atmosphere

MDPI

Article

ADASYN-LOF Algorithm for Imbalanced Tornado Samples

Zhipeng Qing [1,2], Qiangyu Zeng [1,2], Hao Wang [1,2,*], Yin Liu [3,4], Taisong Xiong [1,2] and Shihao Zhang [1,2]

1 CMA Key Laboratory of Atmospheric Sounding, Chengdu 610225, China; 2016021234@cuit.edu.cn (Z.Q.); zqy@cuit.edu.cn (Q.Z.); xts@cuit.edu.cn (T.X.); sxz@cuit.edu.cn (S.Z.)
2 College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China
3 Jiangsu Meteorological Observation Center, Nanjing 210041, China; liuyin200421@163.com
4 State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081, China
* Correspondence: wh@cuit.edu.cn

Abstract: Early warning and forecasting of tornadoes began to combine artificial intelligence (AI) and machine learning (ML) algorithms to improve identification efficiency in the past few years. Applying machine learning algorithms to detect tornadoes usually encounters class imbalance problems because tornadoes are rare events in weather processes. The ADASYN-LOF algorithm (ALA) was proposed to solve the imbalance problem of tornado sample sets based on radar data. The adaptive synthetic (ADASYN) sampling algorithm is used to solve the imbalance problem by increasing the number of minority class samples, combined with the local outlier factor (LOF) algorithm to denoise the synthetic samples. The performance of the ALA algorithm is tested by using the supporting vector machine (SVM), artificial neural network (ANN), and random forest (RF) models. The results show that the ALA algorithm can improve the performance and noise immunity of the models, significantly increase the tornado recognition rate, and have the potential to increase the early tornado warning time. ALA is more effective in preprocessing imbalanced data of SVM and ANN, compared with ADASYN, Synthetic Minority Oversampling Technique (SMOTE), SMOTE-LOF algorithms.

Keywords: tornadoes; class imbalance; machine learning

Citation: Qing, Z.; Zeng, Q.; Wang, H.; Liu, Y.; Xiong T.; Zhang, S. ADASYN-LOF Algorithm for Imbalanced Tornado Samples. *Atmosphere* **2022**, *13*, 544. https://doi.org/10.3390/atmos13040544

Received: 29 January 2022
Accepted: 25 March 2022
Published: 29 March 2022

1. Introduction

Tornadoes are small and medium-scale extreme weather events, usually generated at the bottom of thunderstorm clouds, with destructive power that can tear houses and trees and roll into the sky. Tornadoes occur less frequently in China than in the United States each year, and the majority of tornadoes occur from noon till evening in the summer months (June, July, and August) [1]. A tornado can be classified as EF0 to EF5 level according to the damage degree and wind speed [2,3]. With the upgrade of radar detection capabilities, tornado recognition algorithms went through the following process: tornadic vortex signature (TVS) criteria [4]-mesocyclone detection algorithm (MDA) [5,6]-tornado detection algorithm (TDA) [7]-tornadic debris signature (TDS) [8]. With the upgrading of computer technology in the past few years, artificial intelligence (AI) algorithms and classification models are gradually applied to tornado detection. For example, tornado detection algorithm based on neuro-fuzzy system and fuzzy logic [9,10], the S-band radar adaptive neuro-fuzzy tornado detection system [11], forecasting tornado with random forests [12], using a convolutional neural network (CNN) and image to predict tornadoes [13]. Artificial intelligence in the future tornado detection can reduce the tornado false alarm rate, increase the early warning time, and lower the experience restrictions on weather forecasters.

When applied to detect tornadoes, artificial intelligence algorithms usually suffer from the class imbalance problem. The class imbalance problem means the instances of one class are much more than the instances of another class [14], and the performance of classifiers leans to be partial towards the majority class in the imbalanced data set [15]. The imbalance might make it difficult to develop effective classifiers [16] in many applications such as sensor and detection [17,18]. Imbalanced models result in poor detection and high false

alarm rates based on AI methods in tornado recognition. There were many studies on the class imbalance problem of tornado samples. In 2014, Trafalis et al. used SVM, Logistic Regression, RF, and other models to evaluate the performance of different algorithms under the imbalanced tornado data set. However, they only compared the performance of different models without further research and solution to the problem of class imbalance [19]. When studying rare events, Maalouf et al. tested the imbalanced tornado sample set using a sample-weighted learning method [20–22]. This method essentially did not increase the number of tornado samples, so it may still face the class imbalance problem when applied to other models. In 2021, Basalyga et al. constructed tornado image samples using tornado data set and balanced the image training set using image enhancement technology [13]. Image translation and rotation have limited effects on balancing vector samples. An effective way to solve the class imbalance problem of tornado samples is to use radar to collect more tornado events to increase the number of minority samples. This method takes a considerable cost to monitor future tornadoes. Even worse, to retain classification information, minority and majority samples could be increased simultaneously, which will cause the class imbalance problem to still exist at the tornado sample set.

In order to solve the problem of class imbalance in the tornado sample set, sampling techniques need to be applied to balance the majority and minority. He et al. gave a comprehensive overview of imbalanced learning [23]. Haixiang et al. provided an in-depth review of rare event detection from the perspective of imbalanced learning [24], and Yu et al. studied the solutions to the imbalanced problem [25]. In the field of research on the imbalance of vector sample class, under-sampling, over-sampling, threshold shifting, Synthetic Minority Oversampling Technique (SMOTE, creates artificial data based on feature space similarities between existing minority samples), ADASYN, SMOTEBOOST, SMOTE-LOF, adaptive sampling, and other data balancing methods have been well researched [23,25–31]. The ADASYN used a weighted distribution for different minority class samples according to the samples' level of difficulty in learning. More synthetic data were generated for minority class examples that were harder to learn than those minority examples that were easier to learn [27]. A brief introduction of ADASYN algorithm shows in Section 3.1.

This study aims to solve the class imbalance problem of vector samples for tornadoes by using an adaptive synthetic (ADASYN) sampling approach to increase the number of minority samples. In addition, to verify the effectiveness of synthetic minority samples by the ADASYN approach, the local outlier factor (LOF) algorithm is carried out to identify and filter out noise data. In addition, the score performance of ADASYN-LOF, ADASYN, SMOTE-LOF, SMOTE algorithms is compared.

This research is organized as follows: Section 2 presents the weather radar used to detect tornadoes and how to build the tornado sample set using radar level-II data. Section 3 introduces the ADASYN, LOF, and machine learning algorithms, while Sectin 4 describes the experimental framework. Additionally, Section 5 contains the analysis of results and discussion. Lastly, the conclusions are presented in Section 6.

2. Data

2.1. Weather Radar

Fast-scanning and high-resolution weather radars, such as S, X, Ka-band, and phased array radars, are widely applied to detect and warn tornadoes [32–35]. The S-band China new generation of Doppler weather radar (CINRAD SA) plays an essential role in monitoring and forecasting tornadoes. The CINRAD SA's maximum distance resolution is 0.25 km, and the maximum detection range is 460 km. The radial resolution is 1 degree. The reflectivity (Z) distance resolution is 1km, ranging from 0 to 460 km. The detection range of Doppler velocity (V) and velocity spectrum width (W) is 230 km, and the distance resolution is 0.25 km [36]. The CINRAD SA scans in the volume coverage pattern (VCP) 21, elevation angles from 0.5 to 14.5 degrees, with 8 effective elevation data, 0.5, 1.5, 2.5, 3.4, 4.3, 6.0, 10, and 14.5 degrees. The scale of tornadoes usually ranges from tens meters to two kilometers. High-resolution radar networks can improve the acquisition and retrieval of tornado features [37]. The distance resolution limited CINRAD SA's capability to detect

the structure of tornadoes finely. However, CINRAD SA can warn and identify tornadoes by monitoring mesocyclone and tornado velocity signatures, which means that artificial intelligence tornado recognition algorithms based on CINRAD SA's tornado characteristics are feasible.

