China Instrument and Control Society
Instrumentation

INSTRUMENTATION, Vol. 12, No. 4, December 2025 https://doi.org/10.15878/j.instr.202500299

Article

Lightning Warning Algorithm Based on Multi-Station Atmospheric Electric Field and Data Augmentation

Tiantian Yu[1,2], Haitao Wang[3], Wei Xu[1*], Yan Liu[2,4]

[1] Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing 210044, China

[2] Key Laboratory of Big Data & Artificial Intelligence in Transportation, Ministry of Education (Beijing Jiaotong University), Beijing 100044, China

[3] The PLA Unit 63729, Shanxi 030000, China

[4] Nanjing Innovation Institute for Atmospheric Sciences, Chinese Academy of Meteorological Sciences-Jiangsu Meteorological Service, Key Laboratory of Transportation Meteorology of CMA /Jiangsu Key Laboratory of Severe Storm Disaster Risk, Naning 210041, China

[*] Corresponding author email: xw@nuist.edu.cn

Abstract: An effective lightning warning system can ensure the safety of aircraft and promote the development of a low-altitude economy. Compared with weather radars, ground-based atmospheric electric field mills can monitor electric field variations in low-altitude regions in real-time without being affected by ground clutter. To address current challenges in lightning warning methods based on atmospheric electric field data—such as limited lightning location samples and a high false alarm rate (FAR)—this thesis proposes a lightning warning model that integrates multi-station atmospheric electric field data with meteorological variables such as temperature and humidity, combined with data augmentation techniques. First, temporal and lagging features of the electric field are extracted and fused with multidimensional meteorological data including temperature, humidity, wind speed, and total cloud cover. A spatial-temporal density-based spatial clustering of applications with noise (ST-DBSCAN) is employed to annotate samples across multiple stations. The mode-normalized Wasserstein generative adversarial network with gradient penalty (MN-WGAN-GP) is used to generate synthetic samples with distributions similar to real data. Finally, a lightning warning algorithm is constructed based on categorical boosting (CatBoost). Experimental results show that the model achieves a probability of detection (POD) of 82.89% and a FAR of 27.33% on the test set. The proposed algorithm contributes to the development of refined regional lightning warning technologies and ensures the safety of low-altitude operations.

Keywords: low-altitude safety; atmospheric electric field; automatic weather station; lightning warning; catboost

Citation: Tiantian Yu, Haitao Wang, Wei Xu, Yan Liu."Lightning Warning Algorithm Based on Multi-Station Atmospheric Electric Field and Data Augmentation." Instrumentation 12, no.4 (December 2025). https://doi.org/10.15878/j.instr.202500299

1 Introduction

Lightning is an atmospheric discharge phenomenon that occurs when charged clouds approach the ground or interact with oppositely charged clouds. It is characterized by intense electromagnetic interference and instantaneous high-energy discharge, which can result in the failure of navigation systems, communication disruptions, and damage to critical onboard equipment in low-altitude aircraft. Meanwhile, the sudden and unpredictable nature of lightning poses significant safety hazards and operational risks to low-altitude economic activities, particularly for low-altitude logistics and UAV operations.

Effective lightning warning systems can significantly reduce economic losses. The atmospheric electric field, generated by charge distribution differences between the Earth's surface and the atmosphere, serves as a direct signal source of lightning activity. Its variation is closely associated with charge structures and discharge processes within thunderclouds[1]. Traditional lightning warning methods rely on data from weather radars, lightning locating systems, and atmospheric electric field mills. However, weather radars are limited by low temporal resolution and poor real-time warning capability[2]. In contrast, atmospheric electric field mills can monitor the changes in the ground electric field in real time and are suitable for detecting close-range thunderstorm activities. They have higher real-time performance and spatial resolution[3, 4]. In small-scale regional lightning warning, atmospheric electric field data have good application prospects.

Early lightning warning methods based on atmospheric electric field data primarily relied on statistical analyses, typically involving the use of threshold values to estimate the likelihood of lightning occurrence. In 2008, Murphy et al. conducted thunderstorm warning tests using electric field thresholds of 1 kV/m and 2 kV/m, combined with real-time electric field data[5]. In 2009, Aranguren compared the effectiveness of the polarity reversal method and the amplitude threshold method in lightning warning systems, concluding that the former performed better, achieving a probability of detection (POD) of 47% and a false alarm rate (FAR) of 78%[6]. In 2013, Zeng et al. proposed a short-term warning method for cloud-to-ground (CG) lightning based on the amplitude and variation rate of the atmospheric electric field, as well as radar reflectivity factors; however, POD was found to be highly sensitive to the selected threshold values[7]. Although these approaches are easy to implement, they fail to fully utilize the physical characteristics of electric field signals, leading to limited warning performance.

With advances in signal processing, researchers began employing non-stationary signal analysis techniques to extract features from atmospheric electric field data during thunderstorms. In 2014, Kang introduced an energy analysis method based on the short-time fourier transform (STFT) for lightning warning applications[8]. In 2019, Ju Zeli et al. compared the amplitude characteristics of atmospheric electric fields during thunderstorm and non-thunderstorm conditions. They observed that electric field amplitudes increased significantly under thunderstorm conditions, with a larger proportion of energy concentrated in the high-frequency bands. Using the Hilbert-Huang transform, intrinsic mode functions (IMFs) from two weather types were extracted and combined with lightning location data to construct a predictive warning model, achieving a POD of 78% and a threat score (TS) of 61.5%[9]. However, the method still exhibited a high missed alarm rate (MAR) and FAR, both exceeding 20%. To address the challenge of manually selecting threshold parameters, machine learning methods have been progressively introduced into lightning warning research[10]. In 2020, Xu Wei et al. proposed a lightning warning approach based on ensemble empirical mode decomposition (EEMD) and extreme gradient boosting (XGBoost). Compared to the general voting decision-making method, the approach improved the POD by up to 4.8% and reduced the FAR by 5.2% to 6.4%[11]. In 2023, Li et al. employed enhanced empirical wavelet transform and adaptive filtering to decompose atmospheric electric field signals, and used one-dimensional morphological analysis to extract time-frequency domain features. The resulting model achieved a POD of 77.11% with a 22-minute lead time, but suffered from a high FAR of 40.19% and a critical success index (CSI) of only 0.51[12]. These results indicate that the spatiotemporal features of the atmospheric electric field were still not fully exploited, and the generalizability of the method remains limited.

In summary, current lightning warning methods based on atmospheric electric field data primarily focus on the temporal sequence characteristics of single-station atmospheric electric field data, with insufficient exploration of spatiotemporal features. Moreover, the scarcity of lightning location data within the target region leads to limited training samples for warning models, making it difficult to accurately capture the spatiotemporal patterns of lightning events. The interplay of these dual factors significantly compromises the generalization capability and warning performance of existing prediction models. By incorporating multi-station atmospheric electric field data, it becomes possible to better capture spatial variations in electric field characteristics across the region of interest, thereby enhancing the spatiotemporal resolution of lightning warnings. To ameliorate the problem of insufficient data samples, the mode-normalized Wasserstein generative adversarial network with gradient penalty (MN-WGAN-GP) is utilized to model both numerical and categorical features, generating synthetic samples with distributions similar to real observations. Furthermore, the categorical boosting (CatBoost) algorithm is employed to extract complex spatiotemporal features from integrated observations obtained from atmospheric electric field mills and automatic weather stations (AWS), thereby improving both the predictive accuracy and generalization ability of the warning model. Finally, the effectiveness of the proposed method is validated using actual lightning occurrence times and locations recorded by a lightning locating system.

