[J]. Wea Forecasting，13(2)：352-366.

Palmer R D，Bodine D，Kumjian M，et al，2011. Observations of the 10 May 2010 tornado outbreak using OU-PRIME：potential for new science with high-resolution polarimetric radar[J]. Bull Amer Meteor Soc，92(7)：871-891.

PyTorch Team，2025. PyTorch tutorials. https：//docs.pytorch.org/tutorials/.

Ryzhkov A V，Schuur T J，Burgess D W，et al，2005. Polarimetric tornado detection[J]. J Appl Meteor Climatol，44(5)：557-570.

Sandmæl T N，Smith B R，Reinhart A E，et al，2023. The tornado probability algorithm：a probabilistic machine learning tornadic circulation detection algorithm[J]. Wea Forecasting，38(3)：445-466.

Sessa M F，Trapp R J，2023. Environmental and radar-derived predictors of tornado intensity within ongoing convective storms[J]. J

Opera Meteor,11(5):49-71.

Smith B T, Thompson R L, Dean A R, et al, 2015. Diagnosing the conditional probability of tornado damage rating using environmental and radar attributes[J]. Wea Forecasting, 30(4): 914-932.

Smith B T, Thompson R L, Speheger D A, et al, 2020. WSR-88D tornado intensity estimates. Part Ⅰ: real-time probabilities of peak tornado wind speeds[J]. Wea Forecasting, 35(6): 2479-2492.

Snyder J C, Ryzhkov A V, 2015. Automated detection of polarimetric tornadic debris signatures using a hydrometeor classification algorithm[J]. J Appl Meteor Climatol, 54(9): 1861-1870.

Stumpf G J, Witt A, Mitchell E D, et al, 1998. The national severe storms laboratory mesocyclone detection algorithm for the WSR-88D[J]. Wea Forecasting, 13(2): 304-326.

Taszarek M, Allen J T, Groenemeijer P, et al, 2020. Severe convective storms across Europe and the United States. Part Ⅰ: climatology of lightning, large hail, severe wind, and tornadoes[J]. J Climate, 33(23): 10239-10261.

Vaswani A, Shazeer N, Parmar N, et al, 2017. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 5998-6008.

Wakimoto R M, Wilson J W, 1989. Non-supercell tornadoes[J]. Mon Wea Rev, 117(6): 1113-1140.

Wang Y D, Yu T Y, 2015. Novel tornado detection using an adaptive neuro-fuzzy system with S-band polarimetric weather radar[J]. J Atmos Oceanic Technol, 32(2): 195-208.

Wieler J G, 1986. Real-time automated detection of mesocyclones and tornadic vortex signatures[J]. J Atmos Oceanic Technol, 3(1): 98-113.

Xie J Y, Zhou K H, Chen H N, et al, 2024. Multi-task learning for tornado identification using Doppler radar data[J]. Geophys Res Lett, 51(11): e2024GL108809.

Zrnić D S, Burgess D W, Hennington L D, 1985. Automatic detection of mesocyclonic shear with Doppler radar[J]. J Atmos Oceanic Technol, 2(4): 425-438.

(本文责编:张芳)