2.2. Tornado Samples

An interpolation algorithm was first used to increase Z distance resolution to 0.25 km when constructing the tornado sample set. Z, V, and W were combined at the same moment and elevation angle. Additionally, the combined data was divided into many 4 × 4 blocks. The characteristics related to tornadoes were calculated in each block, such as maximum, minimum, and average of ZVW, tornado velocity signature, and the range of W, et al., 32 features in total, as shown in the Table A1. The time and coordinate information of tornadoes were used to classify samples (class: yes-tornado = 1 (yes), non-tornado = 0 (no)). The small-scale characteristic of tornadoes leads to the tiny tornado area in Plan Position Indicator (PPI) data. The feature of short generation and disappearance time of tornado results in a tiny proportion of tornado data in the radar database. These two characteristics cause a small number of yes-tornado samples (positive samples), and a large number of non-tornado samples (negative samples) in the tornado sample set obtained by the block segmentation, which will lead to a considerable difference in the proportion of the two-class samples forming imbalanced data, as is shown in Figure 1. Negative samples belong to the majority class samples for the sample set, and positive samples belong to the minority samples. Minority samples tend to have higher importance than majority samples in the tornado classification model. The prediction model obtained from an imbalanced sample set will reduce the recognition effect of the minority class in order to obtain high overall classification accuracy [23]. Calculating the historical data of tornadoes recorded by CINRAD SA from 2005 to 2015, there are a total of 3897 samples, 97 tornado samples (minority class samples), and 3800 non-tornado samples (majority class samples). The class imbalance ratio is relatively high, and the results of tornado detection models are flawed.

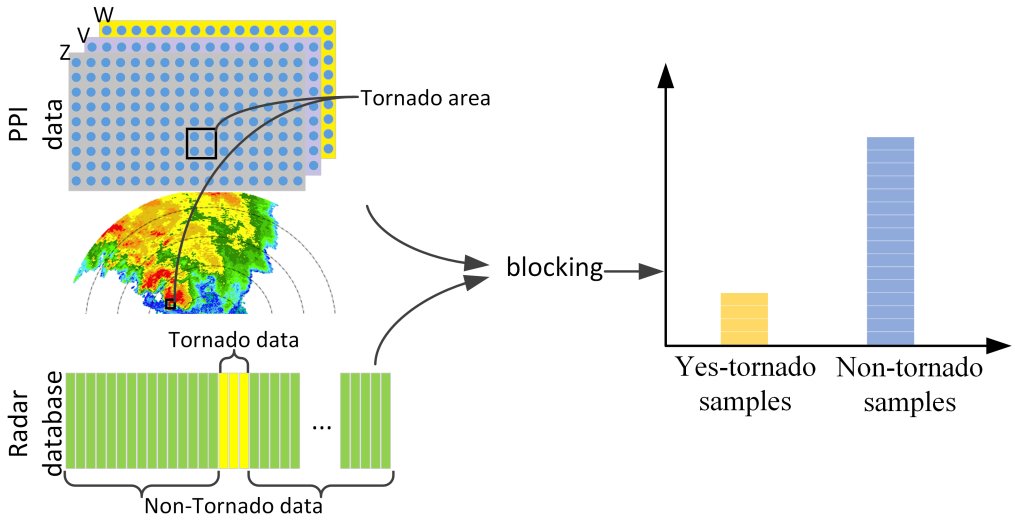

Figure 1. The class imbalance problem of tornado samples.

3. Methods

3.1. ADASYN

One training sample set $D_{tr} = \{\mathbf{x_i}, y_i\}, i = 1, \ldots, m$, where $\mathbf{x_i}$ is a sample vector with n-dimensional features, and $y_i \in Y = \{1, 0\}$, and the m indicates the total number of samples. Firstly, calculate the number of synthetic minority class samples that need to be generated according to Equation (1). The m_s and m_l, respectively, indicate the number of minority class samples and the number of majority class samples in the D_{tr}, $m_s \leq m_l$ and

$m_s + m_l = m$. The $\beta \in [0, 1]$ is used to specify the D_{tr} balance level after the generation of the synthetic samples.

$$G = (m_l - m_s) \times \beta \tag{1}$$

Secondly, find the k-nearest neighbors for each minority sample $\mathbf{x_i}$ according to the Euclidean distance in n-dimensional space, and calculate the ratio r_i according to Equation (2), where Δ_i is the number of majority class samples in the k-nearest neighbors of $\mathbf{x_i}$ and k is equal to the number of k-nearest neighbors. Then, the r_i is normalized to the $\hat{r_i}$ according to Equation (3), where the $\hat{r_i}$ is the density distribution and $\sum_{i=1}^{m_s} \hat{r_i} = 1$.

$$r_i = \frac{\Delta_i}{k}, i = 1, \ldots, m_s \tag{2}$$

$$\hat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{3}$$

Thirdly, calculate the number of synthetic samples needed to be generated for each minority class sample x_i, according to Equation (4).

$$g_i = \hat{r_i} \times G \tag{4}$$

Finally, generate g_i synthetic samples for each minority class sample $\mathbf{x_i}$, according to Equation (5), where the $\mathbf{x_{zi}}$ is randomly selected from the minority samples in the k-nearest neighbors and δ is a random number, $\delta \in [0, 1]$, as is shown in Figure 2 (left).

$$\mathbf{x_{new}} = \mathbf{x_i} + (\mathbf{x_{zi}} - \mathbf{x_i}) \times \delta \tag{5}$$

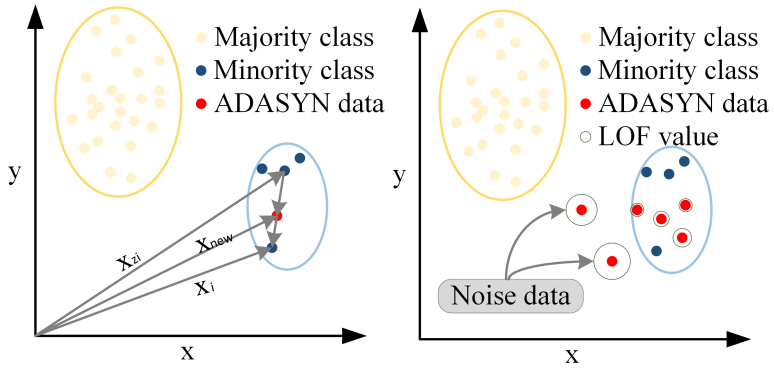

Figure 2. Generation of synthetic samples with the ADASYN approach (**left**), and noise identification with the LOF approach (**right**) (the size of the circle outside of the red sample is the LOF value).

3.2. LOF

After the ADASYN algorithm, the tornado sample set can obtain a balanced ratio, where the number of minority samples: majority samples = 1:1. The synthetic minority samples may have noise samples, and the local outlier factor (LOF) algorithm is used to identify and eliminate noise [31]. The detailed process of the algorithm can refer to reference [38].

For a sample p, the local outlier factor of p is calculated by Equation (6), where the LOF value is the average ratio of the local reachability density of p and those p's k-nearest neighbors. The LOF value of one sample that is not noise is approximately 1. When the LOF value of a sample is significantly greater than 1, it can be labeled as noise, as is shown in Figure 2 (right).

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{\rho_k(o)}{\rho_k(p)}}{|N_k(p)|} \tag{6}$$

3.3. Machine Learning Models

Supporting vector machine (SVM) classification algorithm constructs a hyperplane that separates training samples into binary class, and the SVM is a linear classifier defined in a very high dimensional feature space [39]. The SVM formulation corresponds to the problem of minimizing $||\mathbf{w}||^2/2$ under the constraints $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, \ldots, l$, where the $\mathbf{w}$ is the weight vector that is perpendicular to the separating hyperplane, b is the bias, and l is the number of observations [19]. If the training samples are nonlinearly separable in the feature space, the kernel function is used to increase the dimension of sample space, and the nonlinear problem is converted to a linear problem in a high dimension space, shown in Figure 3 SVM, and Chang et al. developed a library for SVM, including C-SVC, v-SVC, and SVR et al. [40]. The SVM usually outputs classification probabilities by using the Platt scaling method [41].