2 Dataset and Methods

2.1 Introduction of the Dataset

The primary data types used for regional lightning warning included atmospheric electric field data, lightning locating system records, and automatic weather station observations. From March 2022 to September 2023, five atmospheric electric field mills were deployed for continuous ground-level electric field measurements within the target area, with a temporal resolution of 1 Hz. The electric field mill network configuration featured maximum and minimum inter-station distances of 85.22 km and 1.58 km, respectively, while individual mills' AWS demonstrated an approximate detection radius of 20 km.

The lightning locating system recorded real-time data on the geographic coordinates, peak current amplitude, return-stroke height, and location error of lightning events. In parallel, AWS collected meteorological parameters including temperature, relative humidity, wind speed, and total cloud cover. The spatial distribution of the atmospheric electric field mills and AWS stations in the study region is illustrated in Fig. 1.

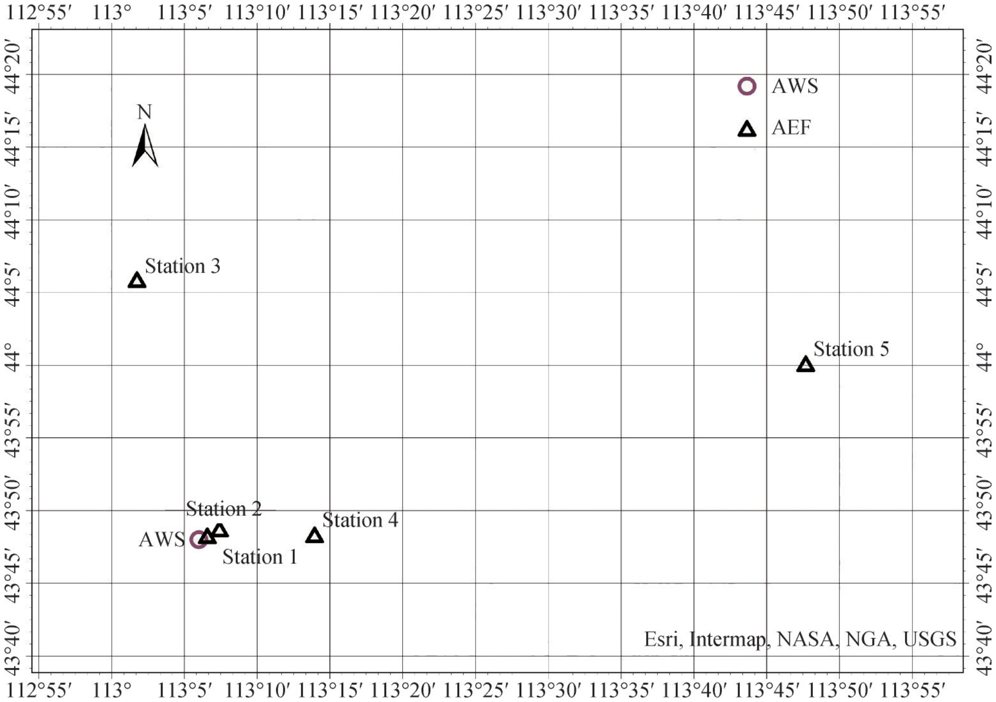


Fig.1 Geospatial distribution of atmospheric electric field mills and AWS within the study area

2.2 Data Preprocessing

The constructed feature space integrates the following components: ① fundamental atmospheric electric field features, including real-time amplitude, rate of change, and rolling window statistics; ② basic meteorological features, such as temperature, relative humidity, wind speed, and total cloud cover; ③ spatiotemporal features, encompassing lagged values and long-term trends of the electric field over mutiole time windows, periodic temproal indicators, and station location information.

As a core variable, the atmospheric electric field amplitude reflects the intensity and variations of the electric field in the atmospheric environment. For each station, basic features-including the mean, maximum, and minimum electric field values-were extracted using a 30-minute time window. In addition, the absolute and relative rates of change between adjacent windows were calculated to capture short-term fluctuations. To characterize future trends, rolling means and rolling standard deviations were computed for the next M, N, and O time windows using a moving window approach. To capture historical variation patterns, statistical features computed from the past L windows were incorporated as lagged sequence features. Observations from multiple stations were integrated to model spatial correlations across the region. Temporal features such as hour of the day, day of the week, month, and season were also included, while station identifiers were used to distinguish among different observation locations. These collectively formed a multidimensional spatiotemporal representation of the electric field data. Meteorological variables—temperature, relative humidity, wind speed, and total cloud cover—obtained from automatic weather stations were temporally and spatially aligned with the electric field features based on timestamps and geographic coordinates.

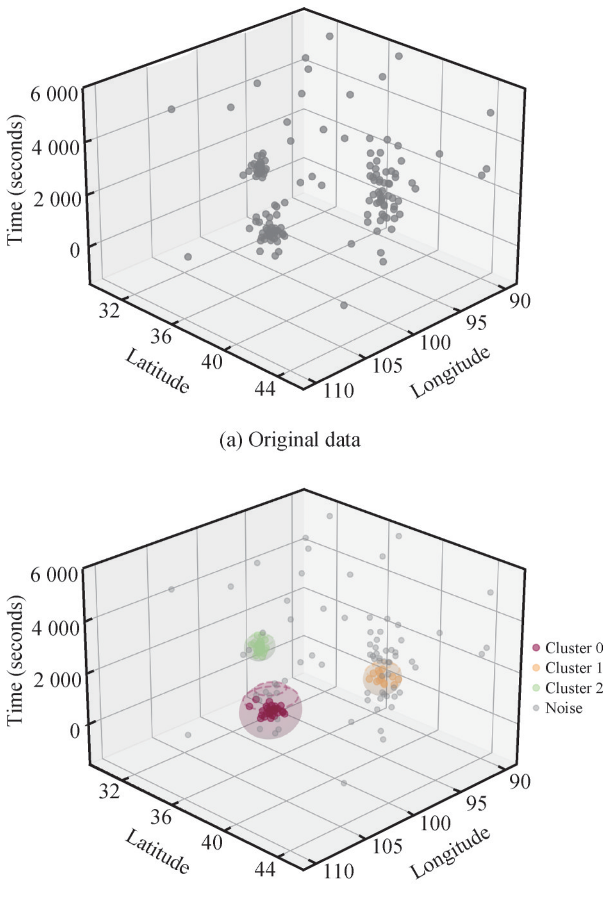
2.3 Label Production

Lightning events typically exhibit spatiotemporal clustering characteristics. Isolated lightning strikes are insufficient to represent the overall intensity and spatial extent of lightning activity. By applying spatiotemporal clustering, lightning events with proximate locations and similar occurrence times can be grouped into the same category, enabling accurate multi-station sample labeling[13]. Spatial-temporal density-based spatial clustering of applications with noise (ST-DBSCAN), an extension of DBSCAN incorporating both spatial and temporal dimensions, clusters data points based on spatial proximity while simultaneously grouping events by their occurrence time[14]. ST-DBSCAN can effectively identify those lightning event clusters that have both spatial clustering and temporal concentration, and is used to label the positive and negative samples of lightning events.

