CUMULONIMBUS SHORT TIME FORECASTING USING ARTIFICIAL NEURAL NETWORK WITH SELECTED RADIOSONDE INDICES IN JAKARTA, INDONESIA

Mr. Agie Wandala Putra

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science and Information Technology
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2014

การทำนายช่วงเวลาสั้นของการเกิดเมฆคิวมุโลนิมบัสด้วยโครงข่ายประสาทเทียมร่วมกับดัชนีที่เลือกจาก เครื่องวิทยุหยั่งอากาศในเมืองจาการ์ตา ประเทศอินโดนีเซีย

นายอากิ วรรณดารา ปุตรา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2557

Thesis Title	CUMULONIMBUS SHORT TIME FORECASTING USING ARTIFICIAL NEURAL NETWORK WITH SELECTED RADIOSONDE INDICES IN JAKARTA, INDONESIA
By	Mr. Agie Wandala Putra
Field of Study	Computer Science and Information Technology
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

.....Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

.....Chairman
(Assistant Professor Suphakant Phimoltares, Ph.D.)

.....Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

.....External Examiner
(Dusadee Sukawat, Ph.D.)

5672608423 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
KEYWORDS: CUMULONIMBUS / ARTIFICIAL NEURAL NETWORK / FEATURE SELECTION / CLASS IMBALANCE PROBLEM

AGIE WANDALA PUTRA: CUMULONIMBUS SHORT TIME FORECASTING USING ARTIFICIAL NEURAL NETWORK WITH SELECTED RADIOSONDE INDICES IN JAKARTA, INDONESIA. ADVISOR: PROF. CHIDCHANOK LURSINSAP, Ph.D., 93 pp.

This research proposed an accurate forecasting of Cumulonimbus (Cb) Cloud development events. Cb plays a vital role for mostly weather hazard in Indonesia, from thunderstorm, lighting, small tornadoes and torrential ice. Neural network has been used in numerous meteorological applications including weather forecasting. By using the data from *radiosonde* observation indices and data cloud classification from *Multifunctional Transport Satellites* (MTSAT), neural network was used as a classifier, also reduced atmospheric indices to select best variables by using *Principle component analysis* (PCA). The location interest of this research is Jakarta Area in Indonesia with data observation from Cengkareng Meteorological Station. An application to classify the Cb occurrences was shown and evaluated.

CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

This research proposed an accurate quantitative forecasting of Cumulonimbus (Cb) Cloud development events. Cb plays a vital role for most weather hazards in Indonesia, namely thunderstorm, lighting, small tornadoes and torrential ice. Neural network has been used in numerous meteorological applications including weather forecasting. In this research, a method was used to predict Cumulonimbus events with the data from radiosonde observation indices and cloud classification data from Multifunctional Transport Satellites (MTSAT). The location of this research is Jakarta, Indonesia and the observation data were derived from Cengkareng Meteorological Station. There has been no study academically reported under this topic and conducted within Jakarta area.

Atmosphere in Indonesia has a different system of circulation. Many parameters can build up the severe weather **[1]** and every severe weather accident always records the Cumulonimbus (Cb) cloud event. Cumulonimbus cloud can bring thunderstorms, gusty, small tornado and cyclone when it is over water and warm air **[2].** In tropical country, Cb development as the impact of convectifity process may happen any time during the day and also any time year-round but it is more commonly during rainy season **[3] .**

Cb are thunderstorm clouds that can grow vertically with strong energy. The cloud bases around 300 m (1000 ft), but it can grow upward to over 12,000 m (39,000 ft) **[4] .** Tremendous amounts of energy are released by the condensation of water vapor within a cumulonimbus. As a result, Cb is associated with lightning, thunder, and even violent tornadoes. Besides toward the industry and other humanity, the most important and hazardous impact of Cb events is toward aviation **[5]**. The only sensible defense against the hazards associated with a Cb is therefore

to avoid flying into one in the first place. Predicting an individual Cb cell is difficult, but it is possible to predict the conditions which will trigger the formation of a Cb.

Artificial neural networks have been studied since long times ago, but efficently for using in weather forecasting just appeared in the last 20 years. With a radiosonde, artificial neural network (ANN) predicts thunderstorm in high latitude **[6]** and it also produces quantitative of precipitation with short term forecasting **[7]**.

Radiosonde ballons are used to measure the vertical profile of atmosphere representing the condition of stability around their track. Using GPS technology, they send the location and data recorded then process the data through ground receiver in the station earth. We can use all the data result for analysisng the weather condition, making the skew-t plot diagram, and also asimilating it with numerical weather forecasting to measure the global weather system.

Current technologies for predicting Cb occurrence mostly were developed from remote sensing system **[8]**. This research tried to analyze the atmospheric indices representing the stability condition of the area where Cb event will be appearied. The analysis was carried out in order to support a weather forecaster and to understand the atmospheric condition of the Cb event.

1.2 Problem Formulation

This research focuses on the following problems:

1. Which are the most appropriate tools for predicting cumulonimbus development?
2. Make the cloud type algorithm using sensor in MTSAT with neural network?
3. What is the suitable artificial neural network model for cumulonimbus forecasting?
4. How do seasonal data affect the predicted results?

1.3 Expected Outcomes

1. A highly accurate predicting model for cumulonimbus event
2. A high accuracy of new cloud type classification

3. A set of appropriate radiosonde indices affecting the formation of cumulonimbus case.

1.4 Scope of the Work

This research will limit the scope to the followings:

1. The sample size of this experiment is radiosonde observation data from the Cengkareng meteorological station, Jakarta, Indonesia, from 1 January 2010 – 30 September 2014 with twice a day operation.
2. Initial 78 indices from radiosonde data are considered.
3. The cumulonimbus records are collected from cloud type product with remote sensing data (MTSAT) provided by Japan Meteorological Agency and our new purposed algorithm.

1.5 Document organization

This research is organized as follows. Chapter 2 explains the theoritical background. Chapter 3 describes the proposed method, and Chapter 4 describes the experiment and results. Then, some evaluation, benefits, final thoughts, and suggestions for future work are given in Chapter 5.