Artificial neural networks (ANN) algorithm has attracted much research in the past few years, and several studies have been applied to the weather radar, such as a study that combined the generative adversarial networks (GNN) and super-resolution reconstruction of weather radar echo images [42]. Another study applied a deep convolutional neural network (DCNN) to NEXRAD PPI scans, and the increased resolution and frequency content improved observation capabilities [43]. The structure of ANN includes: one input layer, several hidden layers, one output layer, and the hidden layers connect the input and output (as is shown in Figure 3 ANN) [44]. The ANN uses functions, such as tanh and sigmoid, to map and activate neurons, and the ANN requires multiple rounds of iterative training to minimize loss and achieve good accuracy [45,46]. Binary ANN usually uses 0.5 as the threshold of classification probability to classify samples.

Breiman proposed the random forest (RF) algorithm in 2001. RF constructs multiple classification trees through randomly sampling samples and randomly selecting features and uses a voting mechanism to make prediction and classification, and outputs probabilities according to the voting results. (shown in Figure 3 RF) [47–49]. The RF usually uses ID3, C4.5, and GINI methods [50,51]. ID3 cannot handle the problem of continuous attributes, but the C4.5 algorithm can handle it. The Gini index reflects the purity of a dataset, and the smaller the value, the higher the purity. The RF is a multivariate nonlinear classification model, avoiding model overfitting with less sensitivity to noise [52]. RF has been widely used in the field of remote sensing [53–55] and extreme weather warnings [12,56,57].

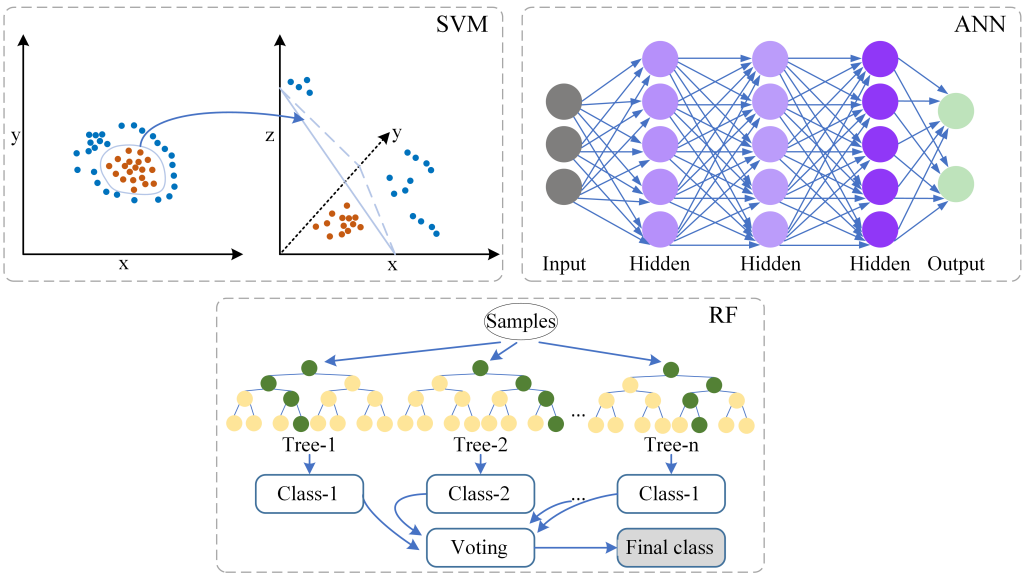

Figure 3. The supporting vector machine (SVM), artificial neural network (ANN), and random forest (RF) algorithms.

4. Experiments

4.1. Experiment 1

In order to obtain qualitative differences between models with and without ADASYN-LOF algorithm, the numerical results need to be compared. The tornado samples were divided into training and testing samples, where the number of training samples: the number of testing samples = 1:1, and the number of positive samples in training samples: the number of positive samples in testing samples = 1:1, and the number of negative samples in training samples: the number of negative samples in testing samples = 1:1 (training set: 1900 negative samples, 49 positive samples, and testing set: 1900 negative samples, 48 positive samples). This experiment steps are shown in Figure 4. We created a copy of the training samples that were directly used to build models (SVM (IBD), ANN (IBD), RF (IBD)). The original training samples were processed by the ADASYN approach, so the number of positive samples was equal to the number of negative samples; then, the LOF algorithm was used to identify the noise of balanced data. After the LOF approach, models (SVM (BD), ANN (BD), RF (BD)) were obtained. ADASYN's $k = 20$, LOF's $k = 20$, and LOF eliminated 93 noise samples during this experiment. The testing samples were directly used to obtain models' quantitative performance, and the binary classification confusion matrix (Table 1) was used. In the confusion matrix, the TP is the number of correct yes-tornado samples predicted by the model, FP is the number of non-tornado samples that the model misclassifies as yes-tornado samples, FN is the number of yes-tornado samples that are misclassified as non-tornado samples, and TN is the number of non-tornado samples correctly classified by the model. According to TP, FP, FN, and TN, the accuracy (7), precision (8), F-score (9), and G-mean (10) can be obtained, and the F-score equals to F1-score when $\beta = 1$ and $Recall = TP/(TP + FN)$. In addition, in order to compare the performance of different models, the Area Under Curve (AUC) score was used. AUC is defined as the area under the receiver operating characteristics curve enclosed by the coordinate axis. The larger the AUC value, the better the average performance of the model. When assessing the weather forecast model, the contingency table was usually used to evaluate the forecast accuracy. So, combining the confusion matrix and the 2×2 contingency table (Table 2), POD (11), FAR (12), and CSI (13) were obtained. The different model performance results show in Table 3.

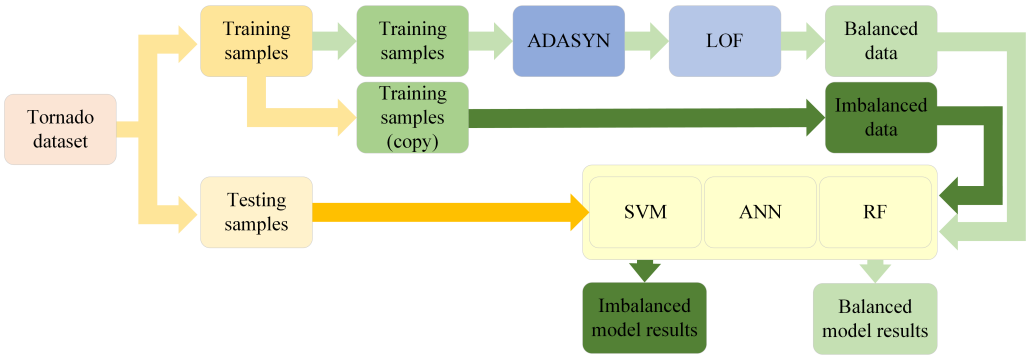

Figure 4. The flow chart of experiment 1.

Table 1. Binary classification confusion matrix.

		True Class	
Model prediction	Y (yes-tornado)	Positive (yes-tornado)	Negative (non-tornado)
	N (non-tornado)	TP (True Positives)	FP (False Positives)
		FN (False Negatives)	TN (True Negatives)
	Column counts	$P_C = TP + FN$	$N_C = FP + TN$

Table 2. 2 × 2 contingency table (X, Y, Z represent the number of times the model correctly warned, missed, falsely warned weather processes, respectively, and W represents the number of times the model correctly warned of no weather processes).

		Warning	
		YES	NO
weather event	YES	X	Y
	NO	Z	W

Table 3. The performance of different models.