The objective of ST-DBSCAN is to consider both spatial and temporal density to identify spatiotemporal clusters. Its objective function is defined as follows:

$$J=\sum_{i=1}^{K}\ \sum_{x\in C_i}\left(\left\| x_s-\mu_{s_i}\right\|^2+\lambda\left\| x_t-\mu_{t_i}\right\|^2\right) \tag{1}$$

In this context, K denotes the number of clusters. C_i represents the i-th cluster. x_s and x_t correspond to the spatial and temporal features of data point, respectively. μ_{s_i} and μ_{t_i} indicate the spatial and temporal centroids of cluster C_i. λ is weighting coefficients balancing spatial-temporal importance. The spatiotemporal clustering results are illustrated in Fig. 2.



(a) Original data

(b) Clustering results

Fig.2 Spatio-temporal clustering effect: (a) Original data, (b) Clustering results

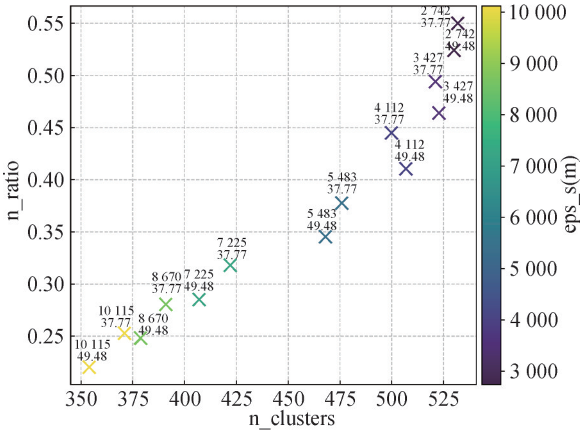


Fig.3 Parameter sensitivity analysis diagram

To establish the spatial radius *eps_s* and temporal radius *eps_t* based on the climatological statistics of storm size and duration in the study region, we extracted the equivalent radius and duration characteristics of storm events from historical lightning location data in the area. Based on these statistical measures, we generated candidate parameter values and validated the rationality of parameter selection through sensitivity analysis. The parameter sensitivity analysis is illustrated in Fig. 3, where the optimal parameters were determined by evaluating the number of clusters and noise ratio.

ST-DBSCAN clustering results are utilized to label each electric field station sample as either a positive or negative instance. The specific procedure is as follows:

Step 1: Set the spatial radius and temporal radius, then apply ST-DBSCAN to cluster lightning events by integrating longitude-latitude coordinates and timestamps, thereby identifying spatiotemporally correlated lightning clusters.

Step 2: Associate each electric field station sample with a specified future time window. Filter lightning clusters that occur within this time window.

Step 3: Specify the latitude and longitude of each electric field station, and compute the spatial distance between each station and the filtered lightning clusters using the Haversine formula. This step determines whether a lightning event occurred within the station's effective spatial influence range.

Step 4: Samples are labeled as positive if a lightning cluster is detected within the station's predefined spatiotemporal range (time window + spatial radius), otherwise as negative.

2.4 Data Augmentation Network

Due to the sparsity and sudden onset of lightning events, the data exhibits significant class imbalance and complex multi-modal distributed characteristics. The Wasserstein generative adversarial network with gradient penalty (WGAN-GP) improves training stability by introducing a gradient penalty term, but struggles to simultaneously preserve the physical plausibility and distributional diversity of minority class samples. To address this issue, we propose an improved generative adversarial network—MN-WGAN-GP—which integrates mode normalization into the WGAN-GP framework. The distribution modeling of numerical features is carried out using the variational Gaussian mixture model (VGMM), and the categorical features are represented via one-hot coding to better generate data with non-Gaussian and multimodal distribution features.

WGAN-GP is an improved version of the traditional generative adversarial network (GAN), designed to address issues of unstable training and poor generation quality. By replacing the Jensen-Shannon (JS) divergence with Wasserstein distance in GAN to measure the distance between the generator and discriminator, issues like training instability and mode collapse can be effectively mitigated[15]. Additionally, WGAN-GP introduces a gradient penalty term L_{gp} into the discriminator's loss function, replacing the original weight clipping strategy. By enforcing constraints on the gradient norm of the discriminator, the gradient penalty mitigates problems such as gradient explosion and vanishing gradients, thereby improving training stability[16]. The objective function of WGAN-GP is defined as follows:

$$min_G max_D V(D,G) = E_{z\sim P_Z(z)}\left[D\left(G(z)\right)\right] - E_{x\sim P_X(x)}\left[\mathrm{D}(\mathrm{x})\right] + L_{gp} \tag{2}$$

$$L_{gp} = \lambda E_{\bar{x}\sim P_{\hat{X}}(\bar{x})}\left[\left(\left\|\nabla_{\bar{x}} D(\bar{x})\right\|_2 - 1\right)^2\right] \tag{3}$$

In the objective function, $\|\cdot\|_2$ denotes the L_2 norm, ∇ represents the gradient operator, and λ is the penalty coefficient, typically set to 10. $P_{\bar{X}}(\cdot)$ denotes linear uniform sampling between data points from the real distribution $P_X(\cdot)$ and the generated distribution $P_{G(z)}(\cdot)$. $\bar{x}$ indicates random interpolation between target and generated data:

$$\bar{x} = \xi x + (1-\xi)G(z) \tag{4}$$

In the equation, ξ is drawn from a uniform distribution over the interval [0, 1]. The model architecture of WGAN-GP is illustrated in Fig. 4.

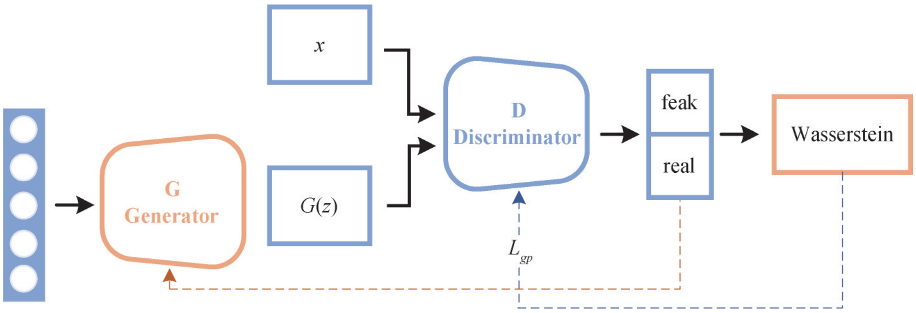


Fig.4 Model structure of WGAN-GP

As shown in Fig. 5, the overall architecture of the MN-WGAN-GP model consists of three main components: a mode normalization module, a generator (G), and a discriminator (D).

The model is designed to address the imbalance in lightning data by incorporating physical constraints into the generative adversarial network framework.

