Whirlwind Classification with Imbalanced Upper Air Data Handling using SMOTE Algorithm and SVM Classifier

D C R Novitasari¹, A Z Foeady¹, R Nariswari², A H Asyhar¹, N Ulinnuha¹, Y Farida¹, D R Santı³, Ilham⁴ and F Setiawan⁵

¹ Department of Mathematics, UIN Sunan Ampel Surabaya, Surabaya, Indonesia
² Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
³ Department of Psychology, UIN Sunan Ampel Surabaya, Surabaya, Indonesia
⁴ Department of Information System, UIN Sunan Ampel Surabaya, Surabaya, Indonesia
⁵ Forecaster and Analyst, Meteorological Climatological and Geophysics Agency, Surabaya, Indonesia

Abstract. Whirlwind is a natural disaster that often occurs and is difficult to predict from some time before. Early identification is needed to prevent a lot of casualties and losses. Whirlwind caused by instability in the atmosphere. Instability in the atmosphere usually occurs at the beginning of the day and the whirlwind can be identified based on the upper air parameter which can represent atmospheric instability. The purpose of this research is to optimize SVM classification with SMOTE algorithm to handling problems in imbalanced data and this research can minimize casualties and losses or also be a breakthrough for disaster-prone areas to be given early warning. The process for identifying whirlwind has several stages, namely pre-processing, imbalanced data handling, and classification. Pre-processing is normalized data. Whirlwind data has a classification problem, namely inter-class data that is not balanced so it needs to be corrected using the SMOTE Algorithm. Research on the identification of whirlwind using a combination of SMOTE-SVM produces the best accuracy is 98.8 %. When compared to data without the SMOTE algorithm the results obtained are better if the SMOTE method is applied. The specificity value is also better when given the SMOTE method. Based on these results it can be concluded that SMOTE can overcome the problem of imbalanced data in the upper air data by increasing the value of the classification of the whirlwind.

1. Introduction
The stability of atmospheric conditions is one of the main factors that affect components in the climate such as rainfall, temperature, or wind speed. When atmospheric changes have a large and irregular scale, they can cause a variety of natural disasters such as whirlwind and storms [1]. A whirlwind is a wind that rotates very strong, has a minimum diameter of 100 meters, and extends to the ground from the cumulonimbus cloud center. The whirlwind wind speed has an average of 30 to 70 mph, even a whirlwind can reach wind speeds of 200 mph [2]. Whirlwind was rated number one in Indonesia in 2018 as the most common disaster. The number of whirlwind disasters in Indonesia in 2018 reached 804 incidents with a rising trend from 2008. The whirlwind incident data was obtained from the BNPB.

Based on the trend of continually rising whirlwind events, identification of whirlwind is needed because the damage and casualties can be anticipated. The identification process can be through prediction, classification, and mathematical modelling based on its spread [3]. In this case, the identification process was carried out by means of classification to indicate the presence of a whirlwind.
or not based on existing parameters. The parameters taken as data include wind speed, temperature, air pressure, and atmospheric parameters (lifted index, severe weather threat index, convective available potential energy, k-index, convective inhibition, total totals, and showalter index) because these parameters can be benchmarks for changes in atmospheric conditions that begin to be unstable [4].

The problem that often occurs in the classification process is the unbalanced amount of data between classes. This problem causes a decrease in the classification results so it needs to be improved on the data. Therefore, the initial process for identifying whirlwind is in handling imbalanced data. To overcome the problem of data imbalance can use 2 methods, which are under-sampling and over-sampling [5]. The undersampling method is a method for handling imbalanced data by reducing the majority data and the majority of the data is the same as the minority data [6]. Methods other than undersampling is the oversampling method, which is a method to overcome data imbalances by adding minority data or equalizing the number of minority data with majority data [7]. In this research, the method used is oversampling because this method can overcome the problem of data imbalance without damaging existing trend data. One method of oversampling is Synthetic Minority Over-sampling Technique (SMOTE). The SMOTE method is a method for balancing the amount of data by adding minority data by using synthesis data and it used in the process of training and testing data [8].

The application of the SMOTE method was also carried out by Piyasak Cheatrakul et al. by using unbalanced data from diabetics, user credit data, SPECT heart data, and Haberman's survival data. In the research, data comparisons between the two classes reached 80% compared to 20%. By using the Complementary Neural Network classification method and using the SMOTE method to overcome data imbalances, the results are very satisfying [9]. Research on SMOTE has also been done by Jae-Hyun Seo and Yong-Hyuk Kim to detect instructions on magma. The research is using SMOTE to balance the amount of data and methods of Support Vector Regression to create a model to predict the intrusion of magma [10]. The whirlwind data is also unbalanced because there is rarely a whirlwind, and the SMOTE method will be used to overcome the problem of imbalanced data [11].

The SMOTE process will produce more balanced data, then the process of classifying data can be carried out into two classes, which are non-whirlwind and whirlwind. The classification method chosen is the Support Vector Machine (SVM) classifier because it can be fairly fast and has good results for separating data with binary classes or two classes when compared to other classification methods. SVM is a classification method that focuses on forming rules to separate two classes called hyperplane [12].

SVM was applied in the text classification. In this research, classified data is document data downloaded in Indonesia. Several labels are given, namely, economic, security, education, health, sports, and politics. In this research, several classification methods were used, which is Naïve Bayes, C.45, K-NN, and SVM. However, the best results in the research are on the SVM method, with an accuracy rate of 90% [13]. Amin Ul Haq et. al. [14] also researched heart disease detection by comparing several classification methods. The best results in this research were obtained from the SVM method using the RBF kernel with an accuracy of 86%. Another research is the combination of SMOTE and SVM because data that will be used is an unbalanced classification on medical imaging data. He compared the 3 classification methods previously carried out by the SMOTE process. These methods are Bayesian, K-NN, and SVM. The best results in this research are in the SVM classification, with an accuracy rate of 80%. The conclusion that can be drawn is that the SMOTE method matches the SVM classification [15].

The SVM method is chosen as a method that will be used to classify data combined with the SMOTE method to improve unbalanced data from the whirlwind. With a combination of these methods, it is expected to help in identifying the occurrence of whirlwind properly and accurately. Identification is carried out at 00 Z to find out the atmospheric conditions and it can conclude that a whirlwind event will occur or not. Z-time or zulu time usually used on meteorological aspect which is maps, radar, and satellite image.
2. Preliminaries

2.1. Whirlwind
Whirlwind is a strong spinning wind that reaches the ground from the cumulonimbus cloud center. The whirlwind has a minimum diameter of 100 meters with an average speed of 30-70 mph with a duration of about 1-10 minutes. Whirlwinds often occur during the day, evening, and evenings during the transition season [16]. The process of whirlwinds usually begins with cumulonimbus clouds, which are dark gray because the atmosphere in the cloud is unstable and causes hot winds and cold winds to hit. This cloud has the potential to cause large winds and heavy rain [17].

2.2. Atmospheric Sounding Data
Atmospheric Sounding Data or commonly known as upper air data, is data taken based on a certain height to determine the level of stability in the atmosphere. Atmospheric Sounding Data has several parameters, such as lifted index, severe weather threat index, convective available potential energy, k-index, convective inhibition, total totals, and showalter index.

(a) Convective Available Potential Energy (CAPE): is a strong indicator of the atmosphere, which indicates the stability of the atmosphere. CAPE is very sensitive in the presence of stability in the atmosphere, such as mixing foreign substances, freezing, and others [18].
(b) Lifted Index (LI): is a parameter used to measure the stability between surface temperature and temperature in the atmosphere. LI is not affected by wind, and in determining it is very easy. Generally, the value of LI, which is less than -9 is an unstable condition; the value between -6 and -9 is said to be unstable, and between -3 and -6 is said to be quite unstable [19].
(c) K-Index (KI): useful for identifying convective rain, heavy rain, or storms. The KI parameter can be seen through the air temperature at altitudes of 500, 700, and 850 hPa and see the dew point at an altitude of 700 and 850 hPa. When the value of KI exceeds 40, it can be concluded that there will be a thunderstorm accompanied by heavy rain in the area [20].
(d) Totals Index (TT): is a parameter that is useful for identifying bad weather. The TT index value is seen through the temperature and dew point at an altitude of 850 and 500 hPa. When the TT value is more than 44, it can be said to be bad weather, but when the TT value is more than 55 a big storm will likely occur until the whirlwind [21].
(e) Severe Weather Threat Index (SWEAT): is a parameter to measure bad weather, but the difference is in the SWEAT parameter considering the speed and direction of the wind to see the potential for bad weather in an area [21].
(f) Showalter Index (SI): is a parameter used to identify atmospheric stability at an altitude of 850 hPa. SI parameters are seen from temperature and dew point at an altitude of 850 hPa [22].
(g) Convective Inhibition (CIN): the value that indicates the amount of energy needed to overcome the negative energy given from the earth's surface to the atmosphere. The atmosphere that has a higher CIN value is said to have a more stable atmospheric level [23].

2.3. Synthetic Minority Over-Sampling Technique (SMOTE)
SMOTE is an oversampling method to overcome imbalances in data by making synthetic or artificial data on minority data and next the amount of data is balanced with majority data. This method can increase the classification to divide into two classes and the accuracy in testing will be greater [8][15]. The steps in using the SMOTE algorithm as follow:

- Identify data that is a minority. The simulation can be seen in Figure 1, with red color is minority data.
2. Find the closest neighbors between minority classes using the K-Nearest Neighbors (KNN) algorithm. This process can be seen in Figure 2 by connecting a line.

2. Calculate the distance between vectors in each neighbor using the Euclidean distance, which can be seen in Equation (1).

\[ \text{Distance} = \sqrt{\sum (x_i - x_j)^2} \]  

where \( x_i \) is the first data and \( x_j \) is the second data.

2. Calculate the data transfer by multiplying the distance with a random value between 0 and 1, which can be seen in Equation (2).

\[ P(i,j) = \text{rand}(0,1) \times \text{Distance}(i,j) \]  

Where \( P(i,j) \) is the transition value.

2. Determine artificial data by adding the main vector with the vector transfer that can be seen in Equation (3). The formation of artificial data can be seen in Figure 3.

\[ X' = x(i,j) + P(i,j) \]  

Where \( X' \) is a new data synthesized.

2.4. Support Vector Machine (SVM)

The SVM method is often referred to as a machine learning method that can divide two classes automatically by using a dividing line called a hyperplane [24]. In SVM, the data closest to the hyperplane is also called the support vector, which functions as a barrier if there is new data that appears
for classification. The initial idea of SVM was obtained from determining the best hyperplane that would divide into two classes. After the hyperplane is formed, the data will be divided into two classes, which is +1 and -1 [25].

The SVM method can classify two types of data sets, such as linear and non-linear data. Based on the target results of the classification, SVM is divided into Binary Classification and Multiclass Classification [26]. SVM also has a kernel to transform data to be used in Lagrange equations in the SVM process. There are 4 types of kernels in SVM, such as linear, quadratic, polynomial, and Gaussian. The first process in SVM is to form the kernel matrix using the kernel equation, which can be displayed in equations (4), (5), (6) and (7).

\[ Linear = K(x, y) = x \cdot y \] (4)

\[ Polynomial = K(x, y) = (x \cdot y + c)^d \] (5)

\[ Gaussian = K(x, y) = \exp \left( -\frac{||x - y||^2}{2\sigma^2} \right) \] (6)

\[ Quadratic = K(x, y) = \frac{1}{\sqrt{||x - y||^2 + C^2}} \] (7)

After the kernel matrix is formed, then search for \( \alpha \), bias \((b)\), and also weight \((w)\) values, which will be included in the hyperplane equation. The value of \( \alpha \) is obtained through Equation (10) by substituting equations (8) and (9) with the target \( y \), data \( n \), and the kernel matrix \( K \).

\[ z = y^T \cdot y \] (8)

\[ A = \frac{1}{n} \sum \sum K(i, j) \cdot Z(i, j) \] (9)

\[ a = A \cdot y \] (10)

Find the values of \( w \) and also \( b \) which can be seen in Equations (11) and (12). And the values are entered into the hyperplane equation in Equation (13).

\[ w = A \cdot K^T \] (11)

\[ b = -\frac{1}{2} \left( \sum w \cdot K \right) \] (12)

\[ x \cdot w + b = 0 \] (13)

