# Data-Based Windstorm Type Identification Algorithm and Extreme Wind Speed Prediction

Wei Cui, A.M.ASCE[1]; Teng Ma, S.M.ASCE[2]; Lin Zhao[3]; and Yaojun Ge[4]

**Abstract:** The extreme wind speed estimation method, which is critical for designing wind load calculation for building structures, should consider windstorm climate types for mixed climates. However, it is very difficult to obtain windstorm climate types from meteorological data records, therefore, it restricts the application of extreme wind speed estimation in mixed climates. This paper first proposes a windstorm type identification algorithm based on a numerical pattern recognition method that utilizes feature extraction and generalization. Subsequently, three sets of model experiments are conducted using data from three meteorological stations on the southeast coast of China from 1990 to 2016, and the prediction of a single station model and a regional model is discussed. The prediction performances of six machine learning algorithms under different experiments are compared. Based on classification results, the extreme wind speeds calculated based on mixed windstorm types are compared with those obtained from conventional methods, and the effects on structural design for different return periods are analyzed. **DOI: 10.1061/(ASCE)ST.1943-541X.0002954.** 

**Author keywords:** Extreme wind speed; Mixed climates; Data-driven method; Pattern recognition; Machine learning.



## Introduction

Wind effects are key factors in structural design, and extreme wind speeds are the starting point. For flexible structures such as long-span bridges, long-span roofs, and high-rise buildings, wind loads are normally the predominant loads. In order to meet both the ultimate safety and performance requirements of wind-resistant structural design, it is necessary to accurately estimate the extreme wind speeds for different recurrence periods.

For significant buildings and infrastructures, it is necessary to estimate the extreme wind speed through probabilistic methods from local wind speed records. The probabilistic method and extreme value theory have already been extensively applied in both building design and structural wind engineering research. The most widely used method of extreme wind speed estimation includes three steps: wind speed sample extraction, probabilistic distribution model selection, and model parameter fitting. There are three major methods for wind speed sample selection: the stage extremum method, the peak-over-threshold method, and the method of independent storm (Palutikof et al. 1999). Through sampling extreme values by extracting peak wind speeds from unit time intervals, the stage extremum method uses the Gumbel distribution (Gumbel 2012) to fit the extreme wind speed distribution and estimates design wind speeds for different return periods. This method is easy to implement but has a low sample utilization rate, so it is suitable for areas with long-term wind speed observation data. It is widely used by design codes in different countries, including the Canadian Building Structure Load Specification (CNBC2015) and the Chinese Building Load Specification (GB50009-2012). In order to overcome the problem of low utilization rate of wind speed observation data, Cook (1982) proposed a new method for selecting wind speed samples, namely, the method of independent storm (MIS). Several subsequent studies (Harris 1999) have shown that if the wind speed observation data comprises continuous data samples and an independent storm segment can be identified, the MIS proposed by Cook has a better data utilization rate than the stage extreme value method. The peak-over-threshold method filters data samples below a predefined threshold from the parent sample as an extreme value sample. In 1975, Pickands (1975) first proposed this extreme value theory based on the generalized Pareto distribution (GPD). In 1996, Simiu and Heckert (1996) used this method to analyze extreme wind speeds in the United States. Mo et al. (2015) rebuilt the extreme wind speed for China using exposure adjusted surface wind observations. In all the previous three methods, it is assumed that the wind speed data are sampled from the same probabilistic distribution, which means wind extremums comes from the same climate pattern, such as monsoons.

However, in mixed climate regions affected by various windstorms, such as China's southeastern coastal region, this assumption is invalid. Samples with the same wind speed values but different windstorm climate types may yield different extreme wind speed estimations. If a single distribution is used for fitting, relatively large deviations will occur. Based on the probability distribution characteristics of wind speeds from mixed climate types,

[1]Assistant Professor, State Key Lab of Disaster Reduction in Civil Engineering, Tongji Univ., Shanghai 200092, China; Key Laboratory of Transport Industry of Wind Resistant Technology for Bridge Structures, Tongji Univ., Shanghai 200092, China. ORCID: https://orcid.org/0000-0001-7489-923X

[2]Ph.D. Candidate, State Key Lab of Disaster Reduction in Civil Engineering, Tongji Univ., Shanghai 200092, China; Key Laboratory of Transport Industry of Wind Resistant Technology for Bridge Structures, Tongji Univ., Shanghai 200092, China.

[3]Professor, State Key Lab of Disaster Reduction in Civil Engineering, Tongji Univ., Shanghai 200092, China; Key Laboratory of Transport Industry of Wind Resistant Technology for Bridge Structures, Tongji Univ., Shanghai 200092, China; State Key Laboratory of Mountain Bridge and Tunnel Engineering, Chongqing Jiaotong Univ., Chongqing 400074, China (corresponding author). Email: zhaolin@tongji.edu.cn

[4]Professor, State Key Lab of Disaster Reduction in Civil Engineering, Tongji Univ., Shanghai 200092, China; Key Laboratory of Transport Industry of Wind Resistant Technology for Bridge Structures, Tongji Univ., Shanghai 200092, China.

Note. This manuscript was submitted on September 1, 2019; approved on October 20, 2020; published online on March 8, 2021. Discussion period open until August 8, 2021; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Structural Engineering***, © ASCE, ISSN 0733-9445.

(Gomes and Vickery 1976; Gomes and Vickery 1978) proposed a composite extreme wind speed analysis method for mixed climates including thunderstorms, hurricanes, and tornadoes, and carried out extreme wind speed estimations for various return periods. Subsequently, Cook et al. (2003) and Cook (2004) improved the method by including the confidence interval analysis of wind speeds with mixed wind climates. The proposed skewed Gaussian distribution is suitable for describing the wind pressure distribution on a building and the error is weakened when fitting the tail part of the Gomes-Vickery method (Cook et al. 2003; Cook 2004). In addition, many other scholars have developed many alternative methods for distinguishing between thunderstorm and nonthunderstorm climates, and combined their probability distributions into mixed distributions for extreme analysis (Riera and Nanni 1989; Twisdale and Vickery 1992; Choi 1999; Choi and Hidayat 2002). Recently, more research proves mixture distribution analysis is suitable for regions with mixed climates (Shin et al. 2016; Ouarda and Charron 2018; Jung and Schindler 2019).

However, most meteorological observatories only record conventional wind climate data such as wind direction and wind speed, and do not record the associated wind climate categories, which makes screening various wind climate types very difficult and tedious. This limitation also constrains research, application, and development of extreme wind speed prediction with mixed climates.

Different methods and benchmark procedures have been proposed to identify different types of windstorms from conventional meteorological data. Riera and Nanni (1989) extracted thunderstorm fragments from conventional meteorological data based on the duration of thunderstorms, the occurrence of lightning, and rainfall (Riera and Nanni 1989). Choi et al. divided windstorms in Singapore into large-scale and small-scale windstorms by visually observing the meteorological data (Choi and Tanurdjaja 2002). Durañona et al. (2007) proposed four indicators to discriminate among extreme winds caused by different wind climates based on extreme wind speed, average wind speed, and wind speed trend (Durañona et al. 2007). Lombardo et al. (2009) achieved an automatic separation of thunderstorms and nonthunderstorms based on the start and end times of thunderstorms, and pointed out that the conditions for ensuring the independence of the two extreme wind speed data points are different (Lombardo et al. 2009). Thunderstorms that need to be separated should have a 4-h sampling rate rather than a 4-day one. However, this method depends on the manual observation records and is confined to the informative windstorm categories provided by the corresponding weather station (Chen and Lombardo 2019).

The preceding methods can achieve rapid identification and extraction of windstorm segments but also have different limitations. The algorithms based on empirical criteria are obtained from raw data analyses (Choi and Tanurdjaja 2002; Durañona et al. 2007). So, it is difficult to conclude a unified criterion through different regions; one method is only applicable to the identification of certain windstorms in a specific area. Thus, it is difficult to propose a general windstorm identification method based on empirical criteria. On the other hand, the history-based method is a general windstorm identification method, but it is limited to the data set that contains information about windstorm categories. At present, there are a large number of weather stations around the world, in which large portions of reported data do not have storm categories, such as NOAA Surface Data Hourly Global (DS3505) (NOAA, n.d.) This article establishes a mapping between historical meteorological data and storm types through machine learning to reduce manual intervention of empirical methods and optimize the data dependency of historical methods. In order to maximize utilization of windstorm data for identification, machine learning algorithms are used to extract windstorm data features that have high correlation with the windstorm type from structured conventional wind climate data, and thus achieve automatic identification of windstorm types. The section "Conclusion" introduces three meteorological observatories on the southeast coast of China and sets up a data set on them. Relying on this data set, three experiments were carried out to discuss the prediction capabilities of different algorithms in storm type recognition. Experiments include: Experiment 1: A single station model predicts unknown data from itself; Experiment 2: The regional model predicts unknown data from known stations; Experiment 3: The regional model predicts unknown data from unknown stations. In each experiment, the performance of six machine learning algorithms is discussed based on the quantified index, and a representative algorithm is used as an example to give its confusion matrix on the test set. Based on identified windstorm data, extreme wind speed calculation results for different return periods of mixed distributions are compared with those obtained from traditional methods to verify the proposed method for extreme wind speed prediction in a mixed climate area. Finally, the applicability and limitations of the automatic identification algorithm for windstorm types are discussed and the main conclusions of this paper are summarized.

## Data Source and Preprocessing

This section introduces the original data sources and filtering, windstorm segmentation, and other data preprocessing steps in detail. Related methods of wind speed time series and surface roughness modification are also introduced in detail in the section "Application Examples" of the actual example because of the strong geographical relevance to meteorological observatories.

### *Meteorological Data Source*

The balance between data acquisition difficulty and data set quality needs to be considered during the data set selection. Generally, there are two sources of data sets for establishing algorithms: private local weather monitoring networks or public global open-source weather databases. Private local data may have a high sampling frequency but with high cost, so it is usually used for regional specific analysis, and can extract a small-scale short duration storm such as downburst (De Gaetano et al. 2014). Open-source databases mostly rely on government weather stations, which have a wider coverage but lower sampling rate. This research tries to propose a storm classification algorithm with wide applicability.

So the original meteorological data originate from the NOAA Global Integrated Surface Database. This database consists of observations from 29,570 meteorological stations around the world, and includes wind direction, wind speed, atmospheric temperature, atmospheric humidity, atmospheric pressure, precipitation, visibility, cloud conditions, and so forth. The region studied in the paper is China's southeastern coast, which is affected by multiple types of wind climates including East Asian monsoons, Northwest Pacific typhoons, and other microscale winds. Therefore, the main windstorm types to be identified in the paper are monsoons, typhoons, and *other* types. The other types are defined as a continuous process with high wind speed with relative short duration and without typical typhoon or monsoon characteristics.