The mode normalization module encodes and decodes input features under physical constraints. During the encoding phase , it maps real multi-modal distributed data X into a unified latent space, enforcing physical plausibility constraints on generator outputs. In the decoding phase , it reconstructs the encoded components into the original feature space to generate final samples. For numerical features X_1, such as electric field amplitude, VGMM is employed for probability density estimation. Each variable is decomposed into three components: the cluster probability x_1, the intra-cluster normalized value x_2, and the cluster indicator x_3. Lightning-related electric field and meteorological variables typically exhibit multi-modal distributions. Directly modeling such features with unimodal distributions often leads to unrealistic synthetic samples. VGMM enables more accurate modeling of each continuous variable, capturing the complex distributional patterns inherent in lightning-related data. For categorical features X_2, such as total cloud cover, one-hot encoding is applied to obtain a binary vector x_4, ensuring category independence in the generative process.

The generator G is implemented as a deep fully connected neural network. It takes a 128-dimensional Gaussian noise vector z as input and transforms it into high-dimensional, complex feature representations through three fully connected layers containing 256, 512, and 1024 neurons, respectively. Each layer is followed by batch normalization and a LeakyReLU activation function ($\alpha = 0.2$). At the output stage, the generator produces multiple branches, structurally constrained by the encoded space of the original data X. Each output feature is regularized by an appropriate activation function to match the target feature type, as summarized in Table 1.

The discriminator D receives mode-normalized encoded data (x and $G(z)$) from both real and generated

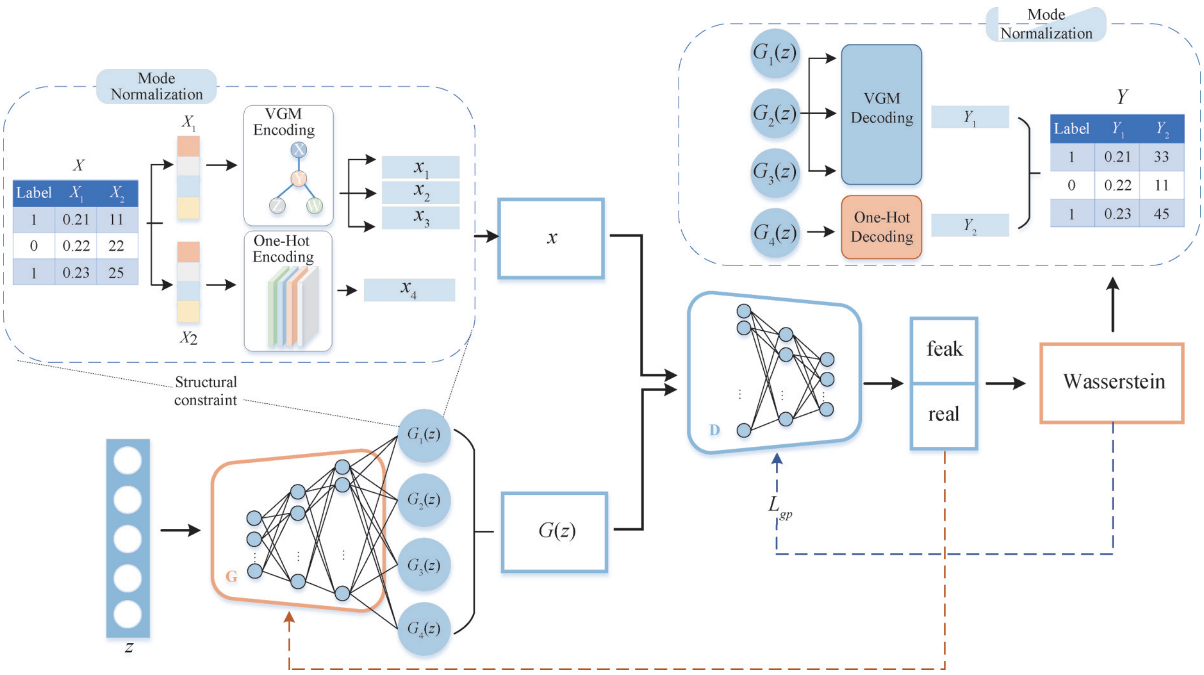


Fig.5 Model structure of MN-WGAN-GP

Table 1 Generator output branch

Output Type	Activation Function	Output Vector
cluster probability	Softmax	$G_1(z)$
intra-cluster normalized value	Tanh	$G_2(z)$
cluster indicator	Sigmoid	$G_3(z)$
one-hot encoding vectors	Softmax	$G_4(z)$

samples, extracting discriminative features through three fully connected layers with 1024, 512, and 256 neurons, respectively. Each layer is followed by a LeakyReLU activation function ($\alpha = 0.2$) and a dropout layer (dropout rate = 0.2) to enhance model robustness. The final output is a linear layer that produces a single scalar value used to compute the Wasserstein distance. During training, the discriminator must not only distinguish between real and generated samples but also satisfy the Lipschitz continuity constraint, which is enforced through the gradient penalty mechanism. After training, the generator's output is post-processed to reconstruct the final synthetic sample *Y*. Specifically, the encoded components $G_1(z)$, $G_2(z)$, and $G_3(z)$ are inverted into the original numerical features Y_1 using the learned parameters of the VGMM. Similarly, the one-hot encoding vector is mapped back to its corresponding categorical feature Y_2. These two parts are then combined to form the final generated sample *Y*.

2.5 Lightning Warning Algorithm Model

2.5.1 Lightning warning model based on the CatBoost algorithm

The core objective of the lightning warning task is to predict the probability of lightning occurrence in a future time window based on input features. This constitutes a typical imbalanced binary classification problem. CatBoost, an efficient gradient boosting decision tree framework, is specifically designed to handle categorical features and improve both model speed and accuracy. Traditional gradient boosting algorithms require preprocessing of categorical features—such as one-hot encoding or target encoding—which may lead to data leakage in high-dimensional scenarios. CatBoost incorporates a unique target encoding technique that prevents information leakage while efficiently processing categorical features for lightning warning classification (such as station IDs or total cloud cover). The framework further introduces ordered boosting, which trains on sequentially ordered data subsets to enhance the model's ability to capture temporal dependencies in lightning warning systems, thereby mitigating target leakage in gradient boosting.

The CatBoost classification process for lightning warning is as follows:

Let the lightning warning dataset consist of N samples, with the feature space $X \in \mathbb{R}^d$ containing both numerical and categorical features, and the label space $Y = \{0, 1\}$ indicating the presence or absence of lightning events. A weighted cross-entropy loss function is employed, assigning higher weights to positive samples:

$$L(y,F) = -\sum_{i=1}^{N}\left[\begin{array}{l} w_1 y_i \log\sigma\left(F(x_i)\right) + \\ w_0\left(1-y_i\right)\log\left(1-\sigma\left(F(x_i)\right)\right) \end{array}\right] + \Omega(F) \tag{5}$$

In the loss function, $F(x)$ denotes the output of the ensemble model, $\sigma(z) = \dfrac{1}{1+e^{-z}}$ is the sigmoid activation function, w_1/w_0 is the class weight used to mitigate sample imbalance by assigning greater importance to positive samples. $\Omega(F)=\gamma T+\frac{1}{2}\lambda w^2$ is the regularization term, where T represents the number of leaf nodes in the tree and w denotes the weight of each leaf.

