Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Weather Classification Based on Hybrid Cloud Image Using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

To cite this article: Yulia Hapsari and Syamsuryadi 2019 *J. Phys.: Conf. Ser.* **1167** 012064

Weather Classification Based on Hybrid Cloud Image Using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

Yulia Hapsari[1*], **Syamsuryadi** [2]

[1]Master of Informatics Engineering, Universitas Sriwijaya, Palembang, 30139, Indonesia
[2]Department of Informatics, Faculty of Computer Science, Universitas Sriwijaya, Palembang, 30139, Indonesia

*E-mail: yuliahapsari@polsri.ac.id

Abstract. Changes in weather and climate conditions have consequences on various sectors of life and greatly affect the activities of human life. Therefore we need a system that can detect weather conditions based on cloud imagery. Finding methods to detect weather conditions at one time with image processing is a new innovation that appears in current weather modeling. This is driven by the high need of various parties to conduct research in detecting a condition carefully and without having to observe it directly. In this study a climate condition classification system will be designed based on cloud imagery using the Hybrid method, namely PCA + LDA. All cloud imagery will be grayscale then feature extraction and cloud classification process using Euclidean Distance. Based on the tests carried out, the system produces an accuracy rate of 96%. The predicted weather conditions are bright, cloudy, and rainy conditions.

1. Introduction

Changes in weather and climate conditions pose a risk if it is not known early. This is driven by different weather conditions between one place and another [1]. Changes in weather and climate conditions have consequences on various sectors of life and greatly affect the activities of human life. Basically natural phenomena are difficult to control except on a small scale. A small phenomenon that occurs is an example when the morning looks bright, but near noon comes heavy rain. Extreme weather was due to vertical cloud expansion and increased rainfall [2].

Based on current geothermal conditions, weather detection is crucial in the application of several scientific disciplines and human activities. In order for this phenomenon to be detected from the very beginning, especially the occurrence of rain which causes extreme weather phenomena, the development of a weather detection system is needed to avoid or minimize the impact of the rain. In addition the weather detection system application can be developed into a short-term prediction system so that it can help BMKG predict daily weather.

The development of information technology today, there are many solutions to the needs faced, including the forecast of weather information. Several articles have discussed the weather forecasting system using the Linear Discriminant Analysis (LDA) method with predicted weather conditions which are sunny, cloudy or rainy by achieving an average system accuracy of 87% [3] and other research on weather detection systems based on cloud imaging using Algorithms. K-Nearest Neighbor (k-nn) with predictable weather conditions is sunny, cloudy, rainy, sunny and rainy night conditions and achieves an average system accuracy of 84.21% [4], but according to the authors the level of accuracy and computation of the system still not maximal. The PCA method has also been used in various fields, including in the financial sector, using PCA in the case of lending by entering around 500 databases from German Bank [5]. And in the field of science to investigate the impact of the application of Principal Component Analysis (PCA) in rainfall clustering on the islands of Java, Bali and Lombok [6], PCA is used as a stage in the process of clustering rainfall. Another reference that is the reference material is a book that explains EOF analysis more deeply on weather data. Each rainfall data is obtained from Meteorology, Climatology and Geophysics (BMKG) data.

Based on the facts above, then in this article the authors conduct research to create a system that can recognize the weather at that time, then the system will classify the weather automatically, process quickly and obtain accurate results. Previous research has been done using a combined method of PCA and LDA in facial recognition applications [7]. This Final Project utilizes digital image processing using the Linear Discriminant Analysis (LDA) method and is hybridized with the Principal Component Analysis (PCA). From an image can contain information that is very important in the imaging of the sky that will occur so that humans can analyze objects in an image without having to observe it directly. With the help of digital cameras, computers and digital image processing, a system that will be based on image processing using Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) methods as a classification method can be realized

2. Methodology

The study began with the study of theoretical development literature and research studies using a hybrid Linear Discriminant Analysis (LDA) Principal Component Analysis (PCA) Method. The results of the literature study will contribute to the progress of the research and determine the theory used.