2.5. Confusion Matrix

The confusion matrix is a method for knowing information that contains actual data and predictive data from a classification system result. In classification, it is expected to be able to classify the data appropriately and it has good results with a small error. Therefore, this method is present to help find out how much success of classification is carried out [27]. There are 3 results that become the benchmark of the confusion matrix, these results are accuracy, specificity, and sensitivity. To get these results, things that need to be identified are True Positive (TP), Positive False (FP), False Negative (FN), and True Negative (TN) [26]. The confusion Matrix table can be seen in Table 1.

| Actual | Classification |
|--------|----------------|
| +      | True Positive (TP) | False Negatives (FN) |
| -      | False Positive (FP) | True Negatives (TN) |

After getting the parameters TP, FN, FP, and TN, then the calculation is done to find the accuracy, specificity, and sensitivity that can be seen in Equations (8), (9), and (10).

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \] (14)
\[
\text{Sensitivity} = \frac{TP}{TP + FN} \\
\text{Spesificity} = \frac{TN}{TN + FP}
\] (15) (16)

3. Research Method

3.1. Types of Research
Research about whirlwind detection using the SMOTE method and SVM method is a quantitative descriptive research because the data used is numerical statistical data. This research was applied with the aim of helping to provide early warning about the occurrence of a whirlwind and that can be anticipated and reduce casualties.

Figure 4. Flowchart Detection Whirlwind

3.2. Data Collection and Analysis
Weather parameter data such as wind speed, temperature, and humidity were obtained from Meteorological Climatological and Geophysics Agency (BMKG), while atmospheric parameter data obtained from Wyoming University Weather Web from January 2018 to March 2019 with 455 data with detailed data 21 data and data are normal or no whirlwind events of 434 data. Validation of whirlwind events was obtained from BNPB. Then the data will be divided into testing data and training data with a comparison of training data and testing data at 60%: 40%, 70%: 30%, and 80%: 20%.

3.3. Testing Data Evaluation
The first step is pre-processing by normalizing the data. The normalization of this data can also be useful to simplify the calculation process because the data value is not so large. The data that will be used as
research has a significant amount of unbalanced data between whirlwind disaster data and normal or no disaster data and the process to overcome the problem needs has been finished using the SMOTE method which will increase the number of minority data by making synthesis data. The balanced data is then carried out by the classification process using the SVM method. The main purpose of the classification process is to divide the data into two classes to determine whether the data will occur or not. After the results of the classification are obtained, a calculation process is needed to determine the accuracy obtained and it can be known what percentage of this research was successful. This research has a goal to be achieved, namely to get the results of a good whirlwind classification. To achieve these objectives, we need steps that are structured. The processing steps can be seen through the flow diagram in Figure 4.

a. Pre-processing
The whirlwind data obtained has a large distance in several parameters. Pre-processing is needed to make the data do not have too large a distance. The processed pre-processing is to use data normalization. The normalization of data will produce data with a range of 0 to 1. Changing values to 0 to 1 is useful to facilitate the calculation to speed up the process of detection of the whirlwind.

b. SMOTE Algorithm
The whirlwind data obtained has a small number of events compared to non-whirlwind condition data. This causes data imbalance between disaster and non-disaster classes. To overcome this problem, the SMOTE method is needed, where this method works by increasing the number of minority data using the synthesis data formation. This synthesis data will later serve as additional data in the training process and testing process. The steps of this algorithm are as follows:

1) Identifying minority data (data that has the fewest labels) and will balance the amount of data.
2) Find the neighbors from data using the KNN method.
3) Calculate the distance of each neighboring data using Equation (3).
4) Calculate the value of the displacement of the matrix using Equation (4).
5) Making synthesis data to repair the number of minority data to the same amount as the majority data using Equation 5.

c. SVM Classifier
The results of the SMOTE algorithm data are then divided into training data and test data with a predetermined comparison system. Then determine the kernel and calculate the kernel matrix according to the kernel selected in Equations (4), (5), (6), and (7). Next calculate the values of $\alpha$, $b$, and $w$ using Equations (10), (11), and (12). From the results of the values $b$ and $w$, that substitute to the hyperplane equation is made in Equation (13) which will be used as the optimum model. The testing phase in SVM uses the test data and the optimum model that has been obtained at the training stage. The results of SVM are then used to calculate the results of SVM using a confusion matrix to determine accuracy, specificity, and sensitivity in equations (14), (15), and (16).

4. Result and Discussion
Several processes are needed to detect whirlwind. The process includes pre-processing, imbalanced data handling, and classification. Pre-processing is used to simplify and speed up the calculation process and improve data that has a large distance between parameters. Handling imbalanced data is used to obtain balanced data and overcome the overfitting condition at the classification stage.

Classification is used to divide into 2 classes, namely whirlwind disaster, and no whirlwind disaster. Samples of whirlwind parameters such as wind speed (WS), air humidity (RH), temperature (T), lifted index (LI), severe weather threat index (SWEAT), convective available potential energy (CAPE), k-index (KI), convective inhibition (CIN), total totals (TT), and showalter index (SI) can be shown in
Table 2 for data that has no whirlwind or normal occurrence and Table 3 for data that occurs in whirlwind.

### Table 2. Sample of Non-Whirlwind Data

| WS | T  | RH | CAPE  | LI | KI  | TT  | SWEAT | SI   | CIN  |
|----|----|----|-------|----|-----|-----|-------|------|------|
| 9  | 27.1| 84 | 135.55| -1.27| 32  | 43.7| 189.78| 0.69 | -107.01 |
| 5  | 27.5| 84 | 618.49| -2.12| 35.7| 44.7| 193   | -0.06| -109.52 |
| 6  | 26.8| 84 | 59.98 | -0.84| 35  | 45.6| 209   | -1.27| -184.24 |
| 6  | 26.3| 84 | 243.2 | 1.4  | 36.3| 44.2| 210.81| -0.27| -65.45  |
| 4  | 27.8| 80 | 96.69 | -0.89| 33.8| 43.6| 197.21| 0.87 | -135.98 |

### Table 3. Sample of Whirlwind Data

| WS | T  | RH | CAPE  | LI | KI  | TT  | SWEAT | SI   | CIN  |
|----|----|----|-------|----|-----|-----|-------|------|------|
| 6  | 27.2| 81 | 1340.11| -3.42| 30.3| 42.4| 146.18| 2.61 | -191.61 |
| 6  | 25.6| 89 | 1251.44| -4.99| 37.9| 46  | 195.99| -0.5 | -93.5  |
| 15 | 27.8| 86 | 1741.15| -4.44| 38.4| 45.2| 250.42| -1.12| -34.84 |
| 5  | 28.2| 78 | 1882.24| -4.84| 35.2| 45.4| 210.81| -0.92| -16.47 |
| 6  | 27.7| 81 | 1484.45| -3.05| 37.5| 45.8| 215.6 | -0.75| -60.67 |

A. Preprocessing

The results of the data normalization process can be seen in Table 4 for non-whirlwind data and in Table 5 for whirlwind disaster data. The results of normalization have a more stable value and are expected to get good results with this stable data.

### Table 4. Sample of Non-Whirlwind Normalize Data

| WS | T  | RH | CAPE  | LI | KI  | TT  | SWEAT | SI   | CIN  |
|----|----|----|-------|----|-----|-----|-------|------|------|
| 0.286| 0.476| 0.771| 0.028 | 0.001| 0.443| 0.128| 0.070 | 0.822| 0.984 |
| 0.286| 0.341| 0.857| 0.082 | 0.001| 0.489| 0.128| 0.071 | 0.822| 0.989 |
| 0.143| 0.207| 0.914| 0.057 | 0.001| 0.477| 0.125| 0.073 | 0.824| 0.990 |
| 0.143| 0.354| 0.743| 0.228 | 0.001| 0.494| 0.144| 0.077 | 0.801| 0.995 |
| 0.357| 0.488| 0.657| 0.005 | 0.001| 0.529| 0.120| 0.076 | 0.824| 0.981 |

### Table 5. Sample of Whirlwind Normalize Data

| WS | T  | RH | CAPE  | LI | KI  | TT  | SWEAT | SI   | CIN  |
|----|----|----|-------|----|-----|-----|-------|------|------|
| 0.286| 0.268| 0.714| 0.187 | 0.001| 0.433| 0.122| 0.047 | 0.858| 0.956 |
| 0.286| 0.073| 0.943| 0.175 | 0.001| 0.509| 0.152| 0.070 | 0.809| 0.978 |
| 0.929| 0.341| 0.857| 0.243 | 0.001| 0.514| 0.145| 0.095 | 0.799| 0.992 |
| 0.214| 0.390| 0.629| 0.263 | 0.001| 0.482| 0.147| 0.076 | 0.802| 0.996 |
| 0.286| 0.329| 0.714| 0.207 | 0.001| 0.505| 0.150| 0.079 | 0.805| 0.986 |

B. SMOTE Algorithm

Synthesis data will be formed as much as the majority of the data until the amount of whirlwind data is almost the same as the non-whirlwind event data. Synthesis data will be formed in each parameter used in the research. Generally, synthesis data will be formed around the whirlwind data and it can assist the training process in determining patterns in the SVM model and can help the training process in increasing accuracy. The sample results from the SMOTE process can be seen in Table 6.
In Table 6, it can be seen that the data formed is in the scope of the whirlwind event data. Then the data is added to existing data by labeling whirlwind events data. The results of the SMOTE process add synthesis data with a total of 875 data produced with details of 434 data normal or non-whirlwind and 441 data on whirlwind events.

| Sample of SMOTE Result |
|-------------------------|
| WS | T | RH | CAPE | LI | KI | TT | SWEAT | SI | CIN |
| 0.181 | 0.257 | 0.635 | 0.112 | 0.001 | 0.277 | 0.066 | 0.025 | 0.858 | 0.945 |
| 0.181 | 0.070 | 0.839 | 0.104 | 0.001 | 0.325 | 0.082 | 0.037 | 0.809 | 0.968 |
| 0.590 | 0.327 | 0.762 | 0.145 | 0.001 | 0.328 | 0.079 | 0.050 | 0.799 | 0.981 |
| 0.136 | 0.374 | 0.559 | 0.157 | 0.001 | 0.308 | 0.080 | 0.040 | 0.802 | 0.985 |
| 0.181 | 0.315 | 0.635 | 0.124 | 0.001 | 0.323 | 0.081 | 0.041 | 0.805 | 0.975 |

C. SVM Classifier

The initial data and synthesis results are input from SVM Classifier. The data is then divided into training data and testing data in accordance with comparisons that have been determined in the research method. Then the training process is carried out and the most optimum SVM model is obtained. After the model is found, the next process is the testing process which will be included in the configuration matrix to find out the results. The results of the confusion matrix can be seen in Table 7. The results use a comparison of linear kernels (L), polynomial (P), gaussian (G), and quadratic (Q).

In Table 7, with the distribution of training data and testing data at 60:40, the best results are found in the polynomial, gaussian, and quadratic kernels with an accuracy rate of 98.8%. These results are influenced by 4 data that should not occur whirlwind identified as a whirlwind. It also affects the level of specificity because the specificity seen from non-disaster data are classified non-correctly. The distribution of training data and testing data at 70:30 the best accuracy is in the gaussian and quadratic kernels with a value of 98.4%. This is also because the whirlwind identifies 4 non-whirlwind data. In polynomial kernels, sensitivity has a value of 99.2% because there is one disaster data identified that is not a disaster. The distribution of training data and testing data is 80:20. The best results are seen in the

| Confusion Matrix Result |
|-------------------------|
| Data | Kernel | TP | TN | FP | FN | Ac | Se | Sp |
| 60:40 | L | 165 | 9 | 0 | 176 | 97.4% | 100% | 94.8% |
| | P | 170 | 4 | 0 | 176 | 98.8% | 100% | 97.7% |
| | G | 170 | 4 | 0 | 176 | 98.8% | 100% | 97.7% |
| | Q | 170 | 4 | 0 | 176 | 98.8% | 100% | 97.7% |
| 70:30 | L | 124 | 6 | 0 | 132 | 97.7% | 100% | 95.3% |
| | P | 125 | 5 | 1 | 131 | 97.7% | 99.2% | 96.2% |
| | G | 126 | 4 | 0 | 132 | 98.4% | 100% | 96.9% |
| | Q | 126 | 4 | 0 | 132 | 98.4% | 100% | 96.9% |
| 80:20 | L | 83 | 4 | 0 | 88 | 97.7% | 100% | 95.4% |
| | P | 84 | 4 | 0 | 88 | 98.2% | 100% | 96.5% |
| | G | 84 | 4 | 0 | 88 | 98.2% | 100% | 96.5% |
| | Q | 84 | 4 | 0 | 88 | 98.2% | 100% | 96.5% |

In Table 7, with the distribution of training data and testing data at 60:40, the best results are found in the polynomial, gaussian, and quadratic kernels with an accuracy rate of 98.8%. These results are influenced by 4 data that should not occur whirlwind identified as a whirlwind. It also affects the level of specificity because the specificity seen from non-disaster data are classified non-correctly. The distribution of training data and testing data at 70:30 the best accuracy is in the gaussian and quadratic kernels with a value of 98.4%. This is also because the whirlwind identifies 4 non-whirlwind data. In polynomial kernels, sensitivity has a value of 99.2% because there is one disaster data identified that is not a disaster. The distribution of training data and testing data is 80:20. The best results are seen in the
polynomial, gaussian, and quadratic kernels with an accuracy rate of 98.2%. Based on the best accuracy results from using the SMOTE-SVM, the method will be compared with a process carried out only using the SVM method without handling the imbalanced data problem with the SMOTE algorithm. A comparison of the confusion matrix result can be seen in Table 8.

| SVM Method | Normal | Whirlwind |
|------------|--------|-----------|
| Normal     | 128    | 2         |
| Whirlwind  | 4      | 2         |

| SMOTE-SVM | Normal | Whirlwind |
|-----------|--------|-----------|
| Normal    | 126    | 4         |
| Whirlwind | 0      | 132       |

In Table 8, the SVM method using SMOTE is better than only the SVM method. Data will be more recognized with a large amount of data. Based on that case, adding synthesis data using the SMOTE algorithm can increase the classification result. The Table shows the whirlwind data using the SMOTE-SVM method don't have a misclassification and compared to the SVM method without the SMOTE there are 4 whirlwind data identified as normal weather condition. These results indicate that SMOTE can overcome imbalanced data problems and improve the performance of the SVM method. The result of the comparison SVM method and SMOTE-SVM method can be seen in Table 9.

| Method                  | Ac   | Se   | Sp   |
|-------------------------|------|------|------|
| SVM (kernel = RBF)      | 95.6%| 33.3%| 98.4%|
| SMOTE + SVM (Kernel = RBF) | 98.4%| 100% | 96.9%|

In Table 8, the results using the SMOTE algorithm to overcome imbalanced data and classification without using the SMOTE algorithm obtained better results in the process by using the SMOTE algorithm. It can be seen that the accuracy in SVM without the SMOTE algorithm only produces an accuracy of 95.6%. The SVM method which is preceded by the SMOTE algorithm has better results with an accuracy value is 98.8%. The addition of the SMOTE method can also be used to increase the value of sensitivity to the classification results. The sensitivity value is useful for determining how much the program can identify the occurrence of a whirlwind. The better the sensitivity value, the best results of classification.

5. Conclusion

The results of implementation and trials using SMOTE-SVM that have been carried out to detect whirlwind disasters, it can be concluded that to identify whirlwind disasters, upper air data, and climate data has been applied. Upper air data is atmospheric data and there are have unnatural atmospheric anomalies, it will also cause unnatural values. SMOTE can improve unbalanced data and can improve accuracy because the scale of synthesis data that formed into data that has the same distribution as whirlwind data, and the combination of SMOTE-SVM can produce good accuracy, namely at 60:40 data distribution with the polynomial kernel, gaussian, and quadratic with an accuracy of 98.8%. If this method compared to data without the SMOTE algorithm, the results obtained are better if the SMOTE method is applied. The specificity value is also better when given the SMOTE method. Based on these results, it can be concluded that SMOTE can overcome the problem of imbalanced data in the upper air data by increasing the value of the classification of the whirlwind.

6. References

[1] A. Cahyadi and U. G. Mada, “Peringatan dini puting beliung di daerah istimewa yogyakarta,” 2017.
[2] R. Davies-Jones, R. J. Trapp, and H. B. Bluestein, “Tornadoes and Tornadic Storms,” Severe Convective Storms, Meteorological Monographs, 28. pp. 167–221.
[3] C. A. Doswell, S. J. Weiss, and R. H. Johns, “Tornado forecasting: A review,” pp. 557–571, 2011.
[4] T. B. Trafulis, I. Adrianto, M. B. Richman, and S. Lakshmivarahan, “Machine-learning classifiers for imbalanced tornado data,” Comput. Manag. Sci., vol. 11, no. 4, pp. 403–418, 2014.
[5] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets: A review,” *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.

[6] R. A. Sowah, M. A. Agebure, G. A. Mills, K. M. Koumadi, and S. Y. Fiawoo, “New Cluster Undersampling Technique for Class Imbalance Learning,” no. 3, 2016.

[7] M. S. Santos, P. H. Abreu, P. J. Garcia-Laencina, A. Simão, and A. Carvalho, “A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients,” *J. Biomed. Inform.*, vol. 58, pp. 49–59, 2015.

[8] N. V Chawla, K. W. Bowyer, and L. O. Hall, “SMOTE : Synthetic Minority Over-sampling Technique,” vol. 16, pp. 321–357, 2002.

[9] P. Jeatrakul, K. W. Wong, and C. C. Fung, “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6444 LNCS, no. 2, pp. 152–159, 2010.

[10] J. Seo and Y. Kim, “Machine-Learning Approach to Optimize SMOTE Ratio in Class Imbalance Dataset for Intrusion Detection,” vol. 2018, 2018.

[11] M. Maalouf and T. B. Trafalis, “Rare events and imbalanced datasets: an overview,” *Int. J. Data Mining, Model. Manag.*, vol. 3, no. 4, pp. 375–388, 2011.

[12] D. Fradkin and I. Muchnik, “Support vector machines for classification,” *DIMACS Ser. Discret. Math. Theor. Comput. Sci.*, vol. 70, pp. 13–20, 2006.

[13] F. Wulandini and A. S. Nugroho, “Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases,” 2009.

[14] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. García-Magarinó, “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,” *Mob. Inf. Syst.*, vol. 2018, 2018.

[15] J. Wang, M. Xu, H. Wang, and J. Zhang, “Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding,” vol. 2, no. 10, pp. 1–4, 2006.

[16] S. B. Iryanthony, “Pengembangan Modul Kesiapsiagaan Bencana Angin Puting Beliung Untuk Mahasiswa Pendidikan Geografi Unnes,” vol. 12, no. 2, pp. 143–154, 2015.

[17] F. Rosdiana, “Putting Beliung , Bencana Regional dengan Sebaran Nasional,” 2013.

[18] T. S. Glickman, *Glossary of Meteorology*. Boston, 2000.

[19] J. Galway, “The lifted index as a predictor of latent instability,” *lifted index as a Predict. latent Instab.*, vol. 37, pp. 528–529, 1956.

[20] J. J. George, *Weather Forecasting for Aeronautics*. New York: Academic Press, 1960.

[21] R. C. Miller, “Notes on analysis and severe storm forecasting procedures of the Air Force Global Weather Center, AWS Tech. Report 200,” 1972.

[22] I. Vasilică, L. Sfîcă, L. Apostol, and B. Florentina, “The status of atmospheric instability indices associated with hail events throughout Moldova,” no. 3, pp. 323–331, 2015.

[23] J. Colby and P. Frank, “Convective Inhibition as a Predictor of Convection during AVE-SESAME II,” pp. 2239–2252, 1984.

[24] A. Z. Foeady, D. C. R. Novitasari, A. H. Asyhar, and M. Firmansjah, “Automated Diagnosis System of Diabetic Retinopathy Using GLCM Method and SVM Classifier,” *2018 5th Int. Conf. Electr. Eng. Comput. Sci. Informatics*, pp. 154–160, 2019.

[25] Y. Ahuja and S. Kumar Yadav, “Multiclass Classification and Support Vector Machine,” *Glob. J. Comput. Sci. Technol. Interdiscip.*, vol. 12, no. 11, pp. 14–19, 2012.

[26] J. Vaişnava, S. Ravi, M. Anousouya Devi, and S. Punitha, “Automatic diabetic assessment for diabetic retinopathy using support vector machines,” *Int. J. Control Theory Appl.*, vol. 9, no. 7, pp. 3135–3145, 2016.

[27] S. Visa, B. Ramsay, A. Ralescu, and E. Van Der Knaap, “Confusion matrix-based feature selection,” *CEUR Workshop Proc.*, vol. 710, pp. 120–127, 2011.