The lightning warning model is built upon a gradient boosting framework, which iteratively optimizes a weighted cross-entropy loss function:

$$F_t(x)=F_t(x-1)+\alpha h_t(x) \tag{6}$$

In this context, $h_t(x)$ represents the newly added decision tree at the t-th iteration, and α denotes the learning rate. In each iteration, the tree h_t is optimized by minimizing a second-order approximation of the loss function:

$$h_t^*=\text{argmin}\sum_{i=1}^{N}\left[g_i h(x_i)+\frac{1}{2}h_i h^2(x_i)\right]+\Omega(h) \tag{7}$$

In this context, $g_i=\partial_{F_{t-1}}L(y_i,F_{t-1})$ and $h_i=\partial^2_{F_{t-1}}L(y_i,F_{t-1})$ denote the first-order and second-order gradients, respectively.

CatBoost adopts a symmetric tree structure where each node follows identical splitting rules. To construct a decision tree $h_t(x)$, the split criterion at each node is determined based on a selected feature j and a threshold τ. The split gain is computed as follows:

$$\text{Gain}=\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}g_i\right)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{\left(\sum_{i\in I_R}g_i\right)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{\left(\sum_{i\in I}g_i\right)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{8}$$

In the above, $I_L=\{i|x_{ij}\le\tau\}$, $I_R=I\backslash I_L$, λ is the split complexity penalty.

For categorical features c, CatBoost allows direct specification through the cat_features parameter. CatBoost automatically applies ordered target encoding:

$$Enc(c)_i=\frac{\sum_{k=1}^{i-1}\mathbb{I}_{x_k=c}y_k+ap}{\sum_{k=1}^{i-1}\mathbb{I}_{x_k=c}+ap} \tag{9}$$

Here, a is the smoothing parameter, and p denotes the prior probability. This encoding method prevents target leakage by introducing random permutations and smoothing during the encoding process.

The hyperparameter settings of the CatBoost model have a direct impact on its classification performance. Among them, the three most influential parameters are the number of iterations, tree depth, and learning rate. To optimize model performance, it is necessary to perform hyperparameter tuning to determine the most appropriate settings for model initialization.

2.5.2 Evaluation metrics

To evaluate the performance of the lightning warning model, four commonly used meteorological evaluation metrics are employed: probability of detection (POD), false alarm rate (FAR), threat score (TS), and equitable threat score (ETS). These metrics are widely adopted in meteorological forecasting and provide a comprehensive assessment of model accuracy and reliability.

In binary classification tasks, the relationship between the true class and the predicted class can be categorized into four types: TP (true positive), FP (false positive), TN (true negative), and FN (false negative). Based on these classifications, the evaluation metrics are defined as follows[17]:

$$POD=\frac{TP}{TP+FN} \tag{10}$$

$$FAR=\frac{FP}{TP+FP} \tag{11}$$

$$TS=\frac{TP}{TP+FN+FP} \tag{12}$$

$$H=\frac{(TP+FN)(TP+FP)}{TP+TN+FN+FP} \tag{13}$$

$$ETS=\frac{TP-H}{TP+FN+FP-H} \tag{14}$$

These evaluation metrics provide an understanding of the model's performance. In particular, when dealing with imbalanced data, relying solely on accuracy may fail to adequately reflect the model's true predictive capabilities.

3 Experiments and Analysis

3.1 Experimental Data

The constructed dataset comprises multi-site atmospheric electric field data, AWS data, and lightning location data. The sample time duration is 30 minutes. Using the lightning location data, samples with lightning events occurring within the subsequent 40 minutes are labeled as positive, while those without lightning events are labeled as negative. The dataset comprises 147 features and 66,256 samples in total . To ensure that the data in the training set is not leaked to the test set and validation set, a time block hierarchical method was adopted. Training was conducted in 2022, validation in early 2023, and testing in the middle of 2023.

3.2 Data Augmentation Processing

MN-WGAN-GP was employed to augment the number of positive samples, thereby balancing the class distribution within the dataset. The model demonstrates stability in handling multimodal distributions. The training set's positive samples were doubled through this augmentation. The key hyperparameters of the model are listed in Table 2. The original dataset contained 1, 758positive samples, which increased to 3, 516 after augmentation.

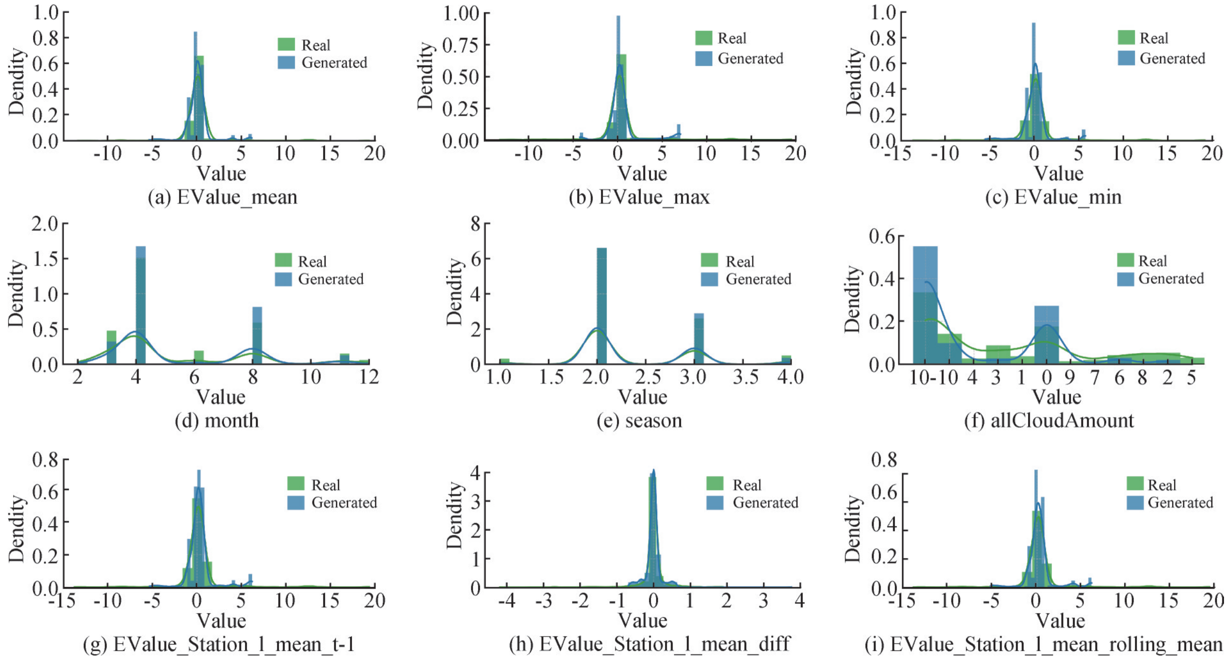
To verify whether the generated data effectively simulates the feature distribution of the original data, nine representative numerical and categorical features were selected for histogram-based comparison. As shown in

Table 2 Key hyperparameter settings

Hyperparameter	Settings
batch_size	64
epochs	2000
gradient penalty weight	20
learning_rate	5e-4
momentum parameter β_1	0.6
momentum parameter β_2	0.8

Fig. 6, green represents the real data, while blue denotes the generated data. For most numerical features, the generated data exhibits similar distribution trends and densities to the real data. For categorical features, the generated samples successfully capture modes observed in the original dataset, demonstrating good diversity. These results indicate that the generated data performs well in feature simulation and can effectively approximate the distribution of the original data.