2.1 Framework

This study uses data from previous researchers and secondary data to capture cloud images. The captured image is used to identify weather conditions. To be able to identify the weather, certain feature extraction techniques are carried out. This feature extraction technique is done by mapping the color of the captured sky image. The mapping results are then grouped and used as variables used as training data / training data. Each cloud condition is captured hundreds of times to get different weather patterns as a basis for the developed system intelligence.

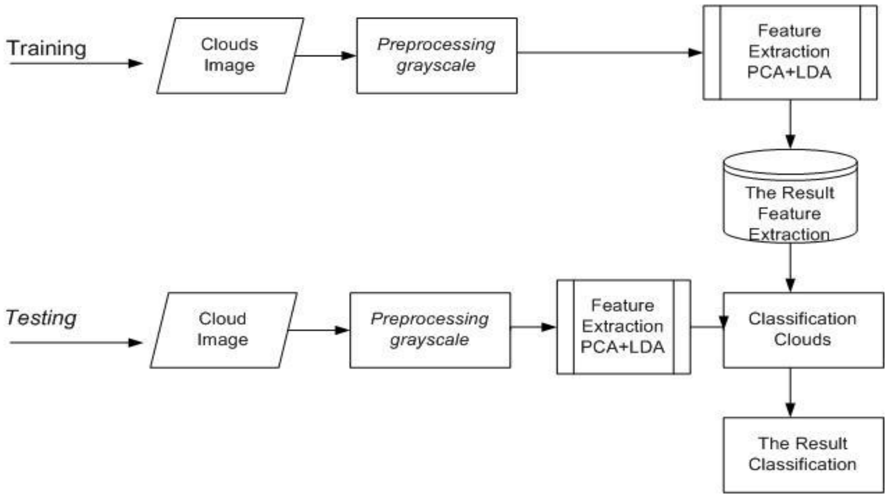

Training
Clouds Image
Preprocessing grayscale
Feature Extraction PCA+LDA
The Result Feature Extraction
Testing
Cloud Image
Preprocessing grayscale
Feature Extraction PCA+LDA
Classification Clouds
The Result Classification

Figure 1. Framework

In system design there are two main processes that will be carried out are the process of training and testing. In the training process. Collections of cloud images are input into the system then the cloud image becomes grayscale and extracted from the features of cloud images using the PCA and LDA algorithms, the results of feature extraction are stored in the system database. Then in the testing process, 1 cloud image inputted into the system then will be classified by cloud using the Euclidean Distance algorithm which is classified from the previous system database and the results of the classification process are weather prediction which is bright, cloudy and rainy

2.2 Feature Extraction Method

The feature extraction method used to generate the characteristics of each cloud image uses a hybrid Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) algorithm. In rainfall data, the PCA method can analyze the evolution of data based on the viewpoint of time and location of observation. When viewed from the stages of analysis carried out, Principal Component Analysis (PCA) is in line with the Empirical Orthogonal Function (EOF) method, or Karhunen-Loeve Transformation (Singular Value Decomposition (SVD) in the matrix. The LDA classification method, which is expected to be able to recognize cloud images and separates well between sunny, cloudy and rainy weather conditions with high accuracy and has faster computation time so that it can be used as a database to help BMKG predict the weather.

2.2.1 Cloud image extraction with PCA + LDA algorithm

The purpose of cloud image extraction or training phase is to get the main features of cloud images, which will be used to be compared with cloud images that will be detected.

2.3 The Basic Theory of Clouds

Clouds have a huge influence on weather and climate. Clouds are a key element of the earth's hydrological cycle, which carries water from the air to the ground and from one region to another [3]. Climate change caused by clouds in turn causes changes in clouds due to climate. This input can be positive (strengthen change) or negative (tends to reduce total change), depending on the process involved. These considerations cause scientists to believe that the main uncertainties in climate model simulations are caused by difficulties in clouds and sufficiently represent the nature of cloud radiation.

2.3.1 Types of Clouds Based on Shape

Cloud forms vary depending on weather conditions and altitude. But its main forms are three types, namely, layered in Latin called stratus, which is fibrous in shape called cirrus, and clumps are called cumulus (Indonesian spelling: stratus, sirus, and cumulus)[8].

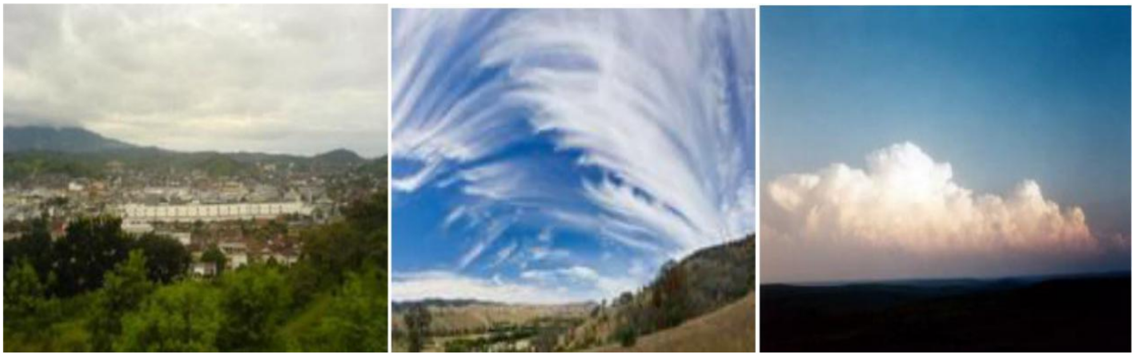


Figure 2. Clouds Stratus, Sirus, Cumulus

2.3.2 Cloud Characteristics.

Weather always changes over time based on cloud conditions. Clouds have characteristics of each weather condition produced. Some examples of cloud characteristics are:

1. Bright Clouds

   Bright clouds indicate bright weather, has a white cloud texture and combines with blue skies and slightly orange due to the influence of sunlight.

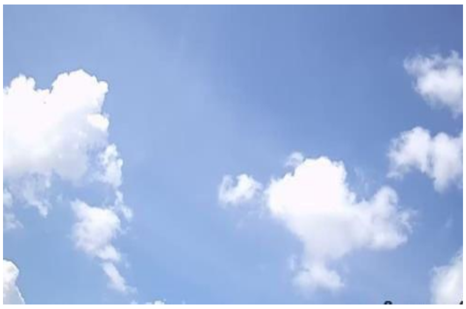


Figure 3. Bright Clouds

2. Cloudy Clouds

   Cloudy clouds can be predicted for rain. The texture of the cloud is gray, white and black. But the dominant color is gray.

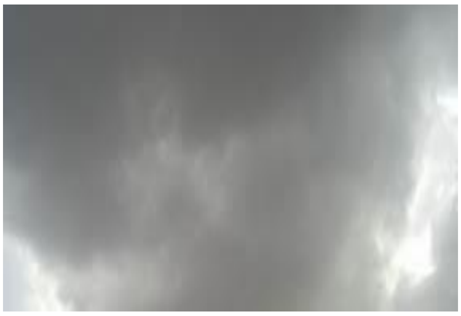


Figure 3. Cloudy Clouds

3. Rainy clouds

   It has a black and gray cloud texture. But the dominant color is the darkest color, black. Dark clouds indicate rain.

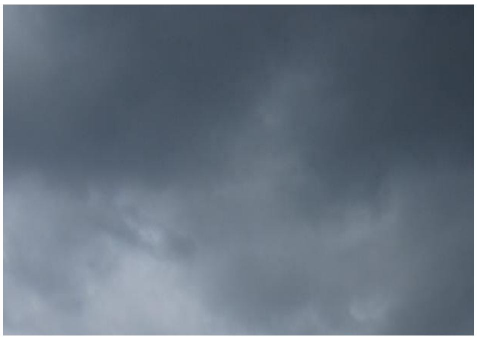


Figure 4. Rainy clouds

2.3 Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) concept is grouping variables that are linearly correlated into 1 main component, so that from the random variable (x1, x2, x3 ......, xp) will be obtained the main component k (k <p) which represents variable variability which exists. The purpose of doing PCA is to reduce the structure of variable relationships into new variables with smaller dimensions. The new variable is able to explain most of the total data variants and are mutually independent of each other. Furthermore, this new variable is called the principal component (PC). Dimensional reduction of data in PCA by transforming original variables that correlate into a set of new variables that are uncorrelated, while maintaining the greatest possible variance that can be explained.

2.4 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is used to determine the low-dimensional features of a high-dimensional space that helps to group photos of the same class and images from different classes. LDA chooses features that maximize the ratio of classes and spread into the class. LDA works based on a matrix analysis that aims to find an optimal projection so that it can project input data in a space with a smaller dimension where all patterns can be separated as much as possible.

2.5 Hybrid LDA and PCA algorithms

In the combination of PCA + LDA algorithm, PCA algorithm is used to reduce the calculation of matrix dimension n x n (n is the number of pixels) to m x m (m is the number of image training). From this matrix, the feature matrix of PCA is obtained. Furthermore, this PCA feature matrix will be used as input for the LDA algorithm.

2.6 *Euclidean Distance*

Euclidean space is a space with limited dimensions that have real value. Euclidean Distance Between two points is the length of the hypotenuse of a right triangle. Where x is the training image, and y is the input test image. If x = x1 + x2 + x3 + ... x_n and y = y1 + y2 + y3 + ... y_n are the two points in Euclidean x to y are:

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + \cdots + (x_n - y_n)^2} \qquad (1)$$

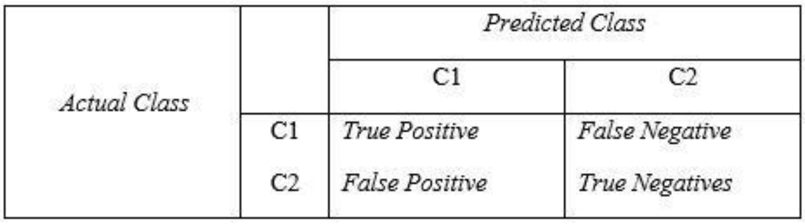
Euclidean distance is the most commonly used metal to calculate the similarity of 2 vectors. Euclidean distance calculates the root of the 2 vector difference square (root of square differences between 2 vectors) .

$$D_{ij} = \sqrt{\sum_{k}^{n} =_{1} \ (x_{ik} - j_k)^2} \qquad (2)$$

2.7 Confusion Matrix

Confusion matrix is a technique used to analyze how well the classifier recognizes data sets from different classes. True positives (TP) and true negatives (TN) provide information when the classifier is correct, while false positivies (FP) and false negatives (FN) tell when the classifier is wrong.

Table 1 *Confusion Matrix*



		Predicted Class	
Actual Class		C1	C2
	C1	*True Positive*	*False Negative*
	C2	*False Positive*	*True Negatives*

Information:
1. TP is the sum of true positives.
2. P is the number of positive columns.
3. TN is the sum of true negatives.
4. N is the number of negative columns.
5. FP is the sum of false positives.

3. Result and analysis

In this section a discussion on weather classification is presented with a discussion of the level of accuracy and processing time of the system. The classification model of learning outcomes and the results of testing the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) methods and compared with weather classifications based on cloud images using Linear Discriminant Analysis (LDA) on cloud images used and analysis of test results. The study was conducted using the java programming language.

In this study analyzed in the form of accuracy or accuracy of the system in classifying data. Therefore, using confusion matrix to test the accuracy of the classification results. Confusion matrix compares the results of manual calculations with results processed by the system.

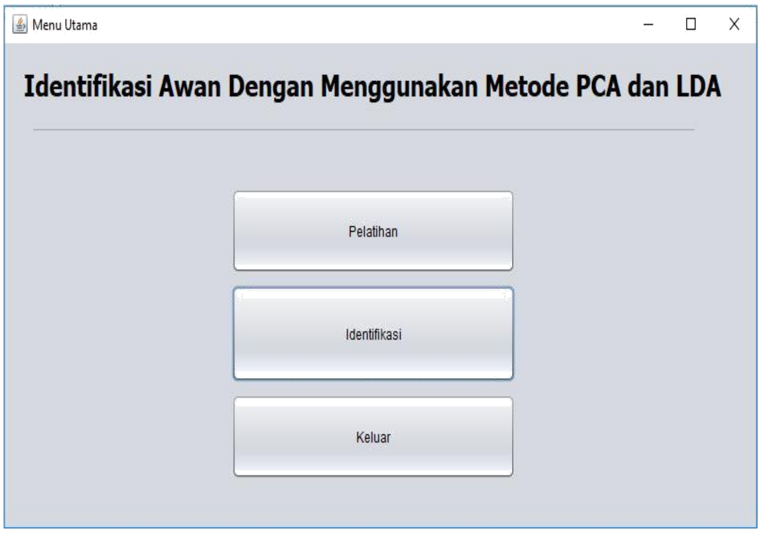
3.1 System planning

In system design there are two main processes that will be carried out are the learning process and the classification process. The learning process is the process of teaching the system how to recognize the characteristics or characteristics of a cloud image that will be input into the training data. The first step is to process the cloud image. The cloud image will go through the next stage, which is feature extraction to get the texture image of the cloud. And the cloud image is changed to grayscale. The next step is to store the data extracted from the feature as training data.

The next stage is the classification process. This process is carried out on testing images and training images, because this process requires training data to determine the results to be released on testing data. Image testing of an unknown cloud image that will be tested on the system. The next process is the classification data testing process of training data that has been stored in the database using the euclidean distance method. The results of the classification process are weather predictions which are bright, cloudy and rainy.

3.2 Interface of cloud identification system using Hybrid PCA and LDA methods

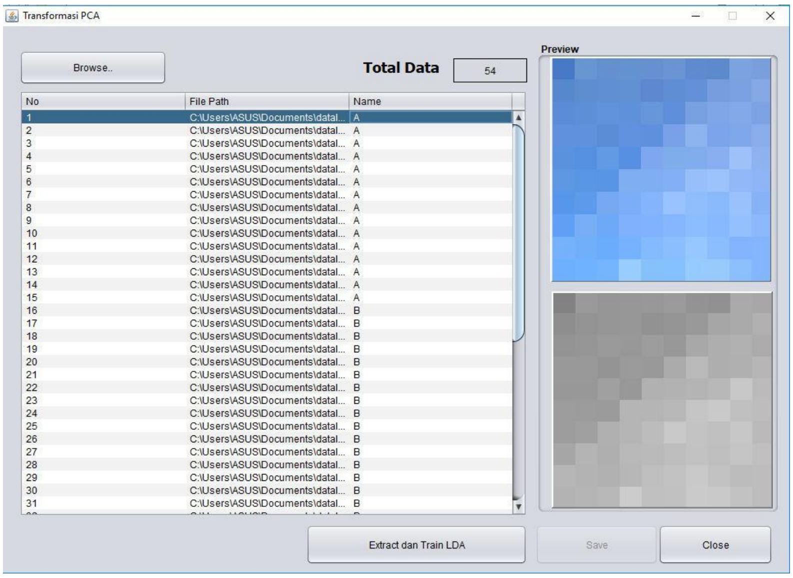
This page is the main page of the cloud identification system using the hybrid Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) methods. Where this page displays a cloud identification system that can provide weather information that indicates bright, cloudy and rainy. This page is the admin work page. There are 2 menu options, namely the menu access to the learning page (training) and the testing page (identification).



Identifikasi Awan Dengan Menggunakan Metode PCA dan LDA

Figure 4 The main interface of the PCA and LDA cloud identification systems

On the learning page, the previous cloud image is stored in a folder labeled A for bright clouds, label B for cloudy and label C for rain clouds. On this page all the clouds are inputted and the input image will go through the next stage, Preprocessing, which is changed to grayscale as shown in 5. Then click extract train lda. The next step is to click save the data extracted from the feature as training data.

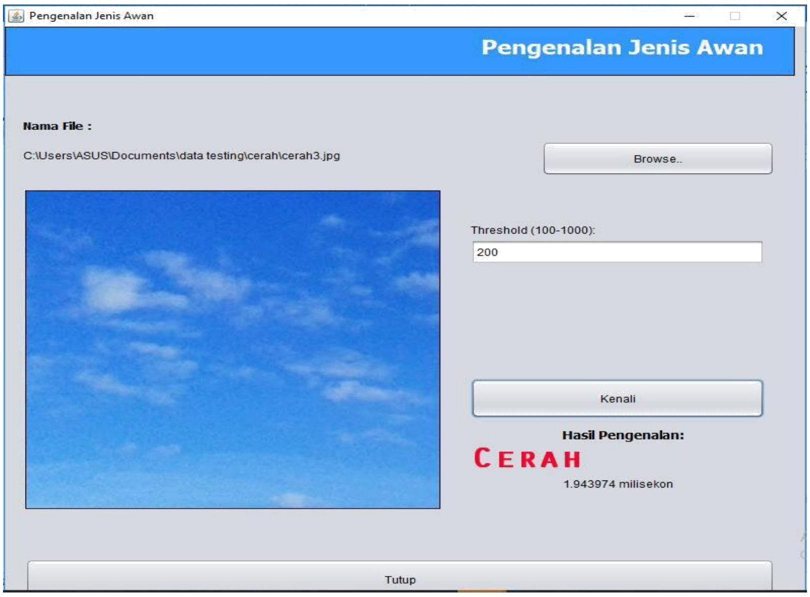


Total Data 54

No	File Path	Name
1	C:\Users\ASUS\Documents\datal...	A
2	C:\Users\ASUS\Documents\datal...	A
3	C:\Users\ASUS\Documents\datal...	A
4	C:\Users\ASUS\Documents\datal...	A
5	C:\Users\ASUS\Documents\datal...	A
6	C:\Users\ASUS\Documents\datal...	A
7	C:\Users\ASUS\Documents\datal...	A
8	C:\Users\ASUS\Documents\datal...	A
9	C:\Users\ASUS\Documents\datal...	A
10	C:\Users\ASUS\Documents\datal...	A
11	C:\Users\ASUS\Documents\datal...	A
12	C:\Users\ASUS\Documents\datal...	A
13	C:\Users\ASUS\Documents\datal...	A
14	C:\Users\ASUS\Documents\datal...	A
15	C:\Users\ASUS\Documents\datal...	A
16	C:\Users\ASUS\Documents\datal...	B
17	C:\Users\ASUS\Documents\datal...	B
18	C:\Users\ASUS\Documents\datal...	B
19	C:\Users\ASUS\Documents\datal...	B
20	C:\Users\ASUS\Documents\datal...	B
21	C:\Users\ASUS\Documents\datal...	B
22	C:\Users\ASUS\Documents\datal...	B
23	C:\Users\ASUS\Documents\datal...	B
24	C:\Users\ASUS\Documents\datal...	B
25	C:\Users\ASUS\Documents\datal...	B
26	C:\Users\ASUS\Documents\datal...	B
27	C:\Users\ASUS\Documents\datal...	B
28	C:\Users\ASUS\Documents\datal...	B
29	C:\Users\ASUS\Documents\datal...	B
30	C:\Users\ASUS\Documents\datal...	B
31	C:\Users\ASUS\Documents\datal...	B

Figure 5 PCA and LDA weather recognition system learning interface

On the testing page (identification), the cloud image will be classified and there will be an introduction to classified as bright clouds. Cloudy or rainy. There is also processing time. For example in figure 6 the cloud image is classified as a bright cloud with a time of 1.943974 milliseconds.



Pengenalan Jenis Awan

Nama File :
C:\Users\ASUS\Documents\data testing\cerah\cerah3.jpg

Threshold (100-1000):
200

Hasil Pengenalan:
CERAH
1.943974 milisekon

Figure 6 Interface of PCA and LDA cloud identification system testing

3.3 Weather Recognition System Testing

To find out the system performance that has been designed, it is necessary to test the system that has been developed. In testing it will be measured how much the success rate of the system designed in analyzing the weather classification system.

3.3.1 Dataset

The data used in this study are secondary data obtained from previous studies entitled weather detection systems based on digital image processing using the K-Nearest Neighbor (k-nn) algorithm with bright, cloudy, rainy, night, rainy night still features has something to do with the problems examined and from literature and other sources guilty of the internet (google image). The dataset before entering into the program is set to image resolution of 320x240 pixels.

3.3.2 Threshold

Threshold is obtained from the average value of the min and the max for each character feature. The threshold test needs to be done with the aim of whether the threshold value range used produces good accuracy or vice versa. Threshold value can be obtained from the distance Euclidean Distance on training and testing data. As an example:

Table 2. Testing bright clouds

Distance									
151	148	144	138	137	133	129	120	117	115
154	152	149	147	145	150	130	124	120	118
158	159	154	151	161	151	137	128	123	122
161	162	161	161	163	155	141	138	132	126
166	166	165	164	166	156	149	145	137	137
169	171	170	168	160	162	157	151	147	145
175	175	175	175	175	170	167	160	158	151
175	178	182	186	185	177	175	169	162	156
183	185	190	198	179	179	178	173	166	160
184	187	201	192	181	179	180	175	168	162

distance 186.10983614820336 (group 0)
distance 335.7604056466456 (group 1)
distance 193.12283655746154 (group 2)

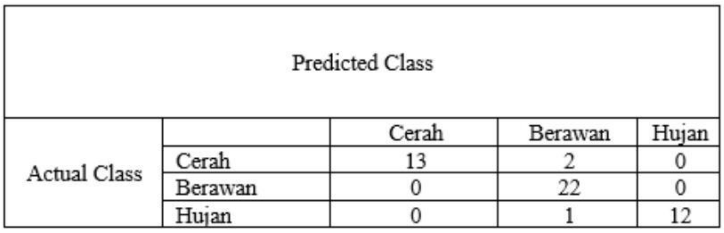
The dataset is entered into 3 groups. Group 0 is bright cloud, group 1 is cloudy cloud and group 2 is rain cloud. Data From the experiment above, we get Euclidean Distance distance that is 186.10983614820336, 335.7604056466456 and 193.12283655746154. The distance taken is the smallest distance for each testing data. So from the example above, the distance is the result of a bright cloud which is group 0. So from the experiment above, a threshold value range of 100-1000 was made to find the best accuracy.

3.4 Testing of the Euclidean Distance Method

In the testing phase of the Euclidean Distance classification, the data used is testing data in the form of an image file format .jpg. In this test used testing data taken from previous research [4] and sources from the internet (google image). The amount of testing data used is 50 data in the category of sunny, cloudy and rainy weather. This testing scenario aims to determine the performance of the Euclidean Distance algorithm in classifying data into predetermined classes. In this trial, the training data in the database will become a reference for each new data testing. the results of the Euclidean Distance process are obtained by giving a value to the confusion matrix to calculate the accuracy value of the test. Using the threshold value and data testing as many as 50 sample data with various weather conditions.

Using the best threshold value and testing data as much as 50 sample data by sharing various weather conditions.

Table 3. Confusion Matrix



		Predicted Class		
		Cerah	Berawan	Hujan
Actual Class	Cerah	13	2	0
	Berawan	0	22	0
	Hujan	0	1	12

Accuration: $A=\frac{47}{50} \times 100\% = 94\%$

The threshold becomes a very important parameter size. After a range of 100 to 1000 and testing, the best threshold is 800, the system is able to produce the expected accuracy.

3.5 System Analysis

The first test was carried out by taking 3 cloud classes which were bright, cloudy and rainy with a total sample of 15 images for bright clouds, 19 images for cloudy clouds and 20 images for rain clouds with testing data total 54 images. Tests on test images obtained by the classification method of Hybrid Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) as in Figure 7 with a threshold value = 100 has an accuracy of 30%, with a threshold value = 200 having a 56% accuracy, with a threshold value = 700 has an accuracy of 92% and with a threshold value = 800, this combined method of PCA and LDA produces an accuracy of 94%.

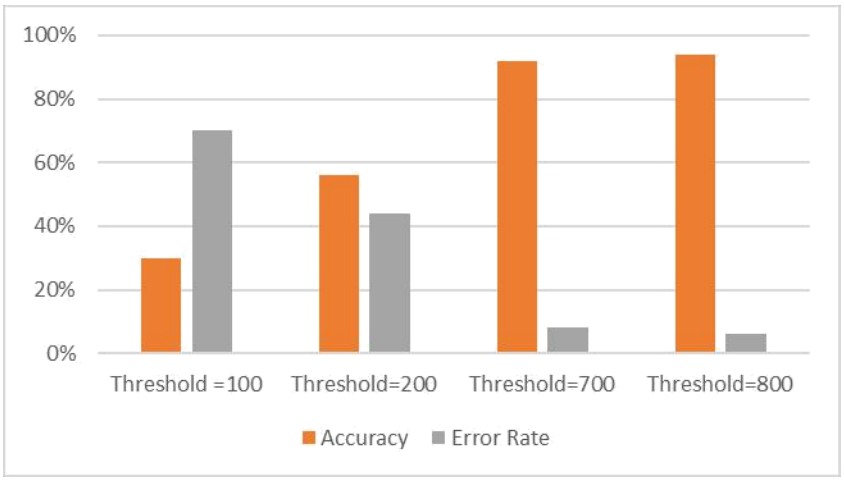


Figure 7. Graph of the level of accuracy and error rate of the hybrid PCA + LDA method

4. Conclusion

From the results of the testing and analysis that has been carried out on the design of the detection and classification system of weather conditions, the following conclusions can be drawn:

1. The design of a system for recognizing weather conditions based on cloud imaging using hybrid Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) produces an accuracy of 94% with a classification of bright, cloudy and rainy.
2. The thresholding value causes system accuracy to affect the detection and classification process.

5. References

[1] Krishnamurthi, Karthik, T. Suraj, K. Lokesh, P. Arum, “Arduino Based Weather Monitoring System,” International Journal of Engineering Research and General Science., vol. 3, issue 2, page.452-458, March-April, 2015

[2] Harper, Christopher W. Blair, John M.Fay, Philip A.Knapp, Alan K.Carlisle, Jonathan D., “Increased rainfall variability and reduced rainfall amount decreases soil CO2 flux in a grassland ecosystem” Global Change Biology, 2005.

[3] Yunita, nourma, Koredianto Usman, Suryo, " Detection and classification of weather conditions based on sky imaging based on digital image processing using the linear discriminant analysis (LDA) method (in english) ",2011

[4] Handoko, Arif, M.Ihsan Zul, Yuli Fitrisia, " weather condition detection system based on digital image processing using the K-Nearest Neighbor (k-nn) Algorithm (in english)",2015

[5] Gulumbe , 2012, "An Assessment of Vegetation Cover Changes across Northern Nigeria Using Trend Line and Principal Component Analysis", *Journal of Agriculture and Environmental Sciences, 1(1), pp. 01-18,2012*.

[6] Juaeni,Ina. Impact of Principal Component Analysis (PCA) Implementation on rainfall clustering over Java, Bali and Lombok Island 13 Juni 2014.

[7] Utami, Yustina Retno Wahyu, Teguh Susyanto, “Performance of Support Vector Machine (SVM) Classification Method with Learning Vector Quantization (LVQ) in Face Recognition Applications (in english)”, 2015.

[8] LAPAN. (2014). *Awan*. Taken on February 12, 2015, from the Field of Atmospheric Modeling: http://moklim.bdg.lapan.go.id/content/awan