Model	Evaluation	ADASYN-LOF	NONE
SVM	ACC	0.9277	0.9317
	PRE	0.7385	0.9211
	F1-score	**0.8421**	0.8046
	G-mean	**0.9467**	0.8388
	AUC	**0.9473**	0.8496
	POD	**0.9796**	0.7143
	FAR	0.2615	0.0790
	CSI	**0.7273**	0.6731
ANN	ACC	**0.9438**	0.9237
	PRE	**0.9070**	0.8750
	F1-score	**0.8478**	0.7856
	G-mean	**0.8832**	0.8354
	AUC	**0.8880**	0.8446
	POD	**0.7959**	0.7142
	FAR	**0.0930**	0.1250
	CSI	**0.7358**	0.6481
RF	ACC	**0.9438**	0.8916
	PRE	0.9268	0.9583
	F1-score	**0.8444**	0.6301
	G-mean	**0.8740**	0.6834
	AUC	**0.8803**	0.7322
	POD	**0.7755**	0.4694
	FAR	0.0732	0.0417
	CSI	**0.7308**	0.4600

$$accuracy(ACC) = \frac{TP + TN}{P_C + N_C} \tag{7}$$

$$precision(PRE) = \frac{TP}{TP + FP} \tag{8}$$

$$F - score = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{9}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{10}$$

$$POD = \frac{X}{X + Y} = \frac{TP}{TP + FN} \tag{11}$$

$$FAR = \frac{Z}{X + Z} = \frac{FP}{TP + FP} \tag{12}$$

$$CSI = \frac{X}{X + Y + Z} = \frac{TP}{TP + FN + FP} \tag{13}$$

4.2. Experiment 2

In order to compare the performance of different models in actual tornado detection, while making full use of all available samples, all tornado samples were used to train models. The experiment steps are shown in Figure 5. Create a copy of the training samples that were directly used to build models (SVM (IBD), ANN (IBD), RF (IBD)). For the original training samples, the ADASYN algorithm was used to balance the ratio of yes-tornado samples and non-tornado samples to 1 : 1. The LOF approach identified the noise samples of synthetic samples. Then, models (SVM (BD), ANN (BD), RF (BD)) were obtained, and models were used to detect tornadoes from 2016–2018, and the results are shown in Section 5.2. ADASYN's $k = 20$, LOF's $k = 20$, and LOF eliminated 246 noise samples during this experiment.

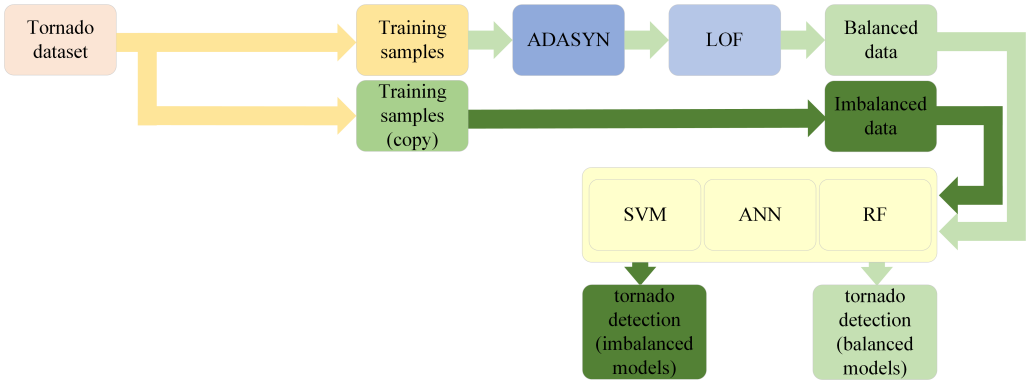

Figure 5. The flow chart of experiment 2.

5. Results and Discussion

The sample set that has the class imbalance problem is imbalanced data, forming imbalanced models, such as SVM (IBD), ANN (IBD), RF (IBD). Similarly, the balanced data, without the class imbalance problem, forms balanced models, such as SVM (BD), ANN (BD), RF (BD).

5.1. Model Performance

In Table 3, the different models' results were compared. The proposed approach by combing ADASYN and LOF in handling training samples is called the ADASYN-LOF approach (ALA), and the NONE indicates that the models were built by the original training samples (copy). After the ALA, the SVM's ACC and PRE decreased, the ANN's ACC and PRE increased, and the RF's ACC and PRE increased. The balanced models had a better F1-score, G-mean, and AUC than imbalanced models, which indicates that the ALA improves the performance of models. For the AUC after the ALA, the SVM's AUC had the maximum performance improvement, and the AUC score order was: SVM > ANN > RF, indicating that the average performance of the balanced model is: the SVM is the best, followed by ANN, and final RF. The balanced SVM's POD was greatly improved, and the CSI increased, but the FAR also increased. The balanced ANN had a better POD, CSI, and FAR than the imbalanced ANN. The balanced RF had a better performance of POD and CSI and worse FAR performance than imbalanced RF. In terms of POD, FAR and CSI after the ALA, the biggest improvement was ANN. The POD order after the ALA was SVM > ANN > RF. Although the POD of SVM was greater (>0.15) than the ANN and RF, the SVM's FAR was much higher than the POD of ANN and RF. The high FAR caused the SVM's CSI to be slightly smaller than the CSI of ANN and RF.

The yes-tornado and non-tornado samples are unequally distributed in the imbalanced sample set, which leads to the models having a high misclassification rate of yes-tornado samples and relatively low G-mean, F1-score, POD, and CSI. After the two class samples are in a balanced distribution, the models' ability to carry out predictive accuracy in

determining the yes-tornado samples is improved, thereby increasing the G-mean, F1-score, POD, and CSI.

5.2. Tornado Detection Results

When using the models to detect tornado cases, the historical tornado events that were not included in the training and testing samples were used from 2016 to 2018. The case requirements are met: the distance between the tornado and the radar center is no more than 130 km, and the Meteorological Bureau has official records about the tornado. In this section, the model detection results are represented by black asterisks, the value is the classification probability of the model, and the results are displayed in reflectivity Z. When there is no black asterisk in the Z, no samples are classified as yes-tornado samples by the model. The detection results of different models after the ALA are called SVM detection results (BD), ANN detection results (BD), and RF detection results (BD). Similarly, the model detection results without the ALA are called SVM detection results (IBD), ANN detection results (IBD), and RF detection results (IBD). The upper and lower subgraphs on the far right are the Doppler velocity V and velocity spectrum width W, respectively, at the same moment and elevation angle as Z.

The first case was the EF4 tornado that touched the ground in Funing, Jiangsu Province, at about 14:30 (Beijing time, UTC+8) on 23 June 2016. The imbalanced models were used to warn tornadoes, using radar 1.5-degree elevation level-II data from 14:00 to 14:30 (Beijing time, UTC+8). The imbalanced models' first tornado warning was at 14:14 (Beijing time, UTC+8), shown in Figure 6 SVM detection results (IBD), ANN detection results (IBD), and RF detection results (IBD). The balanced models were used for tornado warning detection, using radar 1.5-degree elevation level-II data from 14:00 to 14:30 (Beijing time, UTC+8). The balanced models can issue a tornado touchdown warning at 14:14 (Beijing time, UTC+8) in the same area with a greater probability the imbalanced models (SVM: $0.99 > 0.98$, ANN: $0.99 > 0.97$, RF: $0.91 > 0.81$), as shown in Figure 6 SVM detection results (BD), ANN detection results (BD) and RF detection results (BD). In addition, the balanced models' tornado early-warning time was advanced to 14:08 (Beijing time, UTC+8), as shown in Figure 7 SVM detection results (BD) and RF detection results (BD), when the imbalanced models issued no warnings, as shown in Figure 7 SVM detection results (IBD), ANN detection results (IBD), and RF detection results (IBD).