To evaluate the improved performance of the MN-WGAN-GP model, the quality of generated samples was compared with that of the original WGAN-GP model using the same dataset. The WGAN-GP model was used to perform augmentation of positive samples on the original dataset. The differential evolution (DE) algorithm was applied to optimize key hyperparameters of the model, including the number of iterations, batch size, and learning rate. These settings aimed to compare the performance of the two models in terms of data generation quality and similarity.



(a) EValue_mean (b) EValue_max (c) EValue_min

(d) month (e) season (f) allCloudAmount

(g) EValue_Station_1_mean_t-1 (h) EValue_Station_1_mean_diff (i) EValue_Station_1_mean_rolling_mean

Fig.6 Histograms of real and generated data

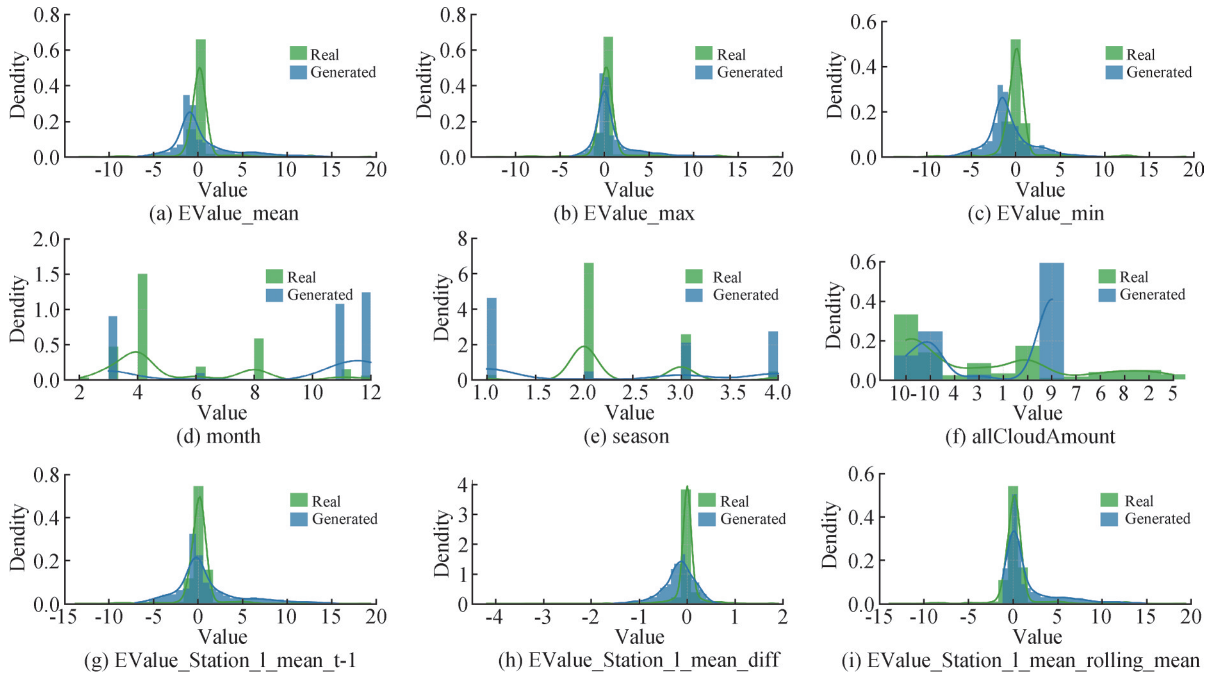
After data generation, histograms of key features from both the original and generated data by the original WGAN-GP model were plotted, as shown in Fig. 7, illustrating how well the WGAN-GP generated data simulates the original data.

As shown in Fig. 7, WGAN-GP is generally able to approximate the distributions of most real features. However, for certain features such as EValue_mean, the distribution of the generated data slightly deviates from the real data. Regarding discrete features, the diversity of data generated by WGAN-GP is limited—for example, features like month and allCloudAmount fail to capture all modes present in the real data. While WGAN-GP is capable of generating continuous-valued features, the density estimates for some feature intervals are less accurate compared to those produced by MN-WGAN-GP, as shown in Fig. 6. This is particularly evident for features such as EValue_min and EValue_Station_1_mean_diff, where the generated data underrepresents certain distributional characteristics observed in the real samples.

From the histogram comparison analysis, MN-WGAN-GP demonstrates superior performance in simulating the real data distribution compared to the original WGAN-GP. The data generated by MN-WGAN-GP closely approximates the distribution of real data across multiple features and exhibits good diversity and continuity. While WGAN-GP generally matches the real data distribution, its performance in capturing diversity in categorical features and density estimation for numerical features is inferior to MN-WGAN-GP.

3.3 Experimental Results Analysis

Using the dataset constructed in Section 2.2, a CatBoost model was built and evaluated. The CatBoost algorithm contains several key hyperparameters that



(a) EValue_mean
(b) EValue_max
(c) EValue_min
(d) month
(e) season
(f) allCloudAmount
(g) EValue_Station_1_mean_t-1
(h) EValue_Station_1_mean_diff
(i) EValue_Station_1_mean_rolling_mean

Fig.7 WGAN-GP generates sample distribution

significantly impact the model's predictive performance. To mitigate the risks of overfitting and underfitting, a grid search combined with 3-fold cross-validation was employed to identify the optimal hyperparameter configuration. This approach systematically explores all parameter combinations within the grid, ensuring comprehensive coverage and selecting the combination that yields the best validation performance. The hyperparameter search space and the optimal hyperparameters are listed in Table 3. After determining the best settings, the lightning early warning model was retrained on the training set using the selected hyperparameters.

Table 3 Hyperparameter search space and optimal hyperparameters

Hyperparameter	Hyperparameter search space	Optimal hyperparameters
learning_rate	0.001,0.01,0.05,0.1	0.1
iterations	500,1000,2000	2000
depth	6,7,8,9,10,11,12	9
subsample	0.5,0.6,0.7,0.8,0.9,1.0	0.9
rsm	0.5,0.6,0.7,0.8,0.9,1.0	0.8
l2_leaf_reg	1,3,5,7,10	10

To evaluate the performance of the lightning warning, the CatBoost model's metrics on the test set were calculated as follows: POD = 0.8289, FAR = 0.2733, TS = 0.6319, and ETS = 0.6187. Additionally, the confusion matrix for the test set was computed, showing the comparison between the true labels and predicted labels, as illustrated in Fig. 8.