### *Windstorm Segmentation*

The proposed data separation algorithm for extreme wind speeds aims to classify the wind climate type from raw data from meteorological observatories. Before the algorithm is formally presented,

because the raw data includes many variables including wind, precipitation, cloud, temperature, and so on, it is first necessary to select relevant ones. Five time series of data with high correlation with wind damage types, wind direction, wind speed, air pressure, temperature, and precipitation, are extracted from the database as original data.

A windstorm represented by meteorological data is a continuous process rather than a single moment. Taking typhoons and monsoons as examples, these two common wind damage types would normally last several hours or even days. Therefore, before identifying the windstorm type, it is necessary to select multidimensional time series segments with high wind speeds from several years of meteorological records.

Data segments with maximum wind speeds greater than 12 m/s within a set period of time are extracted from original raw data. The wind-range frequency spectrum of the natural wind compiled by Van der Hoven (1957) shows that the macrometeorological peak is at a center period of about 4 days, which means a wind process in a natural weather system seldom lasts longer than 96 h (Van der Hoven 1957).

Therefore, we chose a time series with a span of 96 h as the initial windstorm segment with a time step with maximum speed at the center of the 96-h interval. The extracted initial time series data fragments covered all windstorm durations. However, normally some redundant low speeds records are also included because windstorms always last less than 96 h.

In order to avoid the influence of low-speed data on the wind damage classification model, the heuristic segmentation algorithm (BG algorithm) proposed in Bernaola-Galván and Ivanov et al. (2001) is employed to remove it from the initial wind damage fragments (Bernaola-Galván et al. 2001).

This algorithm recursively divides the time series into subsequences. For a time-series $x(t)$ consisting of $N$ points, the statistical value $T(i)$ of the t-test can be used to quantify the mean difference between the two parts of the point $i$

$$T(i) = \left| \frac{\mu_l(i) - \mu_r(i)}{s_D(i)} \right| \tag{1}$$

where $\mu_l(i)$ and $\mu_r(i)$ are the average values of subsequence on the left (or right) of point $i$. And $s_D(i)$ means the combined deviation of the point $i$.

Test statistical values of each point in the time series are calculated by using Eq. (1). The larger $T$ is, the larger difference between the two parts of the left and right points is, and the more likely it is the sequence division point. Calculate the statistical significance $P(T_{\max})$ of the maximum value $T_{\max}$ in $T(t)$

$$P(T_{\max}) = \text{Prob}(T \le T_{\max}) \tag{2}$$

$P(T_{\max})$ indicates the probability that the $T$ value is less than or equal to $T_{\max}$ in the random process. To quantify the screening criteria, two parameters are usually set: the critical probability $P_0$ and the minimum segmentation scale $l_0$. Each subsequence has two recursive terminating conditions: that the statistical significance of the maximum test statistic value is less than critical value $P_0$ and the length of the subsequence is less than the minimum segmentation scale $l_0$.

Through this operation, the initial windstorm segment can be divided into multiple subsegments, and the subsegment including the maximum wind speeds at the middle time point is taken as the final windstorm segment. Considering the duration of the windstorm and the actual segmentation effect, we take $P_0 = 0.7$ and $l_0 = 8$ to segment the initial windstorm segment. In this case, $l_0 = 8$ means that the windstorm segment lasts at least 24 h so that the segment could contain the basic information of windstorm, especially a short duration storm.

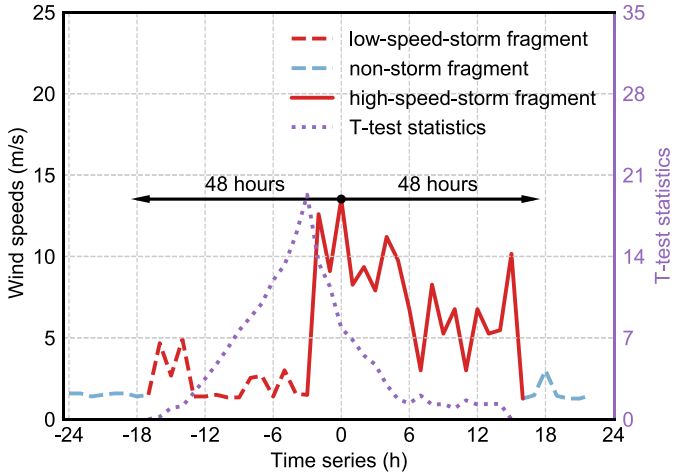


**Fig. 1.** Windstorm fragment screening and segmentation diagram.

Fig. 1 shows an example of the windstorm process. The example wind speed data started from January 9, 2010, at Zhoushan, China. The first segmentation divides the original time series into a dashed line (nonstorm segment) and a red line (storm segment) based on maximum wind speed at center point. Then, using the BG algorithm, high- and low-wind–speed subsequences are identified. The test statistical values are shown as a dotted line and the BG split point is at −3 h, which is the maximum test statistic. The red line is divided into a solid line and a dashed line by BG split point. In this case, the storm segment only has one recursion. Finally, the solid line is used to present this whole storm process.

### *Windstorm Sample Correlation*

An important precondition of extreme analysis is that the data are statistically independent, so each windstorm fragment can only be represented by one wind speed. Each windstorm fragment obtained through the proposed algorithm consists of the maximum wind speed occurring every 3 h. The windstorm usually lasts from several hours to several days, so generally several extreme wind speeds can be obtained from each windstorm fragment. Therefore, in order to ensure the independence of extreme wind speed samples, an appropriate method is required to extract the maximum wind speed from each windstorm and eliminate other wind speeds related to the same windstorm. Simiu and Heckert (1996) believed that sample independence can be guaranteed by the length of the time interval between adjacent extreme sample points, that is, the time interval should be greater than the duration of typical windstorms.

The method for constructing extreme data sets also uses the method of choosing a time interval that ensures the independence of extreme data sets. In the windstorm filtering and segmenting method introduced in the second section, the method of peak-over-threshold is used to preliminarily filter, which ensures that the interval between the maximum wind speed points of the windstorm segments are over 96 h. Segmenting by the BG method ensures that each windstorm fragment belongs to an independent windstorm system. Therefore, each windstorm is represented by the maximum wind speed selected from it. That is to say, an independent wind speed sample with reduced statistical dependence containing information of types is constructed.

## Establishment of Windstorm Identification Algorithms

In the data science field, machine learning algorithms can find regulations just from data sets, and then make predictions. This is the major difference between machine learning data-based models and traditional knowledge-based models. The two common machine learning algorithm categories are supervised learning and unsupervised learning depending on whether the data set contains classification label information. As an initial study on wind climate classification, the proposed machine learning algorithm is a supervised learning algorithm, which means models should comprise labeled data before implementation of the proposed learning model.

### *Feature Extraction*

Typically, the classification of time series is based on the features rather than the actual data value in time series (Bishop 2006). The classifier algorithm is very sensitive to data errors, because data at each time step contributes equally to the classification results, and errors accumulated through the whole time series have a relatively big impact on the model's accuracy (Witten et al. 2016). In addition, in model training and identification, data vectors are required to have the same length. However, windstorm data fragments to be classified have different time lengths because of the BG algorithm. Therefore, an appropriate windstorm feature extraction method is required.

In order to distinguish windstorm types from meteorological data, wind speed, wind direction, atmosphere pressure, air temperature, and precipitation are used by the feature extraction method, which is based on statistical results. According to (Xi et al. 2006), the segmented low-frequency multidimensional time series data can be represented by eight common statistical features for each time series data: mean value $\mu$, standard deviation $\sigma$, skewness $\lambda$, kurtosis $\kappa$, maximum value, minimum value, range value, and median value. Skewness and kurtosis contain information on the shape of the distribution of the time-series values. More precisely, skewness characterizes the degree of asymmetry of values around the mean value. Kurtosis measures the relative peakedness or aptness of the value distribution relative to a normal distribution.

The feature extraction methods for windstorm data in this paper can be separated into three types: first-order features, second-order features, and environmental features (Nanopoulos et al. 2001). First-order features are based on the actual values of the series $x(t)$ and second-order features are based on the differences between nearby values $x'(t)$ that contain the original series varying trends, and can also be used to filter noise. Environmental features include geographic location and windstorm occurrence months, which are not included in original data, but are helpful for model classification.

### *Data Set Establishment*

In this paper, the data set of the windstorm identification algorithm is constructed by a manual classification method. According to the characteristics of wind speed, wind direction, rainfall, temperature, air pressure, and typhoon history database, windstorm data fragments are classified manually (Cook et al. 2003). According to the geographical climate features, monsoons and typhoons are the usual windstorms in the coastal area of southeastern China. The feature of a monsoon windstorm is that the wind direction fluctuates around a certain value in a small variation range (Fig. 2), while the feature of a typhoon windstorm is that before the typhoon lands, air pressure decreases and wind speed increases, and a large amount of precipitation is normally recorded concurrently (Fig. 3).

After the typhoon's passing, air pressure rises and wind deceases. After observation, some storm samples still have high wind speeds but no obvious typhoon monsoon characteristics. Considering that the sampling frequency of NOAA data is not enough to further distinguish small-scale storms, this article classifies such storms as other types (Fig. 4).

In order to reduce the probability of mislabeling and omission of typhoon manual annotation, the typhoon database is used for cross-validation. If a segment of a windstorm does not appear in the typhoon database, the windstorm belongs to another type.

In addition to the label of the data set, imbalance data set is another important problem in machine learning modeling of this study. In terms of the frequency of occurrence, the monsoon is the main windstorm in southeastern China coastal area. Therefore, the original windstorm fragment data set must be unbalanced, and typhoon data is much less than the remaining other two types (Sun et al. 2009). In the field of machine learning, several well-established solutions of imbalanced sample numbers are developed at both data and algorithm levels. In this article, a data-level approach is chosen to address the imbalance because of its effectiveness and ease to be implemented.

Methodologies at the data level include many different forms of resampling, including oversampling, undersampling, and combination method. The resampling method, synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002), is used in the paper, which is based on the random oversampling algorithm, which is employed in this study, because it adopts a simple copying strategy to increase a small number of samples. The idea of the SMOTE algorithm is to synthesize new minority class samples. The synthesis strategy is to randomly select a sample B from its nearest neighbor for each minority class sample A, and then randomly select a point on the line between A and B as the newly synthesized minority class sample. The effect of the oversampling method will be further described in the section "Conclusion."

### *Machine Learning Algorithm*

There are many classic machine learning classification algorithms, algorithm variants, and combination algorithms. Many algorithm variants and combination algorithms are proposed to improve the accuracy of a specific task (Tulyakov et al. 2008). Because the specific algorithms for storm classification is not available, this article will start from the classic algorithms of different classification principles to find suitable algorithms.