CHAPTER 2
THEORITICAL BACKGROUND

2.1 Weather in Tropical Region

2.1.1 Tropical Regions

From the aspect of astronomy and geography, tropics are in the areas between 23.5 degrees north latitude and southern latitude. However, the meteorological definition is inadequate because the weather with no geographic boundaries are fixed. In meteorology, the tropics are defined as the area between the line meetings of the southern and northern hemisphere's trade winds **[9].** Location of the line is not fixed and is not always clear. The line is referred as meteorological equator or the Inter-Tropical Convergency Zone (ITCZ).

Although mentioned as pale, in reality, ITCZ does not show intact form and it is always not fixed. It shifts from the north to the south to follow the motion of the sun, resulting in irregular shift every day that is not same in every place **[1]**. Over the ocean Atlantic and eastern Pacific, ITCZ has small motion friction, unlike the one in the tropics of Asia which shifts between 15 North to 20 South and often unclear.

In climatology, the tropic is approximately equal to the temperature average region, which is higher than 18 ° C. Actually the limit is not appropriate because in the tropical places with high altitudes, the temperature can be less than 18 ° C. Tropical climate regions are identified from a combination of factors such as temperature, rain, wind, vegetation. By considering these factors, tropical climate region is divided into two zones: the tropical zone inside and outside the tropical zone. Tropical zone is characterized by a region that is always moist and has more heavy rain (more than 8 months), and is exposed to equatorial westerly winds.

To distinguish the outer region of the tropics and the criteria used in various ways from climatology, the tropics encompass limited to the 18 ° C isotherm in northern latitudes and in the southern latitudes, with rainfall greater than 800 mm

per year. With atmospheric dynamics criteria, many values can be used to distinguish between the nature of tropical and extra-tropical atmosphere.

2.1.2 Atmospheric Circulation in the Tropical Zone

The term of general circulation has two meanings. Firstly iit is the circulatory of the system when derived from the equations of motion in the earth scale, and secondary it is also the average distribution values of weather elements all over the world. The general circulation of the atmosphere is described by the flow patterns which are meridional and zonal.

2.1.2.1 Hadley Circulation

Hadley circulation models which are based on the deployment of energy from the sun and the earth's rotation are still relevant to be used as an initial discussion. In global scale, Hadley suggested that every part of the earth's surface receives energy from the sun that is not simultaneous and not in equal amounts. The tropics are exposed to the light more than others continuously; even around the equator, the intensity is almost the same throughout the year. Then, the regions closer to the polar surface of the earth receive the amount of the light fewer.

With the energy from the sun and the encouragement of the rotation of the earth, it is resulted in a pattern of atmospheric circulation. The tropics receive a lot of sunlight, so that the radiation in the tropics make the regions move to become low pressure area, and the area outside the tropics receive less light, turning them into the high pressure area. Nonetheless, the areas in the poles are categorized into the low pressure areas because the atmosphere is very thin. The pressure difference between the areas pushes the air in the area of high pressure to flow to the low pressure area.

Because the earth’s rotating system is from west to east, the air around the high pressure area in the northern hemisphere presses toward the equatorial low pressure

area. This is deflected to be the eastern wind, while heading toward the polar regions of low pressure turns east into the west wind, called the polar westerly wind. In the South hemisphere, the air from the subtropical high pressure area toward a low pressure area in the equator is deflected to the left to form the southeast wind, while leading to polar low pressure area turns east wind into the polar westerly wind.

The airflow from the subtropical high pressure area toward the equator blows continuously, known as the trade winds. In tropics and north hemisphere, the trade winds that blow are as northeast trades, while in the tropics and southern hemisphere the trade winds that blow are southeast trades. The air in the tropics receiving a lot of light of the sun tends to move upward, whereas in the polar regions which receive little sunlight, the air tends to move downward. In the area where the air is moving up, many clouds turns into stir, thus forming a vertical circulation pattern.

2.1.2.2 Walker Circulation

Along the equator, there are zonal winds forming a circulatory system known as the Walker circulation. In general circulation over the equatorial Indian Ocean bottom, there is a west wind at the top of the east above the equator, there is Pacific wind at the bottom of the east and at the top of the west. While above the bottom of the Atlantic equatorial, winds blow from the west and across the top of the east. Thus, on a global scale, like in Figure 2.1 shows the atmospheric circulation system in the tropics consists of a component in the direction of the meridional Hadley circulation and Walker circulation in zonal direction.

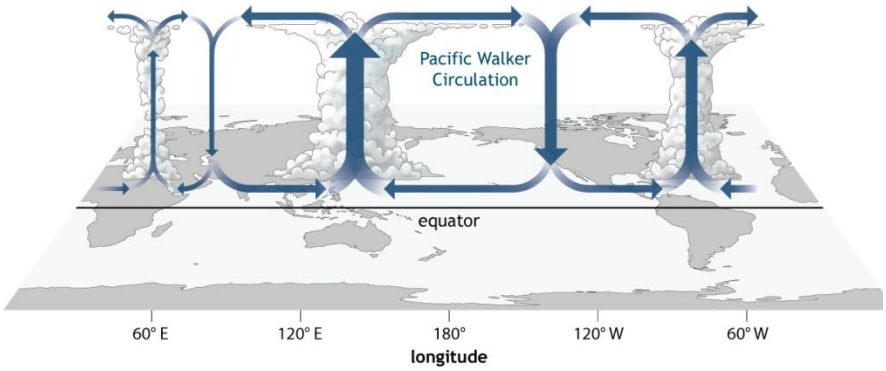

Figure 2. 1 Walker circulation **[13]**.

2.1.2.3 General circulation in Indonesia

Indonesia which is known as the maritime continent region has small-scale wind circulation known as meso-scale circulation. The circulation is associated with the emergence of local air instability. Of further discoveries, general circulation models in Indonesia consist of six components, namely the northern Hadley cell, the southern Hadley cell, the Walker western cell, eastern Walker cell, local circulation cells, and stratospheric oscillation cell (QBO).