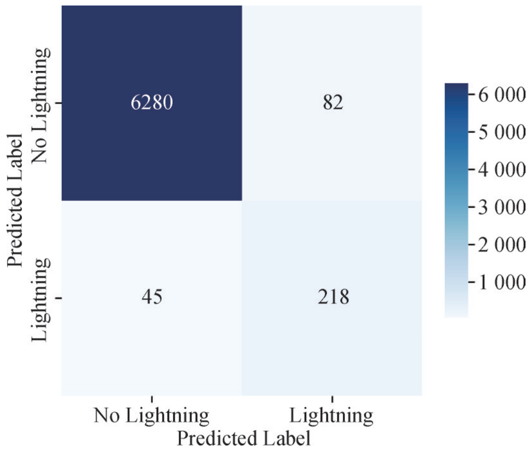


Fig.8 Confusion matrix for the test set

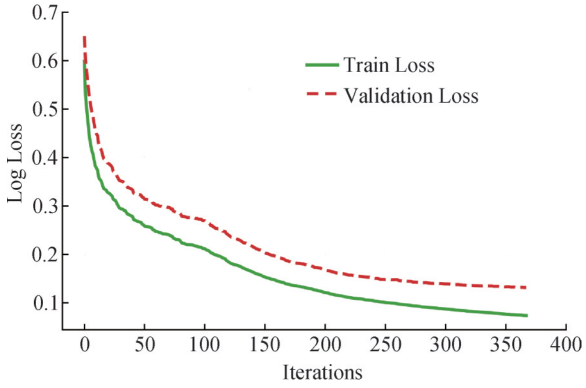


Fig.9 Loss curves for training and validation sets

Fig. 9 illustrates the trends of training and validation loss over iterations during the CatBoost model training process. The green solid line represents the training loss, while the red dashed line represents the validation loss. As shown in the figure, both loss curves exhibit a gradual downward trend as the number of iterations increases, indicating that the model is continuously learning and optimizing. The training loss remains consistently lower than the validation loss, but the gap between them is relatively small, suggesting no obvious overfitting. The CatBoost ensemble learning algorithm demonstrates strong performance in lightning early warning.

3.4 Ablation Experiments

To validate the impact of MN-WGAN-GP generated meteorological data on downstream classification tasks, we conducted ablation experiments based on meteorological evaluation metrics, including POD, FAR, TS, and ETS. The experiments compared three data strategies under identical data splits: no augmentation, WGAN-GP augmentation, and MN-WGAN-GP augmentation. Table 4 shows the comparison of ablation experiment indicators.

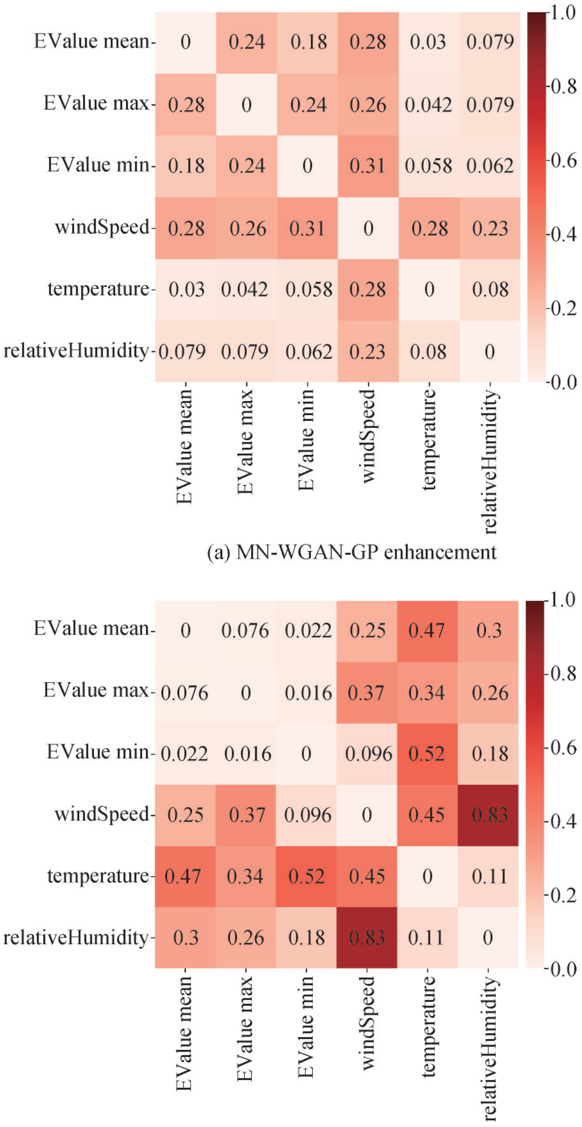
Table 4 Comparison of ablation experiment indicators

Model	POD	FAR	TS	ETS
no enhancement	0.7833	0.2399	0.6280	0.6154
WGAN-GP	0.8023	0.2518	0.6317	0.6190
MN-WGAN-GP	0.8289	0.2733	0.6319	0.6187

The results demonstrate that the MN-WGAN-GP augmentation strategy outperforms the other two approaches, as evidenced by improved POD, TS, and ETS scores, indicating that this method not only enhances detection rates but also strengthens comprehensive forecasting capabilities. In contrast, while traditional WGAN-GP augmentation shows marginal improvements in certain metrics, its exclusive optimization of univariate distributions may lead to the distortion of inter-variable relationships in meteorological data.

For the data correlation analysis, we further investigated the consistency of multivariate relationships between generated data and real data. By computing inter-variable correlation matrices for both real and generated data, we visualized their differences through heatmaps. Fig. 10 presents the difference heatmaps for MN-WGAN-GP augmentation and WGAN-GP augmentation, respectively, providing an intuitive comparison of their performance in preserving variable correlation patterns. Quantitative analysis revealed that for MN-WGAN-GP augmentation, the two correlation matrices exhibited an MAE of 0.1472, an MSE of 0.0291, and a high Pearson correlation coefficient of 0.9541. In contrast, WGAN-GP augmentation showed inferior metrics with MAE= 0.2807, MSE=0.1267, and Pearson correlation=0.6922. These results demonstrate MN-WGAN-GP's superior capability in preserving joint relationships among variables.

The heatmap analysis further showed that the difference matrix of MN-WGAN-GP primarily contained values below 0.3, indicating strong agreement in correlation patterns between generated and real data, with notable discrepancies only occurring in some weakly correlated variable pairs. However, WGAN-GP's difference matrix reached values as high as 0.83, exhibiting significant deviations even in strongly correlated variable pairs. These findings collectively demonstrate that MN-WGAN-GP augmentation not only shows advantages in downstream prediction tasks but also excels in data fidelity.



(a) MN-WGAN-GP enhancement

(b) WGAN-GP enhancement

Fig.10 Differential heat map: (a) MN-WGAN-GP enhancement, (b) WGAN-GP enhancement

3.5 Comparative Experiments

To visually demonstrate the advantages of the CatBoost ensemble learning method in lightning early warning, a comparative analysis was conducted on the predictive performance of three representative ensemble learning models: XGBoost, LightGBM, and Random Forest. All models were trained and evaluated under the same experimental conditions using identical training, validation, and test sets. The training set was augmented using the same data augmentation procedure to ensure consistency.

The predictive performance and prediction time of each model on the test set were computed, with results summarized in Table 5. The comparison of key evaluation metrics is further visualized in Fig. 11. Among the three baseline models, LightGBM achieved the best prediction performance, with POD, FAR, TS, and ETS values of 0.6982, 0.2289, 0.5784, and 0.5675, respectively. XGBoost performed second best, with corresponding values of 0.6532, 0.2602, 0.5311, and 0.5196. The Random Forest model performed the worst, with values of 0.0382, 0.0095, 0.0385, and 0.0341, respectively. In contrast, the CatBoost model has shown significant improvements on both TS and ETS. These results clearly demonstrate the superior performance of CatBoost in the lightning early warning task. Its optimized gradient boosting strategy and capability in handling categorical features make it a more advantageous algorithm choice in this field.