In this paper, six commonly used machine learning algorithms, k-Nearest Neighbor (KNN), Naive Bays (NB) (Domingos and Pazzani 1997), Support Vector Machine (SVM) (Cortes and Vapnik 1995), Gradient Boosting (GBDT) (Friedman 2001), Random Forest (RF) (Breiman 2001) and Logistic Regression (LR) (Press and Wilson 1978) are selected for fitting tests to determine suitable algorithm models for learning windstorm data fragment information.

Table 1 provides the chosen algorithms in this paper that belong to different classification principles, including logic, perception, statistic, instance, and margin (Kotsiantis et al. 2007). In the section of "Application Examples," this article will describe in detail the accuracy and generalization ability of the preceding algorithms in storm identification in order to screen out the suitable algorithm.

For a given supervised machine learning problem, a key question in comparing them is to determine whether one algorithm is fundamentally superior to another, or the former outperforms the latter just because hyperparameters have been better tuned for the former (Luo 2016). So, in this article, we tend to achieve the *optimal state* of each

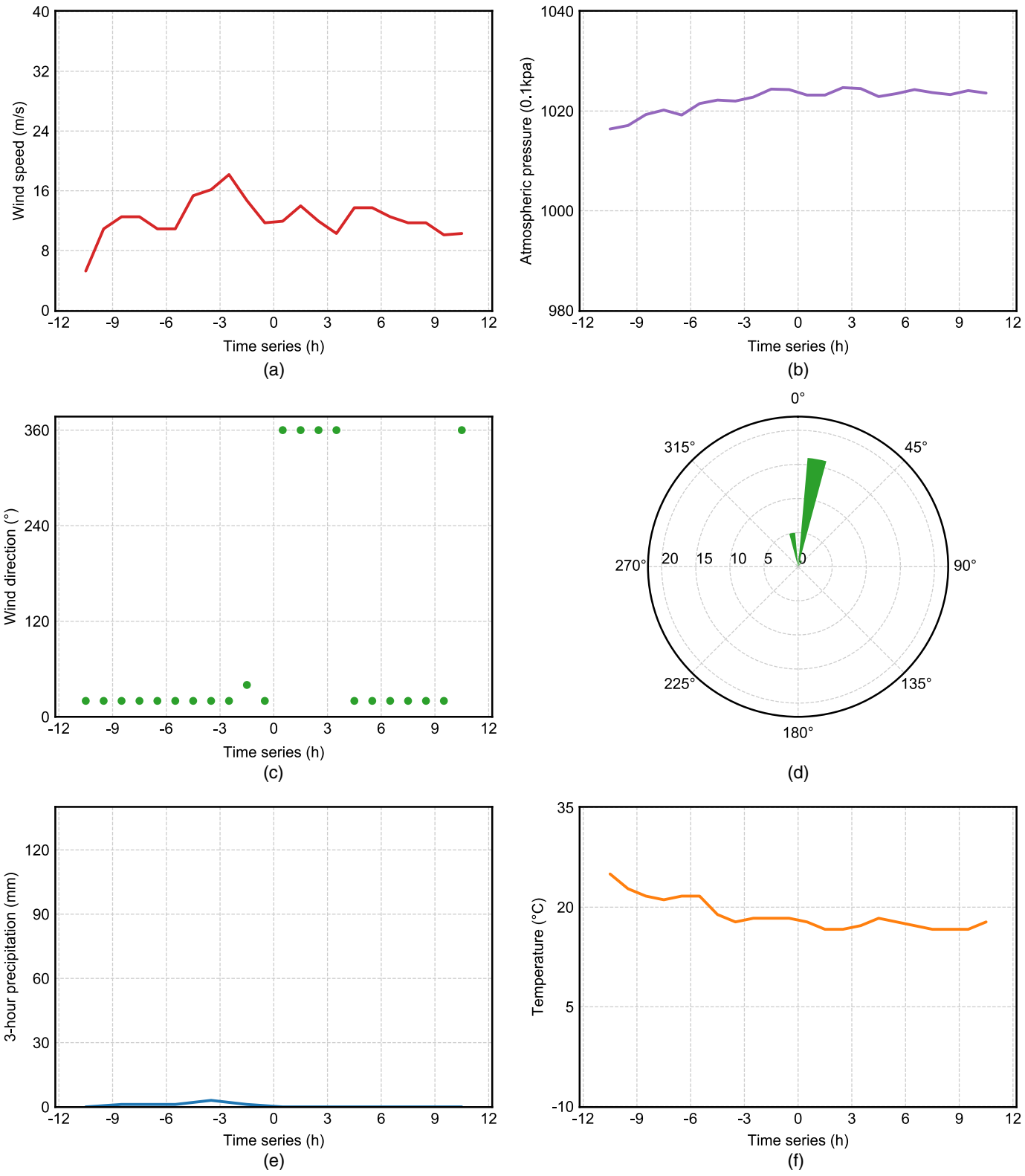


**Fig. 2.** Monsoon-type windstorm data chart starting from October 15, 1999, in Dachen Island meteorological station: (a) wind speed time-history; (b) atmosphere pressure time-history; (c) wind direction time-history; (d) polar histogram of the wind direction; (e) precipitation time-history; and (f) temperature time-history.

algorithm fairly, which means that each algorithm finds out the best hyperparameter group in the same data set. The specific steps are as follows. The data set is divided into three parts according to 6: 2: 2, 60% is the training set to train the model parameters, 20% is the verification set to iteratively optimize the hyperparameters, and another 20% is the test set to examine the model performance. The iterative optimization comparison of algorithm hyperparameters is implemented using the scikit-learn machine learning module (Pedregosa 2011). In this way, the optimal hyperparameters of the six machine learning models in this data set can be obtained.

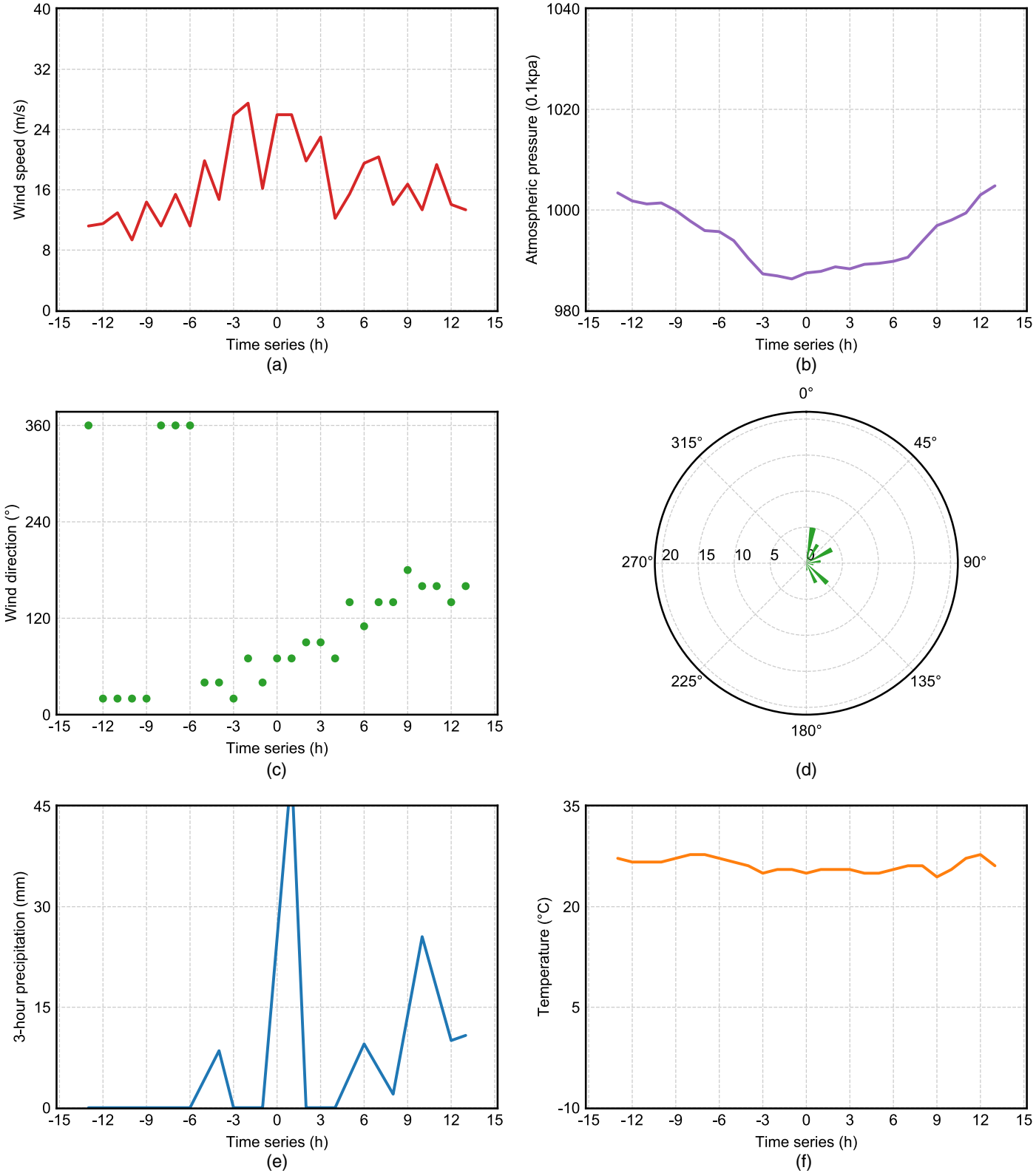


**Fig. 3.** Typhoon Winnie (1997) windstorm data chart starting from August 17, 1997, in Dinghai meteorological station: (a) wind speed time-history; (b) atmospheric pressure time-history; (c) wind direction time-history; (d) polar histogram of the wind direction; (e) precipitation time-history; and (f) temperature time-history.

### Classifiers' Evaluation Metrics and Method

In order to screen out the best machine learning algorithms, it is necessary to introduce some quantitative model evaluation metrics and clarify the evaluation method in this paper.