Because Indonesia is located in the equatorial region, and the region is characterized by monsoon circulation associated with the Asian monsoon system and Australian monsoon, the Hadley cell becomes mingled with the monsoon circulation cell, shifted to north to south seasonally. Walker western changes are associated with seasonal Indian monsoon, while eastern Walker cell is more related with ENSO, local circulation cells associated with daily heating, and quassi biennial osillation (QBO).

2.1.3.4 Circular Pattern of Flow System in Equatorial Regions

In a smaller scale, the pressure and flow of air in the equatorial region have four basic patterns, namely: equatioral duct, bridge equatorial, equatorial step, simple cross equatorial drift. Equatiorial duct is the pattern formed by the high pressure side contiguous equator. Under the circumstances, there is a duct equator where winds around the equator are parallel to the equator from the east and geostrophic. The

pattern is formed frequently in the Atlantic, but it almost never occurs in the Indonesian region.

Equatorial bridge is the pattern formed by the low pressure in the next couple contiguous equator. Under the circumstances, there is a bridge equatorial wind around equator that comes from the west and is geostrophic, then to the east of the area, Low pressure cyclonal turns to the north and to the south of the equator. Such situation often occurs in the western Pacific near Papua, and often arises in connection with the active ENSO.

Equatorial step is the pattern formed by a pair of high pressure in the north and low pressure in the south of the equator. Wind around the equator is quasi geostrophic. Under the circumstances, there is a step equatorial, where equatorial wind passing turns into equatorial westerlies. This so often happens at the beginning of the cold monsoon in Asia during November-December.

Meanwhile, simple cross equatorial drift is the pattern formed by high pressure in the north of the equator and the high pressure and low pressure to the south of the equator. The wind blowing from the north across the equator and is siklonal in characteristic. This situation often occurs in Indonesia at the time of cold surge Asian monsoon in the month of January to March.

2.1.2.5 Monsoon

Many researchers have suggested that the word of monsoon comes from the Arabic "Mousim"; this is the name of Arabic seasonal wind in which in the intervening six months the wind blows from the northeast, and six months later, it blows from the southwest. The term then was used in other areas. Even in Europe, this term is also used to call the period of western wind with "European monsoon". Monsoon is the main cause of the annual variation of temperature difference in mainland broad (continents) and the surrounding seas. The temperature difference is then followed by the increasing pressure on land during the winter which is very low in the summer. The areas which have a monsoon system in the tropics generally are

Northern Australia, Africa, Spain, Texas and the west coast of the United States, Chile. Monsoon is most obviously seen in South and East Asia. In India, the most occurring monsoon is the southwest monsoon. Moreover in Indonesia, the monsoon belongs to south-west monsoon and east monsoon, though this is not always the same in all regions in Indonesia.

Indonesia is not the source of the monsoon area, but there are some areas through which the air flows monsoon. Because of differences in warming period between Indian Ocean and the Asian continent alternately every years, there is a tendency for ocean air mass to flow from the ocean to the continent and vice versa. In summer time, in Asia there is a tendency of the air to flow from the Indian Ocean to Asian continent. Monsoon arising is called as summer Asian monsoon. Because of the tendency of air flow direction from the south to the west, monsoon in India is referred as the southwest monsoon. In contrast to the winter period in Asia, the air tends to flow from the Asian continent to Indian oceans.Then, the monsoon is called cold monsoon, and when the air flow direction is from East Sea, it is called the northeast monsoon. In addition to the Asian monsoon, northern Australia also brings monsoon. During the winter, in the southern hemisphere, the north east wind blows in Australia, while in the summer in southern hemisphere or winter in Asia, westerly wind blows in northern Australia from the Indian Ocean.

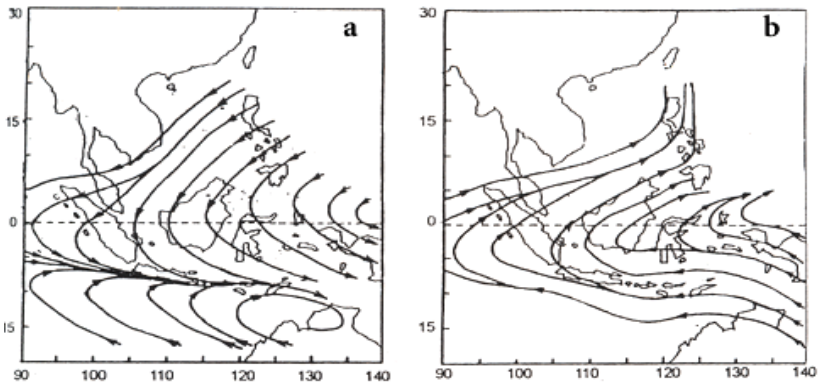

Figure 2. 2 Normal wind 3000 feet (a)rainy season (b)dry season **[1]**.

Australian continent have in same area with high pressure, when this are in summer time then the air flow moving to the lower pressure at the same time Asian continent mostly reach the winter season, also when the asian summer monsoon

bring air mass blow up and strengthen to the south east asia area include Indonesia like shown in Figure 2.2. Since the turn of the stream passes through the territory of Indonesia, the season in Indonesia has changed following the change of monsoon or monsoon. However, since Indonesian territory is quite large, the monsoonal characteristics vary.

The wind comes from the northeast trade winds turns into a west wind; whereas during the summer monsoon, the wind blows over the area from the southeast to the southwest as an extension of the southeast trade winds. On the top of Sumatra, the southern part of Java, East Nusa Tenggara. From the wind circultion system, it can be seen that both the time and the cold monsoon during summer monsoon are different in every region of the state, so that weather phenomena are also different. Jakarta is characterized by the air from South China Sea and the Indian Ocean. The characteristics of the existing air determine the phenomenon occurs, especially to rainfall. Although the value is different, the rainfall in Indonesia experiences annual monsoonal variation, i.e when within six months of a year there are a lot of rain and this period is called rainy season, and in the remaining six months there is less rain called the dry season.

2.2 Cumulonimbus development theory

2.2.1 Solar Radiation

Solar radiation is an electromagnetic wave that is very important for the survival of living beings on earth, and as the main energy source for the atmospheric and physical processes that determine the state of the weather and climate. Acceptance of solar radiation on the earth's surface varies according to place and time, such as the location of the latitude. At the micro-scale, slope will determine the amount of radiation received.