Table 5 Evaluation metrics and prediction times for different models on the test set

Model	POD	FAR	TS	ETS	Prediction time /s
CatBoost	0.8289	0.2733	0.6319	0.6187	0.0269
LightGBM	0.6982	0.2289	0.5784	0.5675	0.0507
XGBoost	0.6532	0.2602	0.5311	0.5196	0.0481
Random Forest	0.0382	0.0095	0.0385	0.0341	0.1342

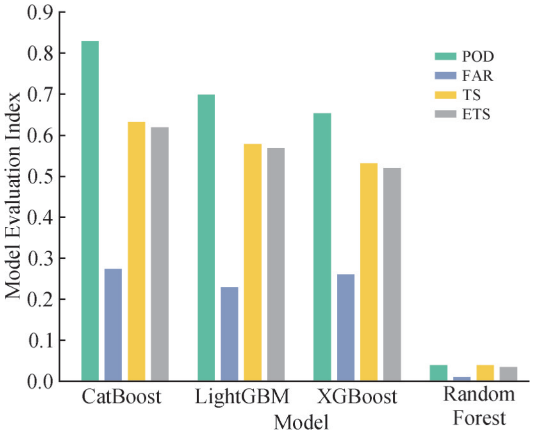


Fig.11 Comparison of model evaluation metrics

In terms of computational efficiency, CatBoost requires only 0.0269 seconds per prediction on the test set, outperforming LightGBM, XGBoost, and random forest. Its prediction speed is approximately 44.07%, 46.94%, and 79.96% faster than these three models, respectively. These results indicate that CatBoost not only excels in prediction accuracy but also offers high prediction speed, making it an ideal choice for real-time lightning early warning systems.

4 Conclusion

To address the challenges of insufficient spatiotemporal feature extraction from single-station atmospheric electric field data and the scarcity of lightning locating samples in lightning early warning, a data augmentation method based on MN-WGAN-GP was proposed. Combined with multi-station atmospheric electric field data and meteorological variables, a CatBoost-based lightning early warning model was constructed. Experimental results led to the following main conclusions:

The MN-WGAN-GP model generates synthetic samples consistent with the distribution of real data, and through temporal sequence feature construction and spatiotemporal density-based clustering for sample labeling, it enhances the spatiotemporal correlations within the dataset, providing richer and more representative training samples for subsequent modeling. Compared to traditional lightning warning methods based on single-station atmospheric electric field data, integrating multi-station electric field data and meteorological factors, along with the construction of, enables the model to capture the spatiotemporal evolution patterns of lightning activity more comprehensively, significantly improving warning accuracy. Compared to ensemble learning methods such as LightGBM, XGBoost, and random forest, CatBoost achieved the best performance on key metrics, with POD, FAR, TS, and ETS of 0.8289, 0.2733, 0.6319, and 0.6187, respectively. Additionally, its single prediction time of 0.0269 seconds outperformed other models, demonstrating higher suitability for real-time early warning scenarios.

In summary, the proposed lightning early warning method based on MN-WGAN-GP data augmentation and CatBoost not only mitigates sample scarcity and insufficient spatiotemporal feature extraction but also shows advantages in prediction accuracy and computational efficiency. Future work may further explore the model's generalization ability across different geographic environments and meteorological conditions.

Author Contribution:

Tiantian Yu: Conceptualization; Writing - original draft preparation; Methodology. Haitao Wang: Data curation; Resources. Wei Xu: Writing - review and editing; Formal analysis; project administration. Yan Liu: Supervision.

Foundation Information:

This research was funded by the National Natural Science Foundation of China, grant number 41605121 and Foundation of Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education (No.BATLAB202402)

Data Availability:

The authors declare that the main data supporting the findings of this study are available within the paper and its Supplementary Information files.

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 24 June 2025; Accepted 18 November 2025; Published online 31 December 2025

References

[1] BAO R, HE Z, ZHANG Z. Application of lightning spatio-temporal localization method based on deep LSTM and interpolation [J]. *Measurement* , **2022**, 189.

[2] BAO R, ZHANG Y, MA B J, et al. An Artificial Neural Network for Lightning Prediction Based on Atmospheric Electric Field Observations [J]. *Remote Sensing* , **2022**, 14(17).

[3] MANSOURI E, MOSTAJABI A, TONG C, et al. Lightning Nowcasting Using Solely Lightning Data [J]. *Atmosphere* , **2023**, 14(12).

[4] WAN Z, FU L, PU Z, et al. Optimization of the lightning warning model for distribution network lines based on multiple meteorological factor thresholds [J]. *Frontiers in Energy Research* , **2023**, 11.

[5] MURPHY M J, HOLLE R L, DEMETRIADES N W S. Cloud-to-ground lightning warnings using electric field mill and lightning observations [Z]. *20th International Lightning Detection Conference, Tucson, AZ* , USA, 21-23 April **2008**.

[6] ARANGUREN D, MONTANYA J, SOLA G, et al. On the lightning hazard warning using electrostatic field: Analysis of summer thunderstorms in Spain [J]. *Journal of Electrostatics* , **2009**, 67(2-3): 507-12.

[7] ZENG Q, WANG Z, GUO F, et al. The application of lightning forecasting based on surface electrostatic field observations and radar data [J]. *Journal of Electrostatics* , **2013**, 71(1): 6-13.

[8] KANG H, LIU C, JIANG X. Weather Recognition Algorithm Based on the Characteristics of Atmospheric Electric Field Signal [J]. *Comput Simul* , **2014**, 31: 312-5.

[9] JU Z, V X, PU L, et al. Method of Lightning Nowcasting Warning Based on Atmospheric Electric Field Characteristics [J]. *Insulators and Surge Arresters* , **2019**, (04): 111-7.

[10] SRIVASTAVA A, MISHRA M, KUMAR M. Lightning alarm system using stochastic modelling [J]. *Natural Hazards* , **2015**, 75(1): 1-11.

[11] XU W, XIA Z, XING H. Lightning warning method based on EEMD and XGBoost [J]. *Chinese Journal of Scientific Instrument* , **2020**, 41(08): 235-43.

[12] LI X, YANG L, YIN Q, et al. Lightning Risk Warning Method Using Atmospheric Electric Field Based on EEWT-ASG and Morpho [J]. *Atmosphere* , **2023**, 14(6).

[13] AO Y, NI X, HUANG F, et al. Lightning duration and area from geostationary Lightning Mapping Imager based on a modified lightning cluster algorithm [J]. *Atmospheric Research* , **2024**, 303.

[14] NICOLIS O, DELGADO L, PERALTA B, et al. Space-time clustering of seismic events in Chile using ST-DBSCAN-EV algorithm [J]. *Environmental and Ecological Statistics* , **2024**, 31(2): 509-36.

[15] HOU W, GUO H, YAN B, et al. Tool wear state recognition under imbalanced data based on WGAN-GP and lightweight neural network ShuffleNet [J]. *Journal of Mechanical Science and Technology* , **2022**, 36(10): 4993-5009.

[16] YUDA E, ANDO T, KANEKO I, et al. Comprehensive Data Augmentation Approach Using WGAN-GP and UMAP for Enhancing Alzheimer's Disease Diagnosis [J]. *Electronics* , **2024**, 13(18).

[17] LU M, JIN C, YU M, et al. MCGLN: A multimodal ConvLSTM-GAN framework for lightning nowcasting utilizing multi-source spatiotemporal data [J]. *Atmospheric Research* , **2024**, 297.