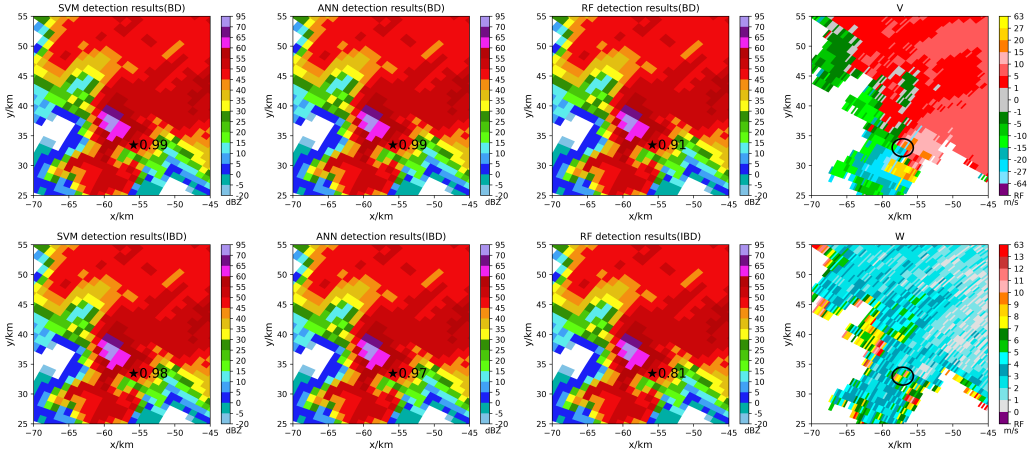

Figure 6. The results of models early warning the tornado at 14:14 (Beijing time, UTC+8) 1.5-degree (the black circle centered at the models warning results with a radius of 1.5 km, the SVM, ANN, RF represent Support Vector Classifier, Artificial Neural Network Classifier, Random Forest Classifier, the V represents Doppler Velocity, and the W represents Doppler Velocity Spectral Width. The BD indicates that the classifier was formed on a balanced tornado dataset, and the IBD indicates that the classifier was formed on an imbalanced tornado dataset.).

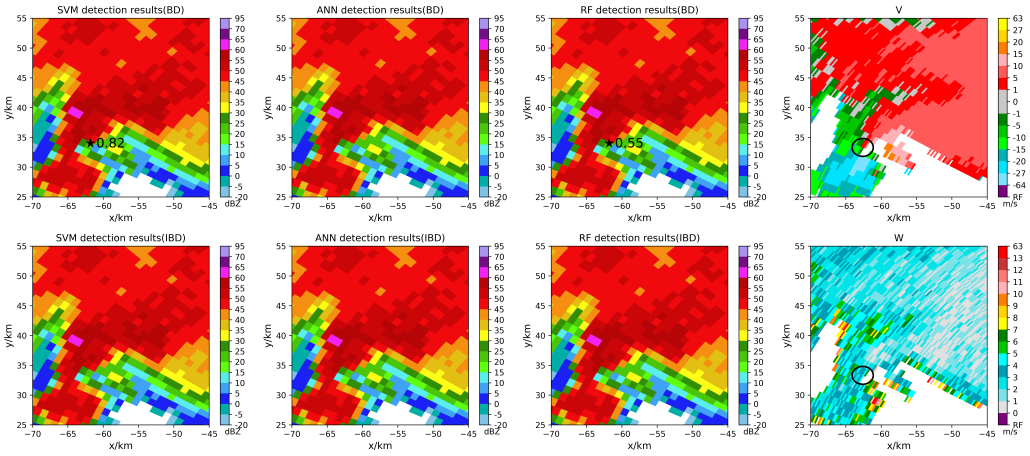

Figure 7. The results of models early warning the tornado at 14:08 (Beijing time, UTC+8) 1.5-degree (the black circle centered at the models warning results with a radius of 1.5 km, the SVM, ANN, RF represent Support Vector Classifier, Artificial Neural Network Classifier, Random Forest Classifier, the V represents Doppler Velocity, and the W represents Doppler Velocity Spectral Width. The BD indicates that the classifier was formed on a balanced tornado dataset, and the IBD indicates that the classifier was formed on an imbalanced tornado dataset.).

The second tornado case occurred in Dongtai, Jiangsu Province at around 11:00 (Beijing time, UTC+8), on 2 July 2017. The tornado was 77 km away from the radar center, and the sample variable value calculated by the block segmentation was small, which caused the imbalanced models to fail to recognize this tornado, as shown in Figure 8 SVM detection results (IBD), ANN detection results (IBD), and RF detection results (IBD). The balanced models were used to detect the tornado, the tornado was identified, with relatively a high probability of being classified as yes-tornado (SVM: 0.99, ANN: 0.99 and 0.98, RF: 0.8), as shown in Figure 8 SVM detection results (BD), ANN detection results (BD), and RF detection results (BD).

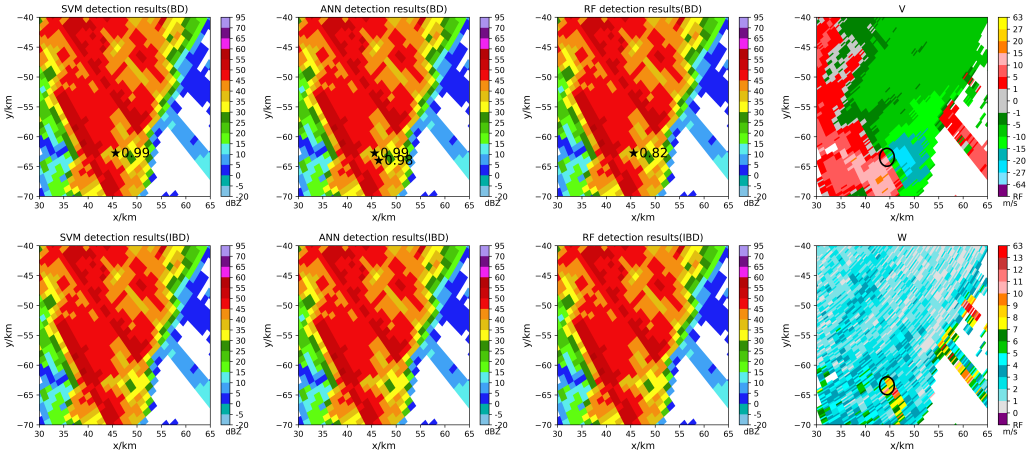

Figure 8. The tornado identification results of models at 11:01 (Beijing time, UTC+8) 0.5-degree (the black circle centered at the tornado location with a radius of 1.5 km, the SVM, ANN, RF represent Support Vector Classifier, Artificial Neural Network Classifier, Random Forest Classifier, the V represents Doppler Velocity, and the W represents Doppler Velocity Spectral Width. The BD indicates that the classifier was formed on a balanced tornado dataset, and the IBD indicates that the classifier was formed on an imbalanced tornado dataset.).

The third case was the tornado that occurred in the outer circulation of Typhoon Wembia No.1815 in 2018, which touched the ground in Xuzhou, Jiangsu Province at around 18:40 (Beijing time, UTC+8), on 18 August. The tornado was far away from the radar center, and the distance was 120.5 km. When the detection range of CINRAD SA is

more significant than 100 km, CINRAD SA suffers from beam broadening and power attenuation, so only partial information of the tornado can be obtained. In Figure 9 V, although $\triangle V = |V_- - V_+| = 26.5$ m/s at the 0.5-degree elevation, the radar TVS product did not issue a tornado warning because the thresholds of TVS were not met. The imbalanced models were used to detect this tornado, and no tornado warnings were issued, as shown in Figure 9 SVM detection results (IBD), ANN detection results (IBD), and RF detection results (IBD). The balanced SVM and ANN model identified this tornado, as shown in Figure 9 SVM detection results (BD), ANN detection results (BD). However, the balanced RF model did not issue this tornado, as shown in Figure 9 RF detection results (BD).

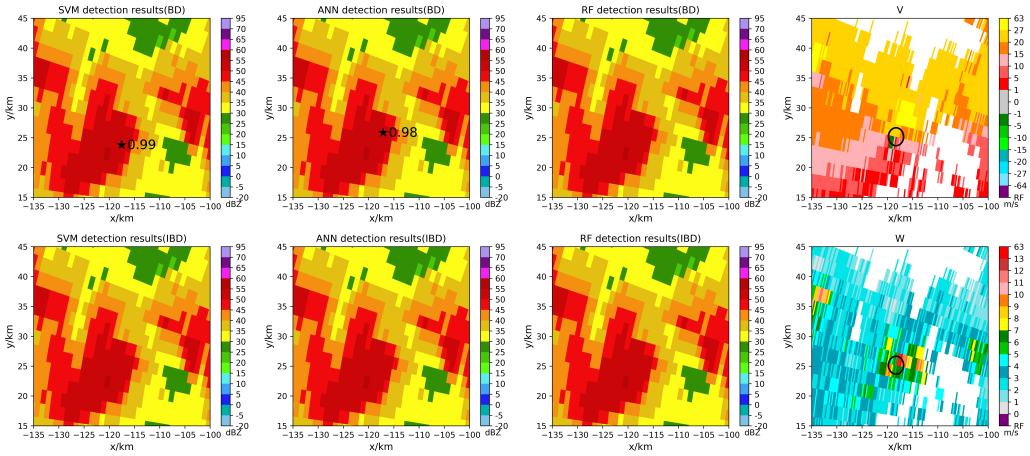

Figure 9. The tornado identification results of models at 18:45 (Beijing time, UTC+8) 0.5-degree (the black circle centered at the tornado location with a radius of 1.5 km, the SVM, ANN, RF represent Support Vector Classifier, Artificial Neural Network Classifier, Random Forest Classifier, the V represents Doppler Velocity, and the W represents Doppler Velocity Spectral Width. The BD indicates that the classifier was formed on a balanced tornado dataset, and the IBD indicates that the classifier was formed on an imbalanced tornado dataset.).

In the first tornado case, the balanced and imbalanced models were used to compare the tornado's early warning time. The first tornado warning of the imbalanced models was at 14:14 (Beijing time, UTC+8), and the first tornado warning of balanced models, SVM and RF, was at 14:08 (Beijing time, UTC+8). The balanced models increased the tornado early warning time from 16 min to 22 min, indicating that the ALA optimizes the distribution of samples and can advance the tornado early warning time. In addition, the balanced models had a higher probability than the imbalanced models (SVM BD: 0.99 > SVM IBD: 0.98, ANN BD: 0.99 > SVM IBD: 0.97, RF BD: 0.99 > RF IBD: 0.81), which indicates that the results of the balanced models have higher credibility than the results of the imbalanced models.

In the second tornado case, the scale of the tornado and the sample features were small, which caused the imbalanced models cannot identify this tornado. The balanced models recognized the tornado, and the balanced models had better F1-score and G-mean score than the imbalanced models in Table 3, which confirms that balanced models have better classification performance and can warn more tornado cases than the imbalanced models, making up for the shortcomings of the imbalanced models. In addition, it is worth mentioning that there were two asterisks in the Figure 8 ANN detection results, this was because: when the models were used to detect tornadoes, the intersection between adjacent blocks was also calculated, as shown in Figure A1.

In the third case, the tornado was far away from the radar. The radar was heavily affected by beam broadening and power attenuation, resulting in the TVS algorithm failing to issue tornado warnings. For similar reasons, in the detection results at 18:45 (Beijing time, UTC+8), the imbalanced models could not identify the tornado, but the balanced SVM and ANN model identified the tornado. The balanced and imbalanced RF models did not

issue any tornado warnings. It is speculated that the negative velocity value of the tornado was small, which caused the failure of RF models. In this tornado case, the performance of the balanced model was: ANN > SVM > RF and the average performance of the balanced model obtained in Table 2 was SVM > ANN > RF. The difference in performance is because the internal classification criteria of different models is different, which indicates that multiple models should be coordinated in the actual tornado warning.

In addition to using specific tornado cases to test the models, this experiment compared the noise immunity performance of balanced models and imbalanced models (figures omitted), and the results show that the balanced models have more robust noise immunity performance than the imbalanced models. Especially when the radar is of poor quality, the balanced models issue fewer or no false warnings than the imbalanced models.

Before studying the ALA, weight and cost methods were used to solve the imbalance. However, due to the small number of positive samples, the methods (adding weights for different class) did not generate new samples and did not improve the problem of missing tornadoes. The study compared the performance of the ADASYN-LOF, ADASYN, SMOTE-LOF, and SMOTE algorithms on the dataset, as shown in Table A2. For the SVM, the ADASYN-LOF's ACC, PRE, F1-score, G-mean, AUC, FAR, and CSI were better than the ADASYN, SMOTE-LOF, and SMOTE. For the ANN, the ADASYN-LOF's AUC, F1-score, G-mean, AUC, POD, CSI were better than the other algorithms. For the RF, the PODs of ADASYN-LOF and SMOTE-LOF were equal. Generally, if using SVM or ANN as a classifier, it is better to use ADASYN-LOF to preprocess imbalanced data. For RF, the SMOTE-LOF could be better.

The LOF algorithm can also be used for unsupervised classification, and it is hoped that subsequent research will apply this method to the detection of tornadoes (outliers).

6. Conclusions

The tornado sample set usually has the class imbalance problem that might cause the machine learning models to have a poor tornado detection effect. The adaptive synthetic (ADASYN) sampling approach is used to solve the problem, and the local outlier factor (LOF) algorithm is applied to identify noise data in synthetic samples. The ADASYN and LOF approach is called the ADASYN-LOF approach (ALA). The SVM, ANN, RF models are used and the main conclusions are as follows.

1. After the ALA, the accuracy and precision are increased or decreased, the F1-score, G-mean, AUC, POD, CSI are significantly improved, the average performance is improved, and models have better noise immunity performance than the models without the approach.
2. Using specific tornado cases to test models, the balanced models have the following advantages after the ALA.
    - In the early tornado warning, the models have the potential to increase the early warning time of tornadoes touching the ground.
    - The balanced models can identify some tornadoes that cannot be identified by the imbalanced models.
    - The models can identify tornadoes that cannot be detected due to the limitation of the tornado velocity signature (TVS) algorithm threshold.
3. Compared with the ADASYN, SMOTE-LOF, and SMOTE algorithms, the ALA performs better in preprocessing imbalacned data if SVM or ANN is used as the classifier. If RF is used, the SMOTE-LOF algorithm could work better.

There are three directions for future research:

- optimize the k value of the ALA and appropriately reduce the dimension of sample features;
- study how to appropriately decrease the majority samples when applying the ALA;
- use more datasets (such as tornado datasets in the United States) to evaluate the ALA and apply outlier detection algorithms to detect tornadoes.

Author Contributions: Conceptualization, Z.Q. and Q.Z.; methodology, Z.Q.; software, Z.Q.; validation, Q.Z., H.W. and T.X.; formal analysis, Q.Z.; investigation, S.Z.; resources, Y.L.; data curation, Q.Z.; writing—original draft preparation, Z.Q.; writing—review and editing, Z.Q., Q.Z. and H.W.; visualization, Q.Z.; supervision, Q.Z.; project administration, Z.Q.; funding acquisition, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (U20B2061), the National Key R&D Program of China (2018YFC01506100), Department of Science and Technology of Sichuan Province (2020ZYD051, 2022YFS0541), the Open Grants of the State Key Laboratory of Severe Weather (2020LASW-B11) and the fund of "Key Laboratory of Atmosphere Sounding, CMA" (2021KLAS01M).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank the reviewers for their constructive comments and editorial suggestions that significantly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
ADASYN	Adaptive Synthetic
LOF	Local Outlier Factor
ALA	ADASYN-LOF algorithm
SVM	Supporting Vector Machine
ANN	Artificial Neural Network
RF	Random Forest
TVS	Tornado Velocity Signature
TDS	Tornado Debris Signature
CINRAD SA	the S-band China New Generation of Weather Radar
Z	Reflectivity
V	Doppler Velocity
W	Velocity Spectrum Width
VCP	Volume Coverage Pattern
PPI	Plan Position Indicator
BD	Balanced Dataset
IBD	Imbalanced Dataset

Appendix A

Table A1. The 32 features of the tornado sample.

Feature	Description	Unit
r_average	the average value in the 4 × 4 Z block	dBZ
r_max	the maximum value in the 4 × 4 Z block	dBZ
r_min	the minimum value in the 4 × 4 Z block	dBZ
v_average	the average value in the 4 × 4 V block	m/s
v_max	the maximum value in the 4 × 4 V block	m/s
v_min	the minimum value in the 4 × 4 V block	m/s
w_average	the average value in the 4 × 4 W block	m/s
w_max	the maximum value in the 4 × 4 W block	m/s
w_min	the minimum value in the 4 × 4 W block	m/s
s_average	the average value of velocity shear in the 4 × 4 V block	1/s
s_max	the maximum value of velocity shear in the 4 × 4 V block	1/s
s_min	the minimum value of velocity shear in the 4 × 4 V block	1/s
l_average	the average value of angular momentum in the 4 × 4 V block	m^2/s
l_max	the maximum value of angular momentum in the 4 × 4 V block	m^2/s
l_min	the minimum value of angular momentum in the 4 × 4 V block	m^2/s
vt_average	the average value of rotation speed in the 4 × 4 V block	m/s
vt_max	the maximum value of rotation speed in the 4 × 4 V block	m/s
vt_min	the minimum value of rotation speed in the 4 × 4 V block	m/s
c4_d_v_max	the maximum value of velocity difference in the 2 × 2 V block	m/s
c4_s_average	the average value of velocity shear in the 2 × 2 V block	1/s
c4_s_max	the maximum value of velocity shear in the 2 × 2 V block	1/s
c4_s_min	the minimum value of velocity shear in the 2 × 2 V block	1/s
c4_l_average	the average value of angular momentum in the 2 × 2 V block	m^2/s
c4_l_max	the maximum value of angular momentum in the 2 × 2 V block	m^2/s
c4_l_min	the minimum value of angular momentum in the 2 × 2 V block	m^2/s
c4_vt_average	the average value of rotation speed in the 2 × 2 V block	m/s
c4_vt_max	the maximum value of rotation speed in the 2 × 2 V block	m/s
c4_vt_min	the minimum value of rotation speed in the 2 × 2 V block	m/s
w_range	the range value of velocity spectral width in the 4 × 4 W block	m/s
w_40	the threshould greater than 40% velocity spectral width in the 4 × 4 W block	m/s
w_60	the threshould greater than 60% velocity spectral width in the 4 × 4 W block	m/s
w_80	the threshould greater than 80% velocity spectral width in the 4 × 4 W block	m/s

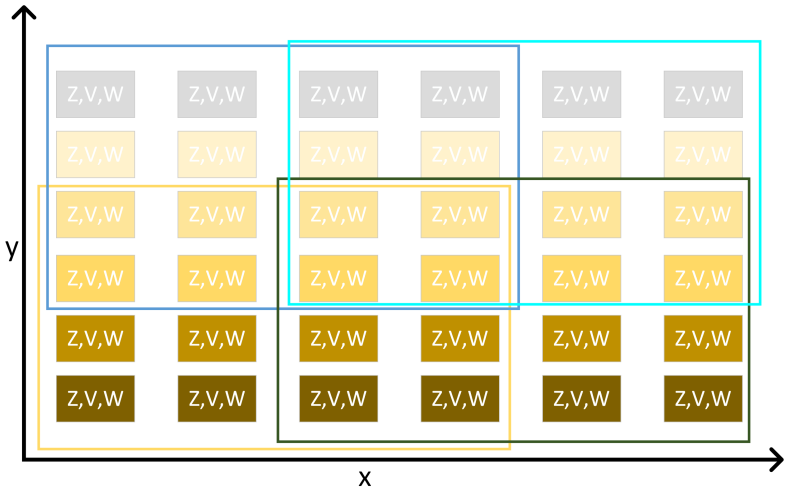

Figure A1. The calculation of intersection between adjacent blocks.

Table A2. The performance of ADASYN-LOF, ADASYN, SMOTE-LOF, SMOTE algorithms.

	Evaluation	ADASYN-LOF	ADASYN	SMOTE-LOF	SMOTE
SVM	ACC	**0.9277**	0.9197	0.9237	0.9116
	PRE	**0.7385**	0.7164	0.7273	0.6957
	F1-score	**0.8421**	0.8276	0.8348	0.8136
	G-mean	**0.9467**	0.9416	0.9442	0.9363
	AUC	**0.9473**	0.9423	0.9448	0.9373
	POD	0.9796	0.9796	0.9796	0.9796
	FAR	**0.2615**	0.2836	0.2727	0.3043
	CSI	**0.7273**	0.7059	0.7164	0.6857
ANN	ACC	**0.9438**	0.9237	0.9398	0.9398
	PRE	0.9070	0.8947	**0.9250**	0.9048
	F1-score	**0.8478**	0.7816	0.8315	0.8352
	G-mean	**0.8832**	0.8246	0.8624	0.8718
	AUC	**0.8880**	0.8369	0.8701	0.8778
	POD	**0.7959**	0.6939	0.7551	0.7755
	FAR	0.0930	0.1053	**0.0750**	0.0952
	CSI	**0.7358**	0.6415	0.7115	0.7170
RF	ACC	0.9438	0.9357	**0.9478**	0.9398
	PRE	0.9268	0.9231	0.9500	**0.9722**
	F1-score	0.8444	0.8182	**0.8539**	0.8235
	G-mean	0.8740	0.8507	**0.8762**	0.8430
	AUC	0.8803	0.8598	**0.8828**	0.8546
	POD	**0.7755**	0.7347	**0.7755**	0.7143
	FAR	0.0732	0.0769	0.0500	**0.0278**
	CSI	0.7308	0.6923	**0.7451**	0.7000

References

1. Chen, J.; Cai, X.; Wang, H.; Kang, L.; Zhang, H.; Song, Y.; Zhu, H.; Zheng, W.; Li, F. Tornado climatology of China. *Int. J. Climatol.* **2018**, *38*, 2478–2489. [CrossRef]
2. McCarthy, D.; Schaefer, J.; Edwards, R. What are we doing with (or to) the F-Scale. In Proceedings of the 23rd Conference on Severe Local Storms, St. Louis, MO, USA, 6–10 November 2006; Volume 5.
3. Doswell, C.A., III; Brooks, H.E.; Dotzek, N. On the implementation of the enhanced Fujita scale in the USA. *Atmos. Res.* **2009**, *93*, 554–563. [CrossRef]
4. Brown, R.A.; Lemon, L.R.; Burgess, D.W. Tornado detection by pulsed Doppler radar. *Mon. Weather Rev.* **1978**, *106*, 29–38. [CrossRef]
5. Zrnić, D.; Burgess, D.; Hennington, L. Automatic detection of mesocyclonic shear with Doppler radar. *J. Atmos. Ocean. Technol.* **1985**, *2*, 425–438. [CrossRef]
6. Stumpf, G.J.; Witt, A.; Mitchell, E.D.; Spencer, P.L.; Johnson, J.; Eilts, M.D.; Thomas, K.W.; Burgess, D.W. The National Severe Storms Laboratory mesocyclone detection algorithm for the WSR-88D. *Weather Forecast.* **1998**, *13*, 304–326. [CrossRef]
7. Mitchell, E.D.W.; Vasiloff, S.V.; Stumpf, G.J.; Witt, A.; Eilts, M.D.; Johnson, J.; Thomas, K.W. The national severe storms laboratory tornado detection algorithm. *Weather Forecast.* **1998**, *13*, 352–366. [CrossRef]
8. Ryzhkov, A.V.; Schuur, T.J.; Burgess, D.W.; Zrnic, D.S. Polarimetric tornado detection. *J. Appl. Meteorol.* **2005**, *44*, 557–570. [CrossRef]
9. Wang, Y.; Yu, T.Y.; Yeary, M.; Shapiro, A.; Nemati, S.; Foster, M.; Andra, D.L., Jr.; Jain, M. Tornado detection using a neuro–fuzzy system to integrate shear and spectral signatures. *J. Atmos. Ocean. Technol.* **2008**, *25*, 1136–1148. [CrossRef]
10. Alberts, T.A.; Chilson, P.B.; Cheong, B.; Palmer, R. Evaluation of weather radar with pulse compression: Performance of a fuzzy logic tornado detection algorithm. *J. Atmos. Ocean. Technol.* **2011**, *28*, 390–400. [CrossRef]
11. Wang, Y.; Yu, T.Y. Novel tornado detection using an adaptive neuro-fuzzy system with S-band polarimetric weather radar. *J. Atmos. Ocean. Technol.* **2015**, *32*, 195–208. [CrossRef]
12. Hill, A.J.; Herman, G.R.; Schumacher, R.S. Forecasting Severe Weather with Random Forests. *Mon. Weather Rev.* **2020**, *148*, 2135–2161. [CrossRef]
13. Basalyga, J.N.; Barajas, C.A.; Gobbert, M.K.; Wang, J.W. Performance Benchmarking of Parallel Hyperparameter Tuning for Deep Learning Based Tornado Predictions. *Big Data Res.* **2021**, *25*, 100212. doi: ARTN 100212 10.1016/j.bdr.2021.100212. [CrossRef]
14. Rout, N.; Mishra, D.; Mallick, M.K. Handling imbalanced data: A survey. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 431–443.