The process of radiation energy transfer between the air above the canopy through convection is a mass movement of the fluid involving a process of material transport and mixing. Turbulence is in progress at any time and plays an important

role in the exchange process and the nature of the material or to the atmosphere. Turbulence accelerates the transfer of water vapor from transpiration and evaporation process as well as the transfer of heat that serves as a canopy for surface temperature controller, so that the process can take optimum plant photosynthesis.

Besides, turbulence may also be known as wind speed profile. The wind speed and the characteristics of a plant canopy become the determining factors for the exchange of water vapor, heat, CO2, and momentum between the air above the canopy.

Solar radiation reaching the earth will eventually be converted into the kinetic energy of the gas in the atmosphere and will lead to movement where the molecule is being fixed. Large wind vector is called wind speed, while the wind direction is the direction from which the wind blows. The rate of surface wind is impaired fast. The development of the disturbance called Gustiness. Standard placement of surface wind gauge is mounted as high as 10 feet above an open field, with a distance of at least 10 times the height of the building - a building or barrier around it. Wind direction is expressed in degrees, measured in the clockwise motion, starting from the northern point of the earth. In addition, the wind speed is expressed in knots, where one knot is equal to 0.5 m / s. Surface wind direction is determined by the wind vane, while the wind speed is measured using anemometer cup counter.

Sensible heat is the amount of heat energy which can cause changes in body temperature. Therefore, to raise or lower the temperature of an object, some amounts of heat energy are required. On the contrary, latent heat is the amount of heat energy that can cause changes in the states of matter. For example, when water is converted into a gas or vapor, then the required amount of heat is called latent heat. In this case, latent heat of vaporization and the latent heat of condensation are distinct. During the process of changing the temperature, the object does not change. The term of sensible heat is used in contrast to the latent heat, which is the amount of energy exchange that is hidden; in other words, it cannot be observed as the temperature changes. For example, during the phase changes such as melting ice,

the temperature of system containing ice and liquid is constant until all the ice has melted.

Sensible heat of a thermodynamic process can be calculated as the product of body mass (m) with company specific heat capacity (c) and temperature change .Sensible heat and latent heat is not a specific requirement of this form of energy, but rather is characterized by the same form of energy, heat, in terms of its influence on the substance or a thermodynamic system. Sensible heat is the heat energy in the transfer process between the system and its surroundings or between two systems with different temperatures. The latent heat is associated with phase changes of atmospheric water vapor, mostly vaporization and condensation, whereas sensible heat is energy transferred that affects the temperature of the atmosphere.

During the day, the portion of the surplus heat that radiates down the atmosphere is known as the heat feels (QH). This heat must pass through, then the heat transfer is directly marked as the temperature gradient. During the day, there will be a negative value (lapse) and positive QH, while at night, positive gradient (inversion) and QH negative value are resulted. Ground flux is controlled by the value of Ks while the air is by KH.

Low temperatures tend to decrease the surface heat flux emanating from these low boundary layers and the temperature which is too cold. As soon as the sun rises, the sunlight heats the earth surface, then the heat is likely to rise, gather at a place and heat the lowest boundary layer.

The transfer of water vapor between the surface and the lowest atmosphere is important to describe the water vapor contained in the air. Evaporation from the surface is going through laminar boundary layer. This process does not only depend on the availability of water but also the availability of energy to change the shape, the presence of water vapor concentration gradient and turbulence to carry water vapor.

Change in moisture between the surface and the atmosphere is defined as humidity. Most of the heat flux noticeably affects the temperature in the lowest

boundary layer. Heat is pumped into the air during the day and returned to the surface at night; the water flux tends to go up in large numbers at all. Water vapor is transferred upward by eddy diffusion that is consistent with the heat.

Solar radiation is an electromagnetic wave generated from the nuclear fusion process that converts hydrogen into helium. Solar radiation reaches the earth's surface, but only about a half of that which receives the top of the atmosphere since most radiation will be absorbed and reflected to the space by the atmosphere, especially by clouds. An average of 30% of the solar radiation is reflected back from earth to the outer space. The sun is also a very important climate control and as a primary energy source that drives the earth's air and ocean currents. Transfer of energy that occurs without the need for a medium for transmitting is called radiation. The diameter of the sun is 1.42 x 106 km, and the surface temperature is 6000 K. The sun's energy is radiated in all directions with the same intensity; most of the energy has been run out to the universe, and only a small fraction is accepted by the earth. Solar energy that falls on the earth's surface is in the form of electromagnetic waves that propagate at the speed of light. The wavelength of solar radiation is very short and is usually expressed in micron (1 micrometer = 10 -6 m).

Natural energy equilibrium between the input of solar radiation, emissivity-dependent wavelength, and the heat transfer produce a diurnal cycle of heating and cooling of the earth's surface and atmospheric boundary layer near the surface. Energy balance is a major factor in the formation of weather, where there is a dynamic equilibrium between the energy input from the sun and the loss of energy by the surface after complex processes. The difference between the input and the output of the system is called the net radiation.

Energy received by the earth's surface will first be used to evaporate soil water and soil moisture, to heat the soil (S) and the rest is to heat the air (A). Very low soil water content and soil moisture (energy for small LE) cause the solar radiation falling into the surface in the form of net radiation (Qn) and used to heat the soil and the air; thus, the temperature increases.

2.2.2 Air Humidity

The state of water in the air as water vapor is expressed as humidity. Air capacity to accommodate the water vapor is determined by the value of the humidity. The link between humidity and air temperature is associated with the process of development and air shrinkage. When the air temperature is higher, the capacity of air to hold water vapor per unit volume of air is also getting bigger. Absolute humidity is expressed by the vapor pressure (e), which shows the water vapor content of the air volume or unity can also be expressed by the mass of water vapor per unit volume of air. The relationship between water vapor pressure, temperature and air volume can be derived from the ideal gas law equation as follows.

$$e = nRT / V$$

Ratio mixture (mixing ratio / w) is the ratio of the mass of water vapor to the mass of dry air per unit volume of air (mV / ms). Specific humidity (q) is the ratio between the mass of water vapor to the total mass of air, while the relative humidity (RH) is the ratio between the amount of steam with a capacity to accommodate water vapor in the air. RH values are expressed in percent.