The concepts of confusion matrix, precision, recall, and average accuracy are introduced to quantify the evaluation performance basically. The classification results can be divided into four categories according to the model's prediction

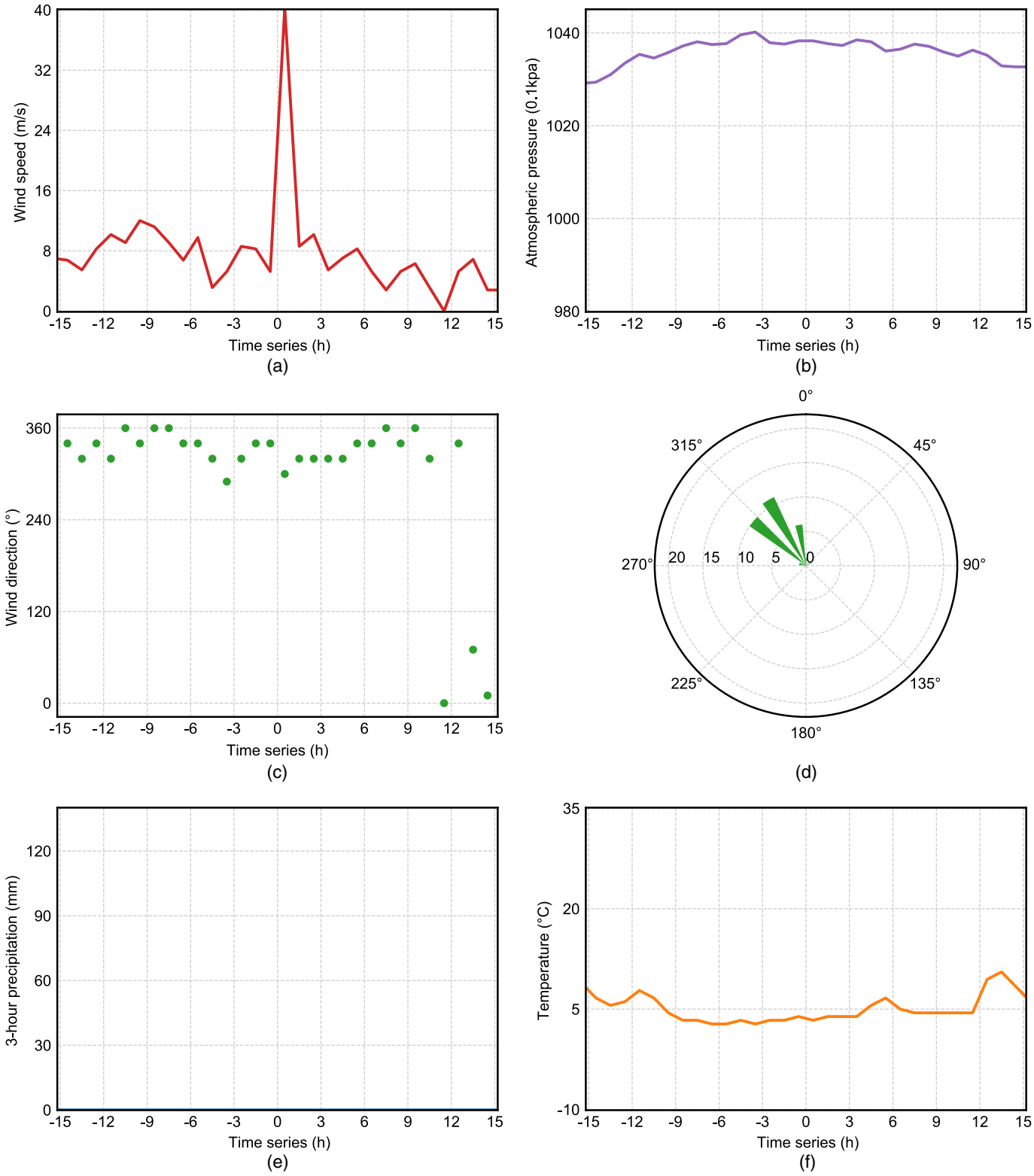


**Fig. 4.** Other-type windstorm data chart starting from December 23, 1995, in Dinghai meteorological station: (a) wind speed time-history; (b) atmospheric pressure time-history; (c) wind direction time-history; (d) polar histogram of the wind direction; (e) precipitation time-history; and (f) temperature time-history.

results and the actual types of windstorm data, as shown in Fig. 5.

In Fig. 5, predicted value means windstorm category is judged by the model through windstorm characteristics and actual value means manually labeled windstorm category. The true positive (TP) samples represent the number of samples that are predicted to be positive and actually labeled as positive in all data. The false negative (FN) samples represent the number of examples that are predicted to be false and actually labeled as positive in all data. The true negative (TN) samples represent the number of examples that

**Table 1.** Classification principles corresponding to machine learning algorithms

| Algorithm | Classification principle |
|---|---|
| k-Nearest neighbor (KNN) | Instance |
| Naive Bayes (NB) | Statistic |
| Support vector machine (SVM) | Margin |
| Gradient boosting (GBDT) | Perception |
| Random forest (RF) | Logic |
| Logistic regression (LR) | Perception |

| | | Actual Value (as manually labelled) | | Index Formula |
|---|---|---|---|---|
| | | positives | negatives | $TP\ rate = \frac{TP}{TP+TN}$ |
| Predicted Value (predicted by model) | positives | TP true positive | FP false positive | $FP\ rate = \frac{TP}{FP+FN}$ |
| | | | | $Precision = \frac{TP}{TP+FP}$ |
| | negatives | FN true negative | TN true negative | $Recall = \frac{TP}{TP+FN}$ |

**Fig. 5.** Second classification problem confusion matrix.

are predicted to be positive and actually labeled as false in all data. The true negative (TN) samples represent the number of examples that are predicted to be false and actually labeled as false in all data. On this basis, performance measurement indexes such as *Precision* and *Recall* are defined. Precision represents the proportion of real cases to all positive predictions, which is equivalent to the degree of accuracy of the measurement model identifying positive samples. Recall represents the proportion of true positive cases to all actual positive cases, which is equivalent to the degree of coverage of the measurement model identifying positive samples.

The storm classification task has the characteristics of unbalanced samples. Although the oversampling method is used to balance the training set when preparing the data set, the test set is still unbalanced, so only accuracy metric is not sufficient to fully evaluate the model performance. Next, some additional quantitative metrics will be introduced for this case, including F1-score and Matthew's correlation coefficient(MCC).

F1-score is the harmonic mean of precision and recall, calculated by Eq. (3). It conveys the balance between the precision and the recall. This article is faced with a multiclassification problem, but F1-score is oriented toward the two-classification problem. Therefore, when using the F1-score evaluation model, the multiclassification problem is decomposed into multiple binary classification problems, which can clearly show the classification effect of each category

$$F1 \cdot score = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

In addition to evaluating the classification effect of each category, a metric is needed to evaluate the model to deal with the imbalance of the sample. In fact, the receiver operating the characteristic (ROC) curve (Bradley 1997) is a good indicator to avoid evaluation bias caused by sample imbalance. The problem is that drawing the ROC curve requires the training model to be sensitive to thresholds. But only the logistic regression and support vector machine are classified by threshold in the test algorithm of this paper. So the ROC curve had to be abandoned, and another metric called Matthew's correlation coefficient (MCC) was used. As Eq. (4) shows, the calculation of the MCC metric uses the four quantities (TP, TN, FP, and FN), which gives a better summary of the performance of classification algorithms. MCC takes values in the interval $(-1, 1)$, with 1 showing a complete agreement, $-1$ a complete disagreement, and 0 showing that the prediction was uncorrelated with the ground truth. Previous studies have proved that it is more reasonable as an evaluation index of the sample imbalanced classifier compared to F1-score (Boughorbel et al. 2017)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(FP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

So, in the next section, this article will use the accuracy, F1-score, and MCC as model performance evaluation metrics.

## Application Examples

This section utilizes data from three meteorological observatories on the southeast coast of China as example data for a series of machine learning experiments; three specific situations are discussed. The first is a single station model, the purpose of which is to use historical data from one station to build a model to predict future data. The second is a regional model, which uses the data from two stations to train a model and predicts the unknown data of the two stations. The third is to use the trained regional model to predict unknown data from the new station. The flow chart of the windstorm classification algorithm experiment is shown in Fig. 6. Each experiment will evaluate the pros and cons of the six machine learning algorithms according to the metrics mentioned in the section on "Classifiers' Evaluation Metrics and Method," and compare the performance of the model on the test set before and after oversampling.

### *Introduction of Meteorological Stations*

Three meteorological stations, Dinghai, Dachen Island, and Shengzhou, are employed in this study. Fig. 7 shows the geographical location and terrain around each station. They are all located near the coastline of the East China Sea and are within 200 km of each other. It can be considered that they are all affected by the same climate pattern, and have the same wind environment. However, the geomorphological profiles around them are different. Dinghai Station is located in a coastal city and has many buildings and urban infrastructures. Dachen Island Station is an off-shore island and is surrounded by the ocean. Shengzhou Station is an inland city about 70 km from the coast. The process of one weather system can affect the preceding three stations with close spatial distance. The different geographic properties of the three stations ensure that their wind speed data is sufficiently disturbed. So, the three stations in the southeast of China were used to show that the method could break through the wind speed data disturbance caused by terrain and actually learn the characteristics of storms along the southeast coast of China from data.

### *Data Set Construction*

The original meteorological data from Dinghai, Dachen Island, and Shengzhou stations all have 27 years of records from 1990 to 2016, with total valid data of 78,105, 77,972, and 71,696 data points, respectively.

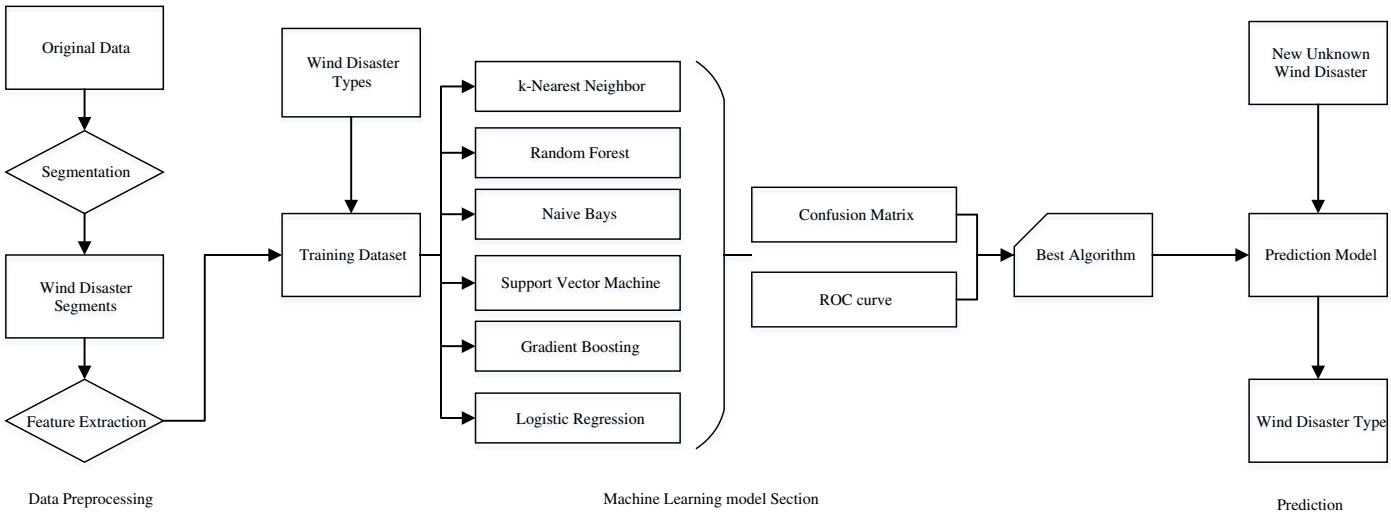


**Fig. 6.** Windstorm classification algorithm experiment flow chart.

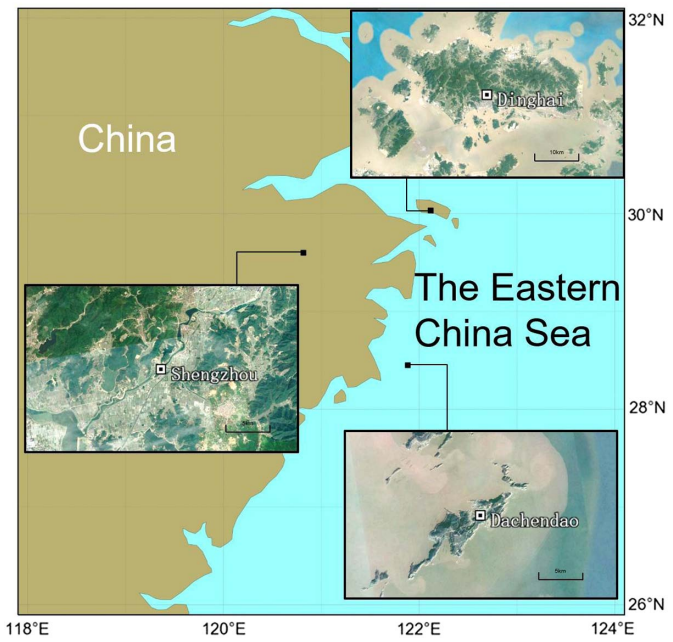


**Fig. 7.** Geographical observation station geographical locations and topographic map. (Map data © 2020 Google, Image by Google, TerraMetrics CNES/ Airbus.)

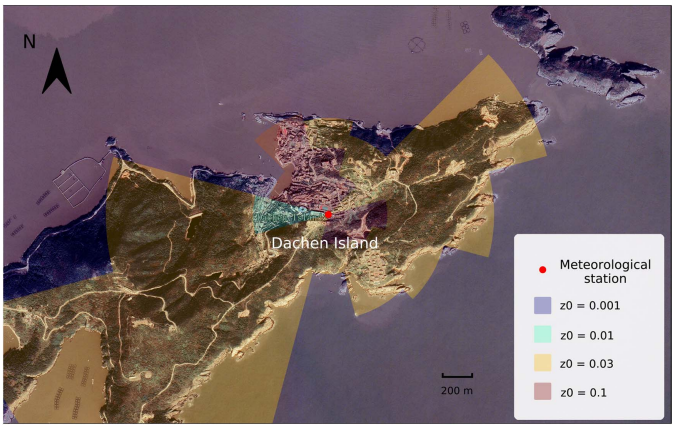


**Fig. 8.** Roughness sectors map around Dachen Island meteorological station. (Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.)

Before the extraction of windstorm fragments, in order to consider terrain interference effects around each station, ground roughness modification to the original wind speed records will be performed. All wind speeds from the three stations should be transformed to uniform ground roughness: 10 m above open terrain ($z_0 = 0.03$ m).

In this paper, the procedure given by ESDU 84011 (ESDU 2004) was used to calculate the ground roughness at the station and the roughness of the terrain over which wind has passed to reach the station (Panofsky and Townsend 1964). The area around the station was divided evenly into 12 sectors in order to consider different wind direction. Moreover, each sector was sorted to different roughness by their urban density and geomorphic features, which were observed on the satellite map. Fig. 8 shows *the roughness map* of Dachen Island as an example of the roughness area segmentation. Distances that terrain roughness extends upwind of site obtained from the roughness map is the input, and the wind speed transforming factors of different wind directions can be obtained by using the program. The results of the three stations are given in Table 2.

After getting the time-histories of wind speed data with standardized ground roughness, the threshold is set to extract the windstorm data fragments from the original data. The wind speed threshold is set to 12 m/s (Simiu and Heckert 1996) to preprocess the data according to the method of the section "Data Source and Preprocessing," and 678, 730, and 289 windstorm fragments are obtained for Dinghai, Dachen Island, and Shengzhou stations. After manually labeling the windstorm types, the three types, typhoon, monsoon, and other, histograms for each station are derived as shown in Fig. 9.

Fig. 9 clearly shows the sample imbalance mentioned in the section "Data Set Establishment." This article uses SMOTE to oversample the data from three stations. In order to more intuitively describe the sample distribution changes caused by oversampling, a principal component analysis (PCA) is used to

**Table 2.** Transforming factor of roughness of different wind directions for Dachen Island, Dinghai, and Shengzhou meteorological stations' data

| Station name | Wind direction | Correction factor | Wind direction | Correction factor | Wind direction | Correction factor |
|---|---|---|---|---|---|---|
| Dachen Island | 30 | 1.029 | 150 | 1.035 | 270 | 0.943 |
| | 60 | 1.093 | 180 | 1.024 | 300 | 1.035 |
| | 90 | 1.052 | 210 | 1.012 | 330 | 1.087 |
| | 120 | 1.903 | 240 | 1.018 | 360 | 1.058 |
| Dinhai | 30 | 0.895 | 150 | 0.900 | 270 | 0.820 |
| | 60 | 0.950 | 180 | 0.890 | 300 | 0.900 |
| | 90 | 0.915 | 210 | 0.880 | 330 | 0.945 |
| | 120 | 0.905 | 240 | 0.885 | 360 | 0.920 |
| Shengzhou | 30 | 0.895 | 150 | 0.900 | 270 | 0.820 |
| | 60 | 0.950 | 180 | 0.890 | 300 | 0.900 |
| | 90 | 0.915 | 210 | 0.880 | 330 | 0.945 |
| | 120 | 0.905 | 240 | 0.885 | 360 | 0.920 |

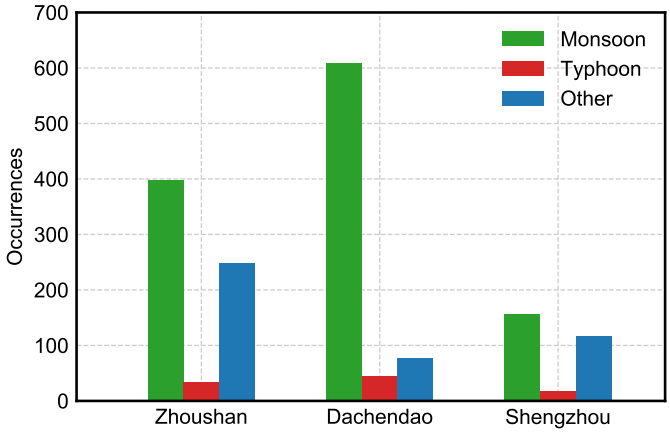


**Fig. 9.** Statistical histogram of windstorm data fragments of example station.

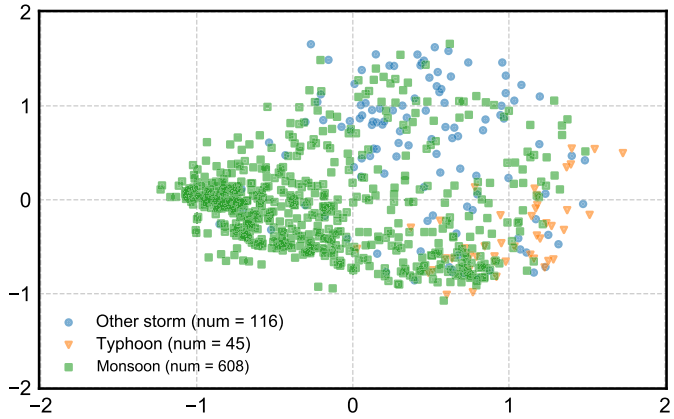


**Fig. 10.** Dimensionality reduction visualization of raw data set of Dachen Island meteorological observatory.

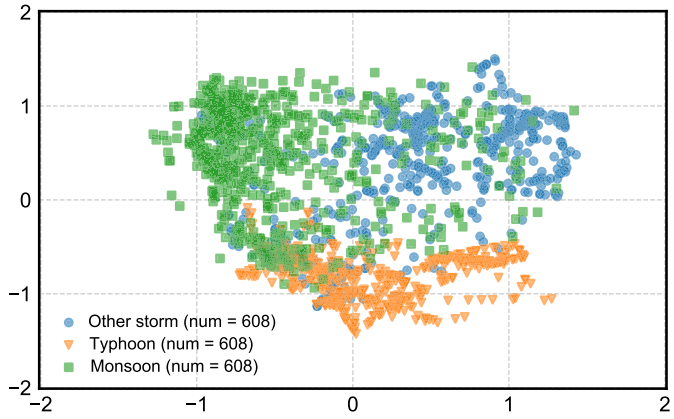


**Fig. 11.** Dimensionality reduction visualization of oversampling data set of Dachen Island meteorological observatory.

reduce the data dimension from 82 to 2 dimensions and retain most of the information (Wold et al. 1987). Figs. 10 and 11 show the data distribution after PCA dimensionality reduction. It can be seen that data oversampling has a greater impact on the sample space. The clustering tendency of samples after oversampling is more obvious. The reason that the cluster boundary is not obvious is that some information will still be lost during the data dimension reduction.

### *Experiment 1: A Single Station Model Predict Unknown Data from the Same Station Data*

This is the most basic case: When the unknown storm data of a weather station needs to be classified, the simplest method is to use the historical data of this station for training. This issue is discussed in this experiment by using data from Dachen Island. The training and validation sets used in the experiment are balanced data sets that have been oversampled by SMOTE, and the test sets are real storm data samples. The results of Experiment 1 are given in Table 3. From the perspective of accuracy and MCC, all six machine learning algorithms have good performance. However, only the F1-score of SVM and LR on three types of storms is higher than 0.6. And the MCC of the SVM is the highest, which means that the SVM has a more balanced confusion matrix.

From the overall effect evaluation, SVM has the best performance among the six algorithms. Fig. 12 shows that the misclassification of the SVM algorithm in the test set mainly appears in the two categories of monsoon and other storms.

In addition to the performance of the algorithm, the effect of oversampling on the model is also worth considering. Based on the preceding experiments, two similar models were established, one was trained and hyperparameters were optimized using raw data, and the other using oversampled equalization data. The performance

**Table 3.** Performance table of six algorithms in Experiment 1 (The number above is the best single performance of each algorithm, and the number below is the worst single performance of current algorithm in Fig. 12)

| Algorithm | F1_score of monsoon | F1_score of other storm | F1_score of typhoon | Accuracy | MCC |
|---|---|---|---|---|---|
| Naive Bayes (NB) | 0.867 | 0.557 | 0.667 | 0.792 | 0.553 |
| k-Nearest neighbor (KNN) | 0.873 | 0.552 | 0.762 | 0.805 | 0.562 |
| Logistic regression (LR) | 0.878 | 0.607 | 0.727 | 0.818 | 0.587 |
| Random forest (RF) | 0.902 | 0.500 | 0.875 | 0.855 | 0.523 |
| Gradient boosting (GBDT) | 0.914 | 0.578 | 0.778 | 0.857 | 0.586 |
| Support vector machine (SVM) | 0.917 | 0.625 | 0.778 | 0.864 | 0.624 |

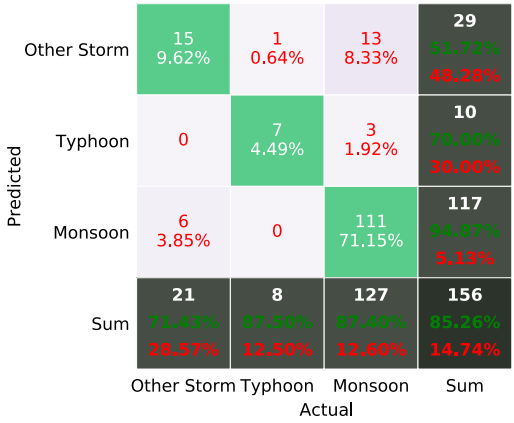


**Fig. 12.** Confusion matrix of SVM algorithm in Experiment 1.

comparison of the two models on the same test set are given in Table 4. Table 4 shows that the oversampling balanced data set does not have an absolute positive impact on model performance. However, oversampling has an impact on the classification boundaries of the model learning and prevents the boundaries from moving closer to the majority to some extent.

### *Experiment 2: Regional Model Predicts Unknown Data from Known Station*

On the basis of a single station model, it is more meaningful and difficult to establish a regional model, which means mixing data from different weather stations in the same area to build a classification model. It should be noted that the training set, the verification set, and the test set are randomly sliced (ratio is 6:2:2) from the station data of Zhoushan and Dachen Island in Experiment 2.

The experimental results are given in Table 5, where SVM, GBDT, and RF all have good performance. Fig. 13 also uses SVM as an example to give a specific confusion matrix.

### *Experiment 3: Regional Model Predicts Unknown Data from Unknown Station*

When a regional model is trained, a new challenge is whether it can predict data accurately from new weather observatories in this area. In this part, the model trained in "Experiment 2" will be retested by a new data set from Shengzhou Station. Shengzhou Station is an unknown station for the model built in "Experiment 2." According to the results in Table 6, it can be clearly seen that the regional model has a prediction performance degradation of unknown station data in the same area. In comparison, SVM performs best in this case. Fig. 14 shows its confusion matrix.

The misclassification of the algorithm mainly manifests itself between other and the monsoon, all algorithms especially show this deficiency, which means that the boundary between the two categories is ambiguous. Improving this problem requires adding more effective features to the model to distinguish them.

The previous experiments show that among the six machine learning algorithms involved in this paper, Support Vector Machine, Gradient Boosting, and Logistic Regression have better performance. The algorithms involved in this article are all classic

**Table 4.** Impact of oversampling on algorithm performance (nonoversampling model is the baseline)

| Algorithm | F1_score of monsoon (%) | F1_score of other storm (%) | F1_score of yyphoon (%) | Accuracy (%) | MCC (%) |
|---|---|---|---|---|---|
| Naive Bayes (NB) | 3.5 | −32.8 | −4.5 | 3.1 | −10.7 |
| k-nearest neighbor (KNN) | 6.7 | −6.7 | 10.5 | 9.7 | 6.9 |
| Logistic regression (LR) | 2.7 | −27.3 | 15.8 | 3.8 | −10.9 |
| Random forest (RF) | 0.8 | −18.4% | 52.4 | 3.1 | −2.1 |
| Gradient boosting (GBDT) | −1.1 | −32.5 | −9.1 | −3.6 | −19.8 |
| Support vector machine (SVM) | 0.2 | −17.4 | 3.8 | −2.2 | −0.5 |

**Table 5.** Performance table of six algorithms in Experiment 2 (The number above is the best single performance of each algorithm, and the number below is the worst single performance of current algorithm in Fig. 13)

| Algorithm | F1_score of monsoon | F1_score of other storm | F1_score of typhoon | Accuracy | MCC |
|---|---|---|---|---|---|
| Naive Bayes (NB) | 0.830 | 0.571 | 0.541 | 0.731 | 0.517 |
| k-Nearest neighbor (KNN) | 0.844 | 0.589 | 0.667 | 0.761 | 0.556 |
| Logistic regression (LR) | 0.878 | 0.657 | 0.730 | 0.811 | 0.620 |
| Random forest (RF) | 0.881 | 0.614 | 0.667 | 0.802 | 0.579 |
| Gradient boosting (GBDT) | 0.892 | 0.634 | 0.722 | 0.823 | 0.609 |
| Support vector machine (SVM) | 0.882 | 0.660 | 0.744 | 0.816 | 0.631 |

algorithms in single classification theory, so this may indicate that the boundary and perceptual algorithms are more suitable for storm classification.

Based on identification results from windstorm classification models, an extreme wind speed for different return periods in a mixed wind climate for structural design is established. Results are compared with those of traditional extreme wind speed estimation methods to verify the influence of mixed climates, which are identified by the proposed model, on the extreme wind speed prediction in mixture climate areas.

### *Extreme Wind Speed Samples for Different Wind Climate Types*

Figs. 15–17 show polar plots of monsoon, typhoon, and other windstorm data obtained from the three stations through the method of construction of extreme samples. The directionalities of the different windstorm types of the three meteorological observatories are shown in the aforementioned figures.

For Dachen Island, the strong wind speeds of monsoon, typhoon, and other storms mostly come from the northeast-southeast region. Typhoon samples account for the majority of high-speed wind samples, while monsoon and other storms account for the majority of low-speed wind samples. Comparison of the numbers of high typhoon wind speed records of three stations shows that the typhoon proportions decrease with distance from station to coastline.

## Extreme Wind Speed Estimation and Comparison

Extreme wind speed samples of typhoon, monsoon, and other windstorms at Dachen Island, Dinghai, and Shengzhou meteorological observatories are constructed using the preceding methods. Since the filtering threshold is set as 12 m/s, except for typhoon wind type, the subsequent calculation of extreme wind speed in this paper is based on the method of peak-over-threshold (POT), i.e., using extreme samples to fit the generalized Pareto distribution (GPD). It should be noted that because the wind speed of most sample points in the parent sample of typhoon exceeds the filtering threshold of 12 m/s, the Gumbel distribution is used to fit the distribution function.

The extremum wind for mixed wind climates is also identified in this paper, and includes all three wind speed types: typhoons, monsoons, and other. Commingled extreme wind speeds are also calculated and compared against China's current building loading code (GB50009-2012). However, from the statistical perspective, a mixture of extreme value samples does not guarantee that the sample points are identically distributed, for example typhoon and monsoon distributions have distinct probabilistic distribution features. Therefore, the extreme value analysis using a mixture of extreme samples is biased for regions influenced by mixed wind climates, such as the three stations analyzed in this paper.

According to Simiu and Yeo (2019), when considering typhoon, monsoon, and other windstorms separately (Simiu and Yeo 2019), the mixture of extreme wind speed cumulative probability distribution function (CDF) is expressed as

$$P(\max(v_M, v_T, v_O) \le V) = P(v_M \le V)P(v_T \le V)P(v_O \le V) \tag{5}$$

In the formula, the left side is the CDF of mixture extreme wind speeds, and the right side is the product of CDFs of typhoons, monsoons, and other windstorm speeds. The premise of Eq. (5) is that typhoons, monsoons, and other windstorms are independent of each other.

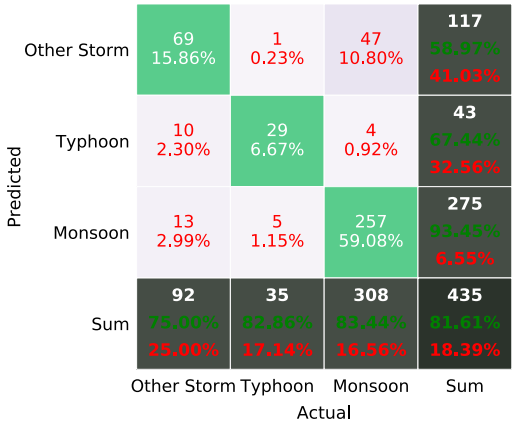


**Fig. 13.** Confusion matrix of SVM algorithm in Experiment 2.

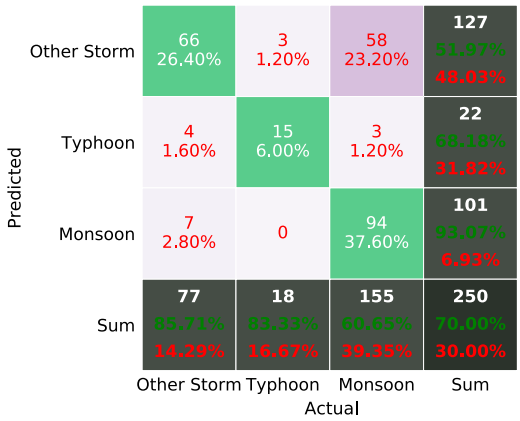


**Fig. 14.** Confusion matrix of SVM algorithm in Experiment 3.

**Table 6.** Performance table of six algorithms in Experiment 3 (The number above is the best single performance of each algorithm, and the number below is the worst single performance of current algorithm in Fig. 14)

| Algorithm | F1_score of monsoon | F1_score of other storm | F1_score of typhoon | Accuracy | MCC |
|---|---|---|---|---|---|
| Naive Bayes (NB) | 0.464 | 0.542 | 0.528 | 0.508 | 0.327 |
| k-nearest neighbor (KNN) | 0.586 | 0.553 | 0.650 | 0.576 | 0.368 |
| Logistic regression (LR) | 0.601 | 0.558 | 0.615 | 0.584 | 0.386 |
| Random forest (RF) | 0.738 | 0.611 | 0.722 | 0.688 | 0.468 |
| Gradient boosting (GBDT) | 0.753 | 0.617 | 0.634 | 0.692 | 0.478 |
| Support vector machine (SVM) | 0.734 | 0.647 | 0.750 | 0.700 | 0.528 |

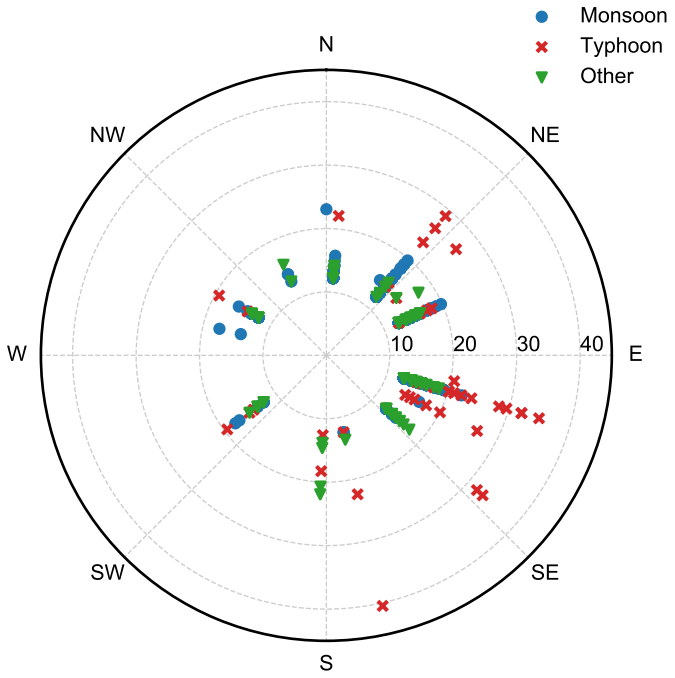


**Fig. 15.** Dachen Island meteorological station extreme wind speed sample wind speed/wind direction polar coordinate map.

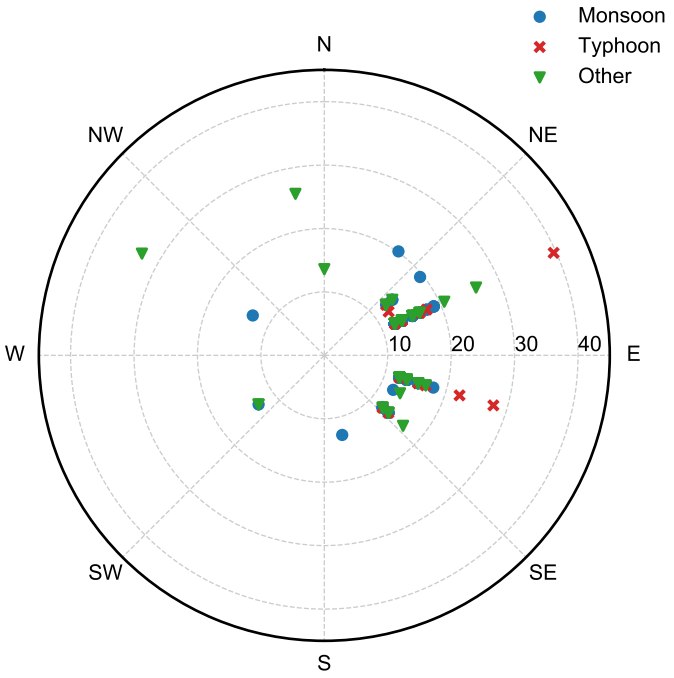


**Fig. 17.** Shengzhou meteorological station extreme wind speed sample wind speed/wind direction polar coordinate map.

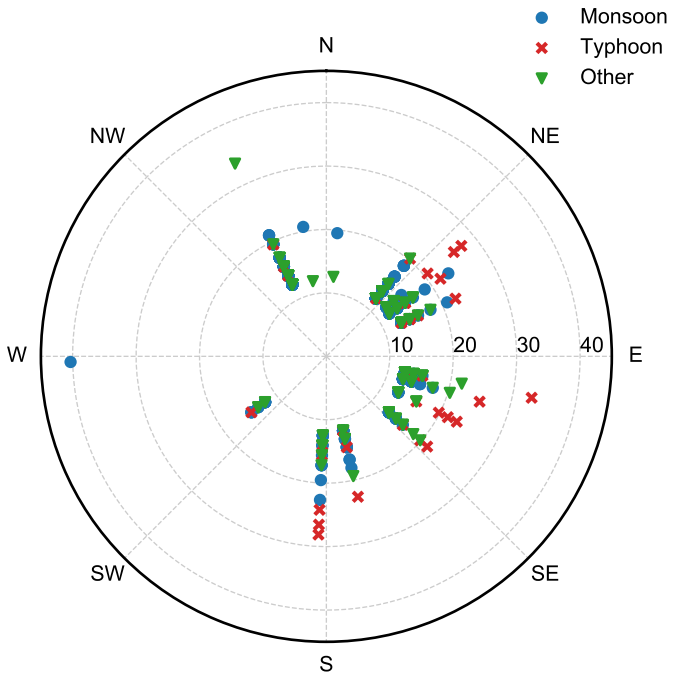


**Fig. 16.** Dinghai meteorological station extreme wind speed sample wind speed/wind direction polar coordinate map.

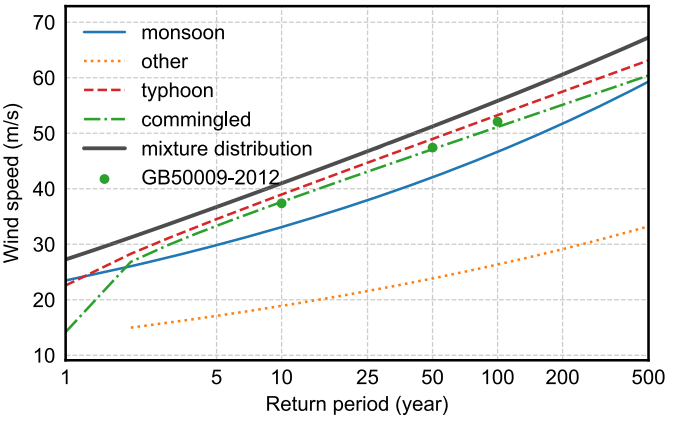


**Fig. 18.** Extreme wind speed-recurrence period curve of Dachen Island meteorological station.

Figs. 18–20 show extreme wind speeds for different return periods obtained by observatories in Dachen Island, Dinghai, and Shengzhou. The solid line and the dotted line are the monsoon and other extreme wind speed curves calculated using the generalized Pareto distribution through POT samples of monsoon and other. The dashed line is the typhoon extreme wind speed curve calculated using the generalized Gumbel distribution through typhoon POT samples. The commingled curve is the dashes-and-doted line shown in Figs. 18–20, which is calculated using the generalized Gumbel distribution through annual extremum samples of all wind speeds. The large solid dots indicate the extreme wind speeds in the Chinese code, which are almost on the dashed and dotted line. The thick black solid line indicates the extreme wind speed from the previous mixture distribution through POT samples of all wind speeds. The curve calculation methods are given in Table 7.

Since the extremum samples are filtered based on the cross-domain method, the relationship between return period and surpass probability is as shown in Eq. (6)

$$P(v_x \geq V) = \frac{1}{T \times N_x} \tag{6}$$

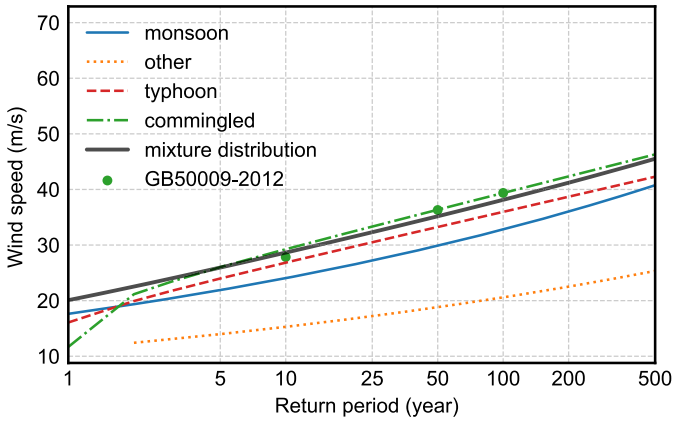


**Fig. 19.** Extreme wind speed-recurrence period curve of Dinghai meteorological station.

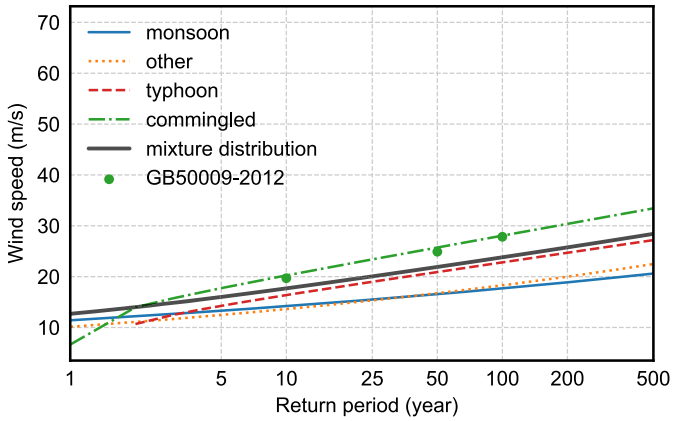


**Fig. 20.** Extreme wind speed-recurrence period curve of Shengzhou meteorological station.

**Table 7.** Extreme wind speed for different return period calculation methods

| Types of extreme wind speed | Sample type | Sample extraction method | Distribution |
|---|---|---|---|
| Monsoon | Monsoon | Peak over threshold | Generalized Pareto distribution |
| Typhoon | Typhoon | Annual extremum | Generalized Gumbel distribution |
| Other | Other | Peak over threshold | Generalized Pareto distribution |
| Commingled | All | Annual extremum | Generalized Gumbel distribution |
| Mixture | All (classified) | Peak over threshold | Mixture distribution |

In the formula, $x$ represents the type of windstorm or extreme wind speed samples, and $N_x$ represents the average number of times such samples occur every year.

In this paper, the curves of extreme wind speed–return period for three types of windstorms are compared. Figs. 18–20 show that in coastal areas, typhoons control extreme wind speeds with high return periods, while monsoon and other storms control extreme wind speeds with low return periods.

From Dachen Island in Fig. 18 to Shengzhou in Fig. 20, typhoons normally dominate the windstorm type for return period intervals longer than 5 years. However, for short return period intervals, the inland areas are affected by monsoons rather than typhoons. Comparing the *commingled* line and the *mixture distribution* line, it can be seen that the extreme wind speed results of the mixture distribution prediction are higher for offshore places, while inland stations are the opposite. This also shows that the mixture distribution better considers the effect of typhoons on extreme wind speeds.

## Conclusion

This paper first proposed the concept of data-driven windstorm type identification for efficient identification of wind weather data from conventional weather stations that is faster and more accurate than traditional experience-driven classification methods. The performance of six classic machine learning algorithms on storm recognition was compared through three machine learning experiments. The results show that Support Vector Machines, Gradient Boosting, and Logistic Regression have better performance. Second, this paper proposed a weather time-course data preprocessing, feature extraction method, and evaluation metrics suitable for windstorm data classification and developed a windstorm identification model based on the extracted features. It is proved through experiments that both the single station model and the regional model have good prediction capabilities, which shows that the machine learning-based storm classification method is a general method for establishing regional storm classifiers. Finally, based on the results of windstorm classification, this paper established separated extreme wind speed samples for different windstorm types. The difference between the extreme and maximum wind speeds of mixed and unclassified mixed extreme samples was compared, and the mixed distribution obtained a conservative solution for the full return period. Therefore, the windstorm type has great significance in estimating the extreme wind speed of different return periods, making the structural design performance of the building more reasonable.

## Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

## Acknowledgments

The authors gratefully acknowledge the support of National Natural Science Foundation of China (52008314, 51678451), Shanghai Pujiang Plan (No. 19PJ1409800), National Key research and Development Program of China (2018YFC0809600, 2018YFC0809604). Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the preceding agencies.

## References

Bernaola-Galván, P., P. C. Ivanov, L. A. Nunes Amaral, and H. E. Stanley. 2001. "Scale invariance in the nonstationarity of human heart rate." *Phys. Rev. Lett.* 87 (16): 168105. https://doi.org/10.1103/PhysRevLett.87.168105.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. Berlin: Springer.

Boughorbel, S., F. Jarray, and M. El-Anbari. 2017. "Optimal classifier for imbalanced data using Matthews correlation coefficient metric." *PLoS One* 12 (6): e0177678. https://doi.org/10.1371/journal.pone.0177678.
Bradley, A. P. 1997. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognit.* 30 (7): 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.
Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.
Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic minority over-sampling technique." *J. Artif. Intell. Res.* 16: 321–357. https://doi.org/10.1613/jair.953.
Chen, G. Z., and F. T. Lombardo. 2019. "A revised automated classification method of thunderstorm and non-thunderstorm wind data based on a neural network." In *Proc., 15th Int. Conf. on Wind Engineering (ICWE)*. Beijing: International Associations for Wind Engineering.
Choi, E. C. 1999. "Extreme wind characteristics over Singapore—An area in the equatorial belt." *J. Wind Eng. Ind. Aerodyn.* 83 (1–3): 61–69. https://doi.org/10.1016/S0167-6105(99)00061-6.
Choi, E. C., and F. A. Hidayat. 2002. "Gust factors for thunderstorm and non-thunderstorm winds." *J. Wind Eng. Ind. Aerodyn.* 90 (12–15): 1683–1696. https://doi.org/10.1016/S0167-6105(02)00279-9.
Choi, E. C. C., and A. Tanurdjaja. 2002. "Extreme wind studies in Singapore. An area with mixed weather system." *J. Wind Eng. Ind. Aerodyn.* 90 (12): 1611–1630. https://doi.org/10.1016/S0167-6105(02)00274-X.
Cook, N. J. 1982. "Towards better estimation of extreme winds." *J. Wind Eng. Ind. Aerodyn.* 9 (3): 295–323.
Cook, N. J. 2004. "Confidence limits for extreme wind speeds in mixed climates." *J. Wind Eng. Ind. Aerodyn.* 92 (1): 41–51. https://doi.org/10.1016/j.jweia.2003.09.037.
Cook, N. J., R. Ian Harris, and R. Whiting. 2003. "Extreme wind speeds in mixed climates revisited." *J. Wind Eng. Ind. Aerodyn.* 91 (3): 403–422. https://doi.org/10.1016/S0167-6105(02)00397-5.
Cortes, C., and V. Vapnik. 1995. "Support-vector networks." *Mach. Learn.* 20 (3): 273–297. https://doi.org/10.1023/A:1022627411411.
De Gaetano, P., M. P. Repetto, T. Repetto, and G. Solari. 2014. "Separation and classification of extreme wind events from anemometric records." *J. Wind Eng. Ind. Aerodyn.* 126: 132–143.
Domingos, P., and M. Pazzani. 1997. "On the optimality of the simple Bayesian classifier under zero-one loss." *Mach. Learn.* 29 (2–3): 103–130. https://doi.org/10.1023/A:1007413511361.
Durañona, V., M. Sterling, and C. J. Baker. 2007. "An analysis of extreme non-synoptic winds." *J. Wind Eng. Ind. Aerodyn.* 95 (9): 1007–1027. https://doi.org/10.1016/j.jweia.2007.01.014.
ESDU (Engineering Sciences Data Unit). 2004. *Wind speed profiles over terrain with roughness changes*. ESDU 84011. London: ESDU.
Friedman, J. H. 2001. "Greedy function approximation: A gradient boosting machine." *Ann. Stat.* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.
Gomes, L., and B. J. Vickery. 1976. "On thunderstorm wind gusts in Australia." *Inst. Eng. Aust. Civ. Eng. Trans. CE* 18 (2): 33–39.
Gomes, L., and B. J. Vickery. 1978. "Extreme wind speeds in mixed wind climates." *J. Wind Eng. Ind. Aerodyn.* 2 (4): 331–344. https://doi.org/10.1016/0167-6105(78)90018-1.
Gumbel, E. J. 2012. *Statistics of extremes*. New York: Columbia University Press.
Harris, R. 1999. "Improvements to the method of independent storms." *J. Wind Eng. Ind. Aerodyn.* 80 (1–2): 1–30. https://doi.org/10.1016/S0167-6105(98)00123-8.
Jung, Christopher, and Dirk Schindler. 2019. "Changing wind speed distributions under future global climate." *Energy Convers. Manage.* 198 (Oct): 111841. https://doi.org/10.1016/j.enconman.2019.111841.
Kotsiantis, S. B., I. Zaharakis, and P. Pintelas. 2007. "Supervised machine learning: A review of classification techniques." *Emerging Artif. Intell. Appl. Comput. Eng.* 160 (1): 3–24.
Lombardo, F. T., J. A. Main, and E. Simiu. 2009. "Automated extraction and classification of thunderstorm and non-thunderstorm wind data for extreme-value analysis." *J. Wind Eng. Ind. Aerodyn.* 97 (3): 120–131. https://doi.org/10.1016/j.jweia.2009.03.001.
Luo, G. 2016. "A review of automatic selection methods for machine learning algorithms and hyper-parameter values." *Network Model. Anal. Health Inf. Bioinf.* 5 (1): 18. https://doi.org/10.1007/s13721-016-0125-6.
Mo, H. M., H. P. Hong, and F. Fan. 2015. "Estimating the extreme wind speed for regions in China using surface wind observations and reanalysis data." *J. Wind Eng. Ind. Aerodyn.* 143 (Aug): 19–33. https://doi.org/10.1016/j.jweia.2015.04.005.
Nanopoulos, A., R. Alcock, and Y. Manolopoulos. 2001. "Feature-based classification of time-series data." *Int. J. Comput. Res.* 10 (3): 49–61. https://doi.org/10.5555/766914.766918.
NOAA (National Oceanic and Atmospheric Administration). n.d.. "Integrated surface hourly data base (3505)." Accessed February 4, 2021. https://www7.ncdc.noaa.gov/CDO/cdopoemain.cmd?datasetabbv=DS3505&countryabbv=&georegionabbv=&resolution=40.
Ouarda, T. B., and C. Charron. 2018. "On the mixture of wind speed distribution in a Nordic region." *Energy Convers. Manage.* 174 (Oct): 33–44. https://doi.org/10.1016/j.enconman.2018.08.007.
Palutikof, J., B. Brabson, D. Lister, and S. Adcock. 1999. "A review of methods to calculate extreme wind speeds." *Meteorol. Appl.* 6(2): 119–132.
Panofsky, H. A., and A. A. Townsend. 1964. "Change of terrain roughness and the wind profile." *Q. J. R. Meteorol. Soc.* 90 (384): 147–155. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49709038404.
Pedregosa, F., et al. 2011. "Scikit-learn: Machine learning in python." *J. Mach. Learn. Res.* 12 (Oct): 2825–2830.
Pickands, J., III. 1975. "Statistical inference using extreme order statistics." *Ann. Stat.* 3 (1): 119–131.
Press, S. J., and S. Wilson. 1978. "Choosing between logistic regression and discriminant analysis." *J. Am. Stat. Assoc.* 73 (364): 699–705. https://doi.org/10.1080/01621459.1978.10480080.
Riera, J., and L. Nanni. 1989. "Pilot study of extreme wind velocities in a mixed climate considering wind orientation." *J. Wind Eng. Ind. Aerodyn.* 32 (1–2): 11–20. https://doi.org/10.1016/0167-6105(89)90012-3.
Shin, J. Y., T. B. Ouarda, and T. Lee. 2016. "Heterogeneous mixture distributions for modeling wind speed, application to the UAE." *Renewable Energy* 91 (Jun): 40–52. https://doi.org/10.1016/j.renene.2016.01.041.
Simiu, E., and N. Heckert. 1996. "Extreme wind distribution tails: A 'peaks over threshold' approach." *J. Struct. Eng.* 122 (5): 539–547. https://doi.org/10.1061/(ASCE)0733-9445(1996)122:5(539).
Simiu, E., and D. Yeo. 2019. *Wind effects on structures: Modern structural design for wind*. Hoboken, NJ: Wiley-Blackwell.
Sun, Y., A. K. C. Wong, and M. S. Kamel. 2009. "Classification of imbalanced data: A review." *Int. J. Pattern Recognit. Artif. Intell.* 23 (4): 687–719. https://doi.org/10.1142/S0218001409007326.
Tulyakov, S., S. Jaeger, V. Govindaraju, and D. Doermann. 2008. "Review of classifier combination methods." In *Machine learning in document analysis and recognition*, 361–386. Berlin: Springer.
Twisdale, L. A., and P. J. Vickery. 1992. "Research on thunderstorm wind design parameters." *J. Wind Eng. Ind. Aerodyn.* 41 (1–3): 545–556. https://doi.org/10.1016/0167-6105(92)90461-I.
Van der Hoven, I. 1957. "Power spectrum of horizontal wind speed in the frequency range from 0.0007 to 900 cycles per hour." *J. Meteorol.* 14 (2): 160–164. https://doi.org/10.1175/1520-0469(1957)014<0160:PSOHWS>2.0.CO;2.
Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.
Wold, S., K. Esbensen, and P. Geladi. 1987. "Principal component analysis." *Chemometr. Intell. Lab. Syst.* 2 (1–3): 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.
Xi, X., E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana. 2006. "Fast time series classification using numerosity reduction." In *Proc., 23rd Int. Conf. on Machine Learning*. New York: Association for Computing Machinery.