15. Kaur, H.; Pannu, H.S.; Malhi, A.K.J.A.C.S. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.* **2019**. *52*, 1–36. [CrossRef]
16. Lin, W.C.; Tsai, C.F.; Hu, Y.H.; Jhang, J.S. Clustering-based undersampling in class-imbalanced data. *Inf. Sci.* **2017**. *409*, 17–26. [CrossRef]
17. Choi, W.; Heo, J.; Ahn, C. Development of Road Surface Detection Algorithm Using CycleGAN-Augmented Dataset. *Sensors* **2021**, *21*, 7769. [CrossRef] [PubMed]
18. Setiawan, B.D.; Serdült, U.; Kryssanov, V. A Machine Learning Framework for Balancing Training Sets of Sensor Sequential Data Streams. *Sensors* **2021**. *21*, 6892. [CrossRef]
19. Trafalis, T.B.; Adrianto, I.; Richman, M.B.; Lakshmivarahan, S. Machine-learning classifiers for imbalanced tornado data. *Comput. Manag. Sci.* **2014**, *11*, 403–418. [CrossRef]
20. Maalouf, M.; Trafalis, T.B. Robust weighted kernel logistic regression in imbalanced and rare events data. *Comput. Stat. Data Anal.* **2011**, *55*, 168–183. [CrossRef]
21. Maalouf, M.; Siddiqi, M. Weighted logistic regression for large-scale imbalanced and rare events data. *Knowl.-Based Syst.* **2014**, *59*, 142–148. [CrossRef]
22. Maalouf, M.; Homouz, D.; Trafalis, T.B. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Comput. Intell.* **2018**, *34*, 161–174. [CrossRef]
23. He, H.B.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. doi: 10.1109/Tkde.2008.239. [CrossRef]
24. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
25. Yu, L.; Zhou, N. Survey of Imbalanced Data Methodologies. *arXiv* **2021**, arXiv: 2104.02240.
26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
27. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
28. Susan, S.; Kumar, A. SSOMaj-SMOTE-SSOMin: Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets. *Appl. Soft Comput.* **2019**, *78*, 141–149. [CrossRef]
29. Ye, X.C.; Li, H.M.; Imakura, A.; Sakurai, T. An oversampling framework for imbalanced classification based on Laplacian eigenmaps. *Neurocomputing* **2020**, *399*, 107–116. [CrossRef]
30. Tripathi, A.; Chakraborty, R.; Kopparapu, S.K. A Novel Adaptive Minority Oversampling Technique for Improved Classification in Data Imbalanced Scenarios. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 10650–10657.
31. Maulidevi, N.U.; Surendro, K. SMOTE-LOF for noise identification in imbalanced data classification. *J. King Saud-Univ.-Comput. Inf. Sci.* **2021**. [CrossRef]
32. Heinselman, P.; LaDue, D.; Kingfield, D.M.; Hoffman, R. Tornado warning decisions using phased-array radar data. *Weather Forecast.* **2015**, *30*, 57–78. [CrossRef]
33. Jun, L.; Honggen, Z.; Jianbing, L.; Xinan, L. Research on the Networking Strategy of Tornado Observation Network in Northern Jiangsu Based on X-band Weather Radar. In Proceedings of the 2019 International Conference on Meteorology Observations (ICMO), Chengdu, China, 28–31 December 2019; pp. 1–4.
34. Adachi, T.; Mashiko, W. High temporal-spatial resolution observation of tornadogenesis in a shallow supercell associated with Typhoon Hagibis (2019) using phased array weather radar. *Geophys. Res. Lett.* **2020**, *47*, e2020GL089635. [CrossRef]
35. Yoshida, S.; Misumi, R.; Maesaka, T. Early Detection of Convective Echoes and Their Development Using a Ka-Band Radar Network. *Weather Forecast.* **2021**, *36*, 253–264. [CrossRef]
36. Zhang, X.; He, J.; Zeng, Q.; Shi, Z. Weather Radar Echo Super-Resolution Reconstruction Based on Nonlocal Self-Similarity Sparse Representation. *Atmosphere* **2019**, *10*, 254. [CrossRef]
37. Chen, H.; Chandrasekar, V. Real-time wind velocity retrieval in the precipitation system using high-resolution operational multi-radar network. In *Remote Sensing of Aerosols, Clouds, and Precipitation*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 315–339.
38. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
39. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
40. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. *Acm Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]
41. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
42. Chen, H.; Zhang, X.; Liu, Y.; Zeng, Q. Generative adversarial networks capabilities for super-resolution reconstruction of weather radar echo images. *Atmosphere* **2019**, *10*, 555. [CrossRef]
43. Geiss, A.; Hardin, J.C. Radar super resolution using a deep convolutional neural network. *J. Atmos. Ocean. Technol.* **2020**, *37*, 2197–2207. [CrossRef]

44. Maind, S.B.; Wankar, P. Research paper on basic of artificial neural network. *Int. J. Recent Innov. Trends Comput. Commun.* **2014**, *2*, 96–100.
45. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717–727. [CrossRef]
46. Zhang, Z., Artificial neural network. In *Multivariate Time Series Analysis in Climate and Environmental Research*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1–35.
47. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
48. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
49. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
50. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: London, UK, 2017.
51. Elaidi, H.; Benabbou, Z.; Abbar, H. A comparative study of algorithms constructing decision trees: Id3 and c4. 5. In Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications, Rabat, Morocco, 2–5 May 2018; pp. 1–5.
52. Peerbhay, K.Y.; Mutanga, O.; Ismail, R. Random Forests Unsupervised Classification: The Detection and Mapping of Solanum mauritianum Infestations in Plantation Forestry Using Hyperspectral Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3107–3122. [CrossRef]
53. Dong, Y.N.; Du, B.; Zhang, L.P. Target Detection Based on Random Forest Metric Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1830–1838. [CrossRef]
54. Dong, L.; Du, H.; Mao, F.; Han, N.; Li, X.; Zhou, G.; Zheng, J.; Zhang, M.; Xing, L.; Liu, T. Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—Subtropical area for example. *IEEE J. Sel. Top. Appl. Earth Obs. Sens.* **2019**, *13*, 113–128. [CrossRef]
55. Wang, Z.Y.; Zuo, R.G.; Dong, Y.N. Mapping of Himalaya Leucogranites Based on ASTER and Sentinel-2A Datasets Using a Hybrid Method of Metric Learning and Random Forest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1925–1936. [CrossRef]
56. McGovern, A.; John Gagne, D.; Troutman, N.; Brown, R.A.; Basara, J.; Williams, J.K. Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Stat. Anal. Data Mining ASA Data Sci. J.* **2011**, *4*, 407–429. [CrossRef]
57. Herman, G.R.; Schumacher, R.S. Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Weather Rev.* **2018**, *146*, 1571–1600. [CrossRef]