Diurnal distribution pattern in near surface air humidity reaches a minimum level in the afternoon. The temperature reaches a maximum value and becomes maximum RH at the surface in the early morning because there is air deposition process that decreases the value of the ice.

2.2.3 Atmospheric Stability

The main factor is the stability of the atmospheric temperature with altitude relationship. The rate at which the temperature varies with altitude is called the lapse rate. Lapse rate has a significant influence on the vertical motion of the air. The mechanism by which air is moved vertically is attached to the concept of adiabatic lapse rate.

- Neutral condition, in which the actual lapse rate and the dry adiabatic lapse rate are the same, so that a parcel of air that move (either up or down) will

have the same temperature as the surrounding air, the density is the same, and will be in balance.

- Unstable conditions, where the actual lapse rate is greater than the dry adiabatic lapse rate. When the parcel is up, the temperature is greater than the surrounding air, its density becomes smaller and will remain on the move upwards. When the parcel is moving up, the temperature difference increases and accelerates the rising air parcel.
- Stable condition, in which the actual lapse rate is less than the dry adiabatic lapse rate. When there is an increase in parcels, the air temperature is less than usual, therefore its density is greater and the parcel will come straight down where the temperature is the same as the surrounding air.

The phenomenon is easy to recognize. The stability of the atmosphere is seen as Cumulonibus (CB) growth of convective clouds. This cloud formation at the start of the atmospheric conditions is unstable due to heating from solar radiation below to raise ground tempetature. Unstable air results in interference which in turn causes convective processes. Bad weather is partly because the cloud forms a storm or thunder and lightning.

Atmospheric stability is the tendency of the atmosphere to resist vertical movement or to suppress turbulence, affecting the ability of the atmosphere to disperse the pollutants that emit into the atmosphere.

Static instability in the atmosphere causes the convection of vertical mixing in the form of heat and possibly cumulus clouds. Vertical mixing can occur in a stable environment, particularly in the form of wave. These waves are the main cause of high turbulence, especially above the planetary boundary layer or near the jet streams, which often results in turbulence (CAT) as feared by aviators. Evolution of turbilen has been described mathematically by Kelvin and Helmholtz. Atmospheric stability conditions can be estimated using the Richardson number (Ri).

To find out the development of convective cloud, it needs to look at the conditions inside the atmosphere, generally measuring the stability of atmosphere. In tropical weather, a severe activity on atmosphere is always asssociated with the presence of cumulonimbus cloud. This cloud storm can bring such thunder, lighting, gusty and icing. It requires good understanding to the condition of convectivity system, starting from preprocessing on surface radiation received by the earth. In the first stage, we see the growth condition with deep moist covection (DMC). This is always related with unstable conditions, where a flow on the beginning starts with a positive value on the parcel. In case there is only a certain level, the topographic conditions are also sometimes helpful. In general principle, convective available potential energy (CAPE) is converted to the kinetic energy. On the instability processing, there is a condition which needs some forces before it turns into a condition called latent instability.

The basic concept of CAPE is how the mechanism lifts up the parcel of atmosphere to higher level. Furthermore, we can note the instable movement is resulted when the parcel is less dense than the environment condition. Then, the parcel will release the latent heat by accelarating the speed and making the real instability.

Thus, the main occurrence of deep moist convection is when the atmosphere notes the CAPE on their system, and also records the force to release the energy with the latent heat. However, in several atmosphere systems impacted by the topography condition, they cannot refer to the occurrence of CAPE **[23]**. Thus in every layer, by looking its lapse rate, the length between wet and dry adiabatic lapse rates can be measured. In some cases, no CAPE is found in convectivity event; the condition brings to that event with strong vertical wind shear noted as dynamic instability has important aspect to convect storm to release CAPE.

Atmospheric stability allows determining the trend of the vertical movement of a mass of air in the atmosphere. Small differences in the vertical movement are necessary to explain or predict the formation of convective clouds, rain and low pressure areas. Unstable air allows the formation of clouds, especially clouds that

have a striking vertical size and that usually lead to bad weather. In contrast with the sunny weather, no cloud results in stable air.

Atmospheric instability is theoretically atmospheric properties characterized by a condition when the air force at the time is put into the atmosphere. If the air force is put into the atmosphere, then there are three possibilities; one possibility is raising air force and the force likely to continue to rise. The second possibility is the air parcel remains in place and is likely to continue to put in place and adjusts the environment condition. The third possibility is the air force drops to a lower downward.

If the first possibility occurs, the air force is moving up and likely to continue to rise to reveal that the atmosphere at the time is in an unstable state. If the second possibility occurs, the air force is silent on any where the air force is released, in other words, the atmosphere is neutral (indifferent); and if the third possibility happens, the air force tends to fall down, in other words, the atmosphere is in a steady state (stable).

To mark the nature of atmospheric instability, small-scale (meso local) associated with particular phenomena is the quantitative value of the index used. The index generally consists of the value of the temperature (T) or potential temperature , which is expressed with such moisture dew point temperature (Td) and or the wet bulb temperature (Tw). The data used are temperature, wind, humidity obtained from observations with radiosonde. Cutting edge tool for radiounting is equipped with computer software that can directly calculate the various values of the index.

From the observed data, it shows that every element (including temperature) value changes after altitude. Following temperature changes altitude (dT / dz) and is called the "rate of shrinkage temperature". The rate of temperature decrease is expressed by the notation. If the air force put into the atmosphere is theoretically done with adiabatic process, the process of putting the air force into the atmosphere results in no heat coming into or out of the group. Then, the rate of shrinkage of the

air temperature in the cluster is called "adiabatic temperature shrinkage rate". If the air in the form of clusters of air saturated water vapor, the rate of temperature decrease is called "saturated adiabatic temperature shrinkage rate" and is expressed by the notation Υ; magnitude of about 0.3 ° C / 100 m. If the air in the form of clusters of dry air (containing no water vapor), the rate of temperature decrease is called "dry adiabatics temperature shrinkage rate" and is expressed by the notation d; magnitude of about 0.6 ° C / 100 m. Furthermore, the value of s and d is used as criteria for marking unstability of boundary layer of the atmosphere. This is done by comparing the rate of shrinkage which atmospheric temperatures are derived from the observed data radiosonde to the value of the rate of temperature decrease saturated adabat and on the rate of shrinkage temperature dry adiabat (d).

2.3 MTSAT-2

2.3.1 Remote Sensing using Satellite Meteorology

Observations in the field of atmosphere and meteorology are conducted using two basic approaches, which are direct observation and indirect observation called remote sensing. Remote sensing is defined as the science, technique, or art of obtaining information or data about the physical condition of an objects, targets, goals, areas and phenomena without touching or getting direct contact with the objects.

Satellite is a remote sensing instrument launched into the outer space to monitor the earth from a distance. Satellites measure indirectly via electromagnetic radiation coming from the surface underneath. Data or information obtained from the target object or signal wave electro magnetic reflectance and emission are recorded by the sensor. The term meteorological satellite then is used to refer a satellite orbiting the earth with a remote sensing instrument to obtain data on the atmosphere and oceans.

According to the cross-orbit, meteorological satellites are divided into two types: Polar Operational Environmental Satellites (POEs) and geostationary Operational

Environmental Satellites (GOESs). GOESs orbit above the equator at the same speed with the speed of the earth rotation, so that it can continuously transmit data from a particular region. Geostationary satellites orbit above a point on the earth's surface at an altitude of about 36,000 km.

POEs orbit around the earth with a north-south direction or move from pole to pole with certain height and inclination angle. The advantages of these satellites are the ability to reach all parts of the earth and to produce high-resolution images and atmospheric profiles due to a lower orbit. Meanwhile, the drawback of these satellites is the inability to observe a certain area continuously.

2.3.2 Electromagnetic Waves, Radiation and Satellite Sensors

All information about the earth and the atmosphere received by the satellite are in the form of electromagnetic waves. It is important to know and understand the mechanisms that drive the solar radiation and all the processes that occur when in the atmosphere.

An electromagnetic wave consists of electric and magnetic fields. The second vector field is perpendicular to each other, as well as on the spreading process. Radiation will be more easily addressed from the form of wavelength ($\boldsymbol{\lambda}$), which is the distance from wave crest to the next wave crest of the electric or magnetic wave.

The relationship between frequency (f) and wavelength ($\boldsymbol{\lambda}$) is as written below:

$$f = c/\boldsymbol{\lambda}$$

Where c is the speed of propagation of electromagnetic waves. That is worth the same as the speed of light in vacuum; speed of light is 2.9979 x 108 m/s. Nonetheless in the atmosphere, these waves move slower because of the interaction with the atmosphere.

2.3.3 Radiation on Black Body

Black body is an object that absorbs all radiation that comes when the temperature is lower than the surrounding temperature and transmits the entire energy of radiation, when the temperature is higher than the surrounding temperature. All objects at temperatures above absolute zero will emit electromagnetic radiation continuously. Therefore, the object on the earth and the atmosphere is also a source of radiation. The amount of energy radiated by the object is directly proportional to the fourth power of the temperature of the object itself. This is expressed by the Stefan-Boltzmann law:

$$w = \sigma.T4$$

where:

W = energy emitted per second from the area of the unit black body with temperature (T),

σ = Boltzmann constant (5.6697 x 10-8 Wm-20 K-4)

The nature of the sun in emitting radiation is like a black body radiator, which is the absolute temperature of 5784 K. The distribution of solar radiation length is about 6000 K. Due to the large temperature difference between the sun (6000 K) with the earth's surface (300 K), the spectrum of the sun and the earth is easily separated. Almost all (98%) energy radiation that reaches the earth's atmosphere lies in the visible zone (wavelengths 0:15 - 4 m), and 99% of the earth's radiation is in the infrared zone (wavelengths of 4-120 m). Both types of radiation experience different influences when they pass through the earth's atmosphere.

According to Planck's law of radiation, the energy distribution of the radiation is a function of temperature and wavelength. The statement can be written in equation as follows.

$$\mathbf{E(\lambda)} = \frac{\mathbf{C_1}}{\boldsymbol{\lambda}^{-5}\left[\exp.\frac{\mathbf{C_2}}{\boldsymbol{\lambda}\mathbf{T}} - \mathbf{1}\right]}$$

Where :

$E(\lambda)$	=	energy as wavelength radiation ($Wm^{-1}\mu m^{-1}$)
λ	=	wavelength (m)
T	=	temperature (K)

c_1 dan c_2 = constant $c_1 = 2\pi hc^2$ dan $c_2 = hc/K$

c = light speed (m/s)

h = Planck (6,62 x 10^{-34} Js)

K = Stefan Boltzman (5,6697 x 10^{-8} $Wm^{-2}K^{-4}$)

2.3.4 Multi-functional Transport Satellite (MTSAT)

MTSAT Satellite is one of the meteorological satellites observing the Asia Pacific region, including Indonesia. The satellite is operated by JMA (Japan Meteorological Agency).

MTSAT Satellite uses five channels like shown in Figure 2.3 , which are:

- Channel IR1 is located at a wavelength of 10.3 to 11.3 mm.
- Channel IR2 is at a wavelength of 11.5-12.5 mm.
- Channel IR4 is at a wavelength of 3.5-4.0 mm.
- The canal is visible at a wavelength of 0.55 to 0.8 mm
- Channel WV (Water Vapor) is based on the availability of water vapor in the atmosphere, which is at a wavelength of 6.5 to 7.0 mm.

The spatial resolution of IR and WV channel is 4 km x 4 km, while the spatial resolution of visible channel is 1 km x 1 km.

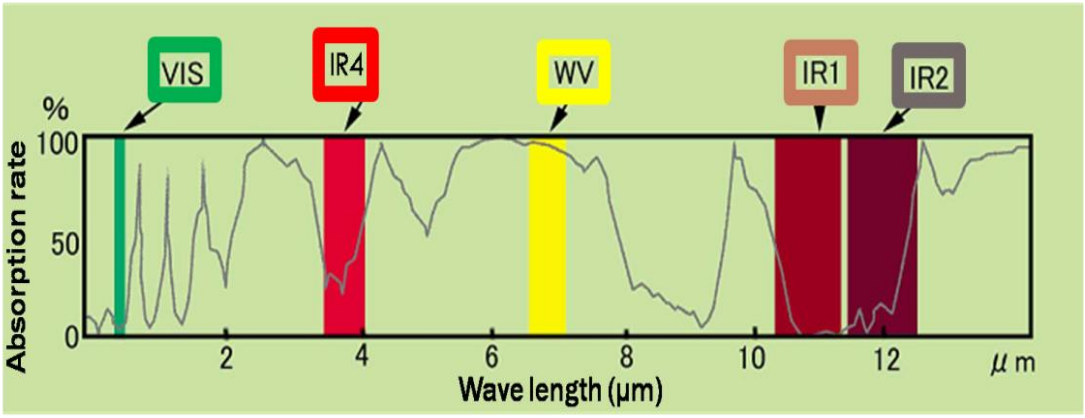

Figure 2. 3 MTSAT Channels **[22]**.

MTSAT satellites are satellites used by the Japan Meteorological Agency (JMA), Ministry of Land, Infrastructure and Transport. In February 2005, the MTSAT-1R was launched and then followed by MTSAT-2 in February 2006. MTSAT geostationary satellite is a type in which the satellite has a fixed position for observing the earth and been oriented around the equator; the earth's surface reaches a height of 35 786 km. MTSAT satellite orbits at a fixed point, namely the 1400BT positioned directly above Biak town, Papua, with a height of ± 36,000 km from the earth's surface. The satellite monitors the same place, that is 1/3 of the earth's area. The data recorded by the satellite are received regularly every hour.

This satellite has two sensors. The first sensor is an infrared sensor which consists of the IR-1 with 10.3 to 11.3 lm wavelength and 5 km resolution (10 bits) which is applied to monitor the cloud, the movement of the wind, and storm; IR-2 has a wavelength of 11.5 to 12, 5 lm with a resolution of 5 km (10 bits) which is applied to observe the sea surface temperature and monitor the volcano; IR-3 or so-called as Water Vapor (WV) has 6.5 - 7 lm wavelength with 5 km (10 bits) resolution which is applied to monitor the movement of water vapor in the middle layer of the atmosphere; the IR-4 called Near Infrared (NIR) has a wavelength of 3.5 - 4.0 lm with a resolution of 5 km (10 bits) which is applied to identify heat source and cloud at night. The second sensor is the Visible (VIS) with 0:55 - 1:25 wavelength and a resolution of 0.90 lm km (6 bits) and it is applied to observe the cloud, ice and snow during daytime.

VIS Imagery contains data / information about the reflection of sunlight by object present in the atmosphere and on the surface of the earth. Visible imagery is resulted from visible sensor. VIS sensors on satellites record data based on the difference in albedo normally reflected by the target cloud, land and sea. VIS sensors only work during the day due to its dependence on sunlight. Atmospheric particles are observed depending on:

a. Cloud illumination (sun angle)

b. Angular position of the camera toward clouds and sun

c. The reflection is very dependent on the cloud: the particle size distribution, particle composition, and surface characteristics on the cloud

The color varies from black to white, where the black shows low albedo and white indicates high albedo. The denser the cloud albedo is, the intensity of white color in the image will be much higher **[16]**.

IR Imagery contains data / information about the thermal radiation emission of existing object in the atmosphere and on the surface of the earth in an area with long infrared waves. IR sensors are obtained from the radiation emitted by the earth into the atmosphere in the thermal-IR wavelengths (10-12 m). These sensors provide information about the surface temperature or the cloud. IR sensor works based on the temperature difference, so that satellites can use this sensor to record objects during the day and night. The color ranges from black to white, indicating temperature difference.

Thermal IR sensor is able to show the land temperature, and ocean as well as the peak temperature of the clouds that exist on it. Mild temperature (0-30° C) generally is the averaged temperature of land or sea without cloud cover. Cold temperature from the top of the cloud is very high, and can show strong convective storm activity.

Image WV contains data / information emission of water vapor in the atmosphere. The image of IR4 (commonly called as image 3.5-4.0μm canal ~ 3.8 m) is a special channel that contains data / information on the patch area between the reflection of solar radiation and the radiation emitted by the earth's surface and clouds .Thus at night, radiation emission object contains information of the earth's surface, and in daylight, it contains a mixture of information between the reflection of sunlight radiation and emission of radiation from the earth's surface, where the reflection of sunlight radiation is more dominant.

The analysis of multiple satellite image / bispectral is a technique for observating the earth's surface or object with low cloud temperature (fog) or volcanic dust by using a combination of two or more different channels so as to produce

images that can distinguish types of clouds, or between clouds and snow / ice surface.

2.3.5 Cloud Type

In general, geostationary satellites are used to detect the types of clouds and cloud movement. By combining the interpretation of Brightness Temperature (BT) and canal reflectance of IR and VIS, cloud types can be identified. In the Figure 2-4 we can know the special charachteristic within each type of cloud from their reflectance.

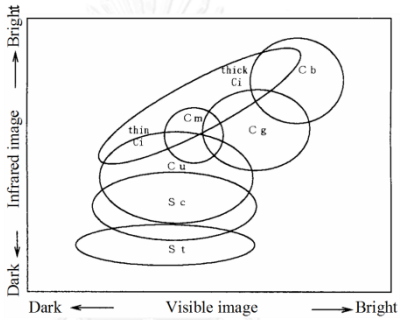

Figure 2. 4 Cloud type in Channel MTSAT reflectance **[22]**.

The first scheme shows how to predict the cumulonimbus cloud, with the cumulonimbus characteristics that can reach tropopause level if strong developed, so the water vapour counts as small value; it is because the radiance observed with thermal window channel in IR1 and water vapour channel is almost equal. This can be used by calculating the equivalent blacbody temperature (TBB) from each channel from the difference, so the treshold for cumulonimbus cloud can be found by citing around 1.5 K.

For other case as in multilayer cloud, there are different levels with high and midle level. The equation form radiative transfer value in semi transparent high cloud is derived, while blackbody below high cloud with thermal window radiance and water vapour are to be observed. On the cirrus case, if it is constant, water vapour and infra red also become constant at certain level. With the base cirrus calculated as the top level of cloud below dense and deep condition of stratus, this

type is represented by the different temperatures that are higher than other type of cloud which is easy to be identified.

2.4 Artificial Neural Network

2.4.1 Radial Basis Function Network

A Radial Basis Function (RBF) network is a special type of neural network that uses a radial basis function as its activation function. RBF networks are very popular for maximum functions of such as curve fitting, time series prediction, control and classification problems **[10].** In RBF networks, determination of the number of neurons in the hidden layer is very important because it affects the network complexity and the generalized capability of the network. In the hidden layer, each neuron has an activation function. The Gaussian function, which has a spread parameter that controls the behavior of the function, is the most preferred activation function. The training procedure of RBF networks also includes the optimization of spread parameters of each neuron. Afterward, the weights between the hidden layer and the output layer must be selected appropriately. Finally, the bias values which are added to each output are determined in the RBF network training procedure.

RBF network is a type of feed hyplane forward neural network composed of three layers, namely the input layer, the hidden layer and the output layer. The input layer is made up of sensory unit that connects the network to its environment. The second layer applies a nonlinear transformation from the input space to the hidden space with high dimensionality. The output layer is linear, supplying a response of the network.

An RBF network approximates f (x) can be expressed as:

$$\boldsymbol{f(x) = \sum_{i=0}^{n} w_i \, \phi(r)}$$

where $\boldsymbol{r = \|x - c_i\|}$, x is an input vector, $\boldsymbol{c_i}$ is the centroid at first layer weight, $\boldsymbol{w_i}$ is the connection weights in the second layer, and $\boldsymbol{\phi}$ is activation function radially symmetric. In this research, $\boldsymbol{\phi}$ is defined as follows.

$$\boldsymbol{\phi}(\mathbf{r}) = \mathbf{exp}\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c_i}\|}{\mathbf{2\sigma^2}}\right)$$

Variable input at this research is clustered with the technique of K-Mean. Since the error function is a quadratic function of the vector **w**, a pseudo inverse can be used to determine the optimal **w** to minimize the value of the error function.

$$\boldsymbol{w} = \boldsymbol{\phi}^{+}.\boldsymbol{T}$$

where $\boldsymbol{\phi}^{+} = (\boldsymbol{\phi}^{T}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^{T}$ is pseudo inverse of $\boldsymbol{\phi}^{T}$ and T=$\boldsymbol{\phi}\mathbf{w}^{T}$. In practice, we tend to use singular value decomposition to avoid possible ill conditioning of $\boldsymbol{\phi}$.

2.4.2 Support Vector Machine

Support Vector Machine (SVM) is a learning system that uses a hypothesis space form of linear functions in a feature space (feature space) with high dimension and trained with a learning algorithm based on optimization theory by implementing learning bias derived from statistical learning theory.

The theory underlying the SVM itself has evolved since the 1960s, but the newly introduced theory **[11]** in 1992 and has grown SVM rapidly. SVM is a relatively new technique compared to other techniques, but it is better performance in various fields of applications such as bioinformatics, handwriting introduction, text classification and so forth.

The process of learning the SVM aims to obtain hypotheses from the interface. This is good as it does not only minimize the empirical risk with an average error on the training data, but also has a good generalization. Generalization is the ability of a hypothesis to classify data that is not contained in the training data correctly. .

SVM aims to ensure the upper limit of generalization in the data test in a way of controlling the "capacity" (flexibility) of the hypothesis as learning outcomes. To measure this capacity, Vapnik-Chervonenkis dimension (VC) or a property of space hypothesis {f (α)} is used. The value of the VC dimension based on statistical learning theory will determine the value of the error hypotheses on testing data. More specifically, a large error on test data / actual risk R (α) with a probability of 1-η,

0≤η≤ 1, on a dataset consisting of n data can be seen in equation (2.1), where (α) is the error R emp on the training data and h is the VC dimension.

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h\left(log\left(\frac{2l}{h}\right)+1\right) - log\left(\frac{\eta}{4}\right)}{l}}$$

VC confidence value (the value of the second element on the right-hand side (2.1)) is determined by the hypothesis / function of learning outcomes.Thus, the principle of SRM is to find a subset of the space hypotheses chosen, so that the upper limit of the actual risk is minimized by using a subset. SRM aims to minimize the actual risk in a way to minimize the error on the training data as well as the VC confidence. However, the SRM is not implemented by minimizing the equation (2.1) because the VC dimension of the hypothesis space {f (α)} is difficult to quantify and there is little hypothetical model known to calculate its VC dimension. In addition, although the VC dimension can be calculated, it is not easy to minimize equation (2.1), since SRM implementation on SVM is using linear functions.

2.4.3 Fisher Linear Discriminant function

Classification using this method with make the projection from high dimensional data to a line and performs the classification in to one dimensional space, then the projection will compute the distance with the means of the two classed with minimizing the variance on each class. With mazimising over all linear projection as fisher criterion. This technique has strong realation with the linear percepton with optimizing cost function on the training set as the threshold.

2.4.4 Multi Layer Percepton

This network is based on feed-forward neural network is fully connected. The number of nodes in the input layer together with a number of features that are presented by the data, while the number of nodes in the output layer is equal to the number of classes is the map data. At least one hidden layer must be added to

the architecture in order to treat the non-linear separation between classes. Some networks with one and two hidden layers, with different number of nodes in each hidden layer, has been used.

2.4.5 Naïve Bayes Classifier

A Naïve Bayes classifier is a simple probabilistic classification is based on the application of Bayes' theorem (Bayesian statistics) for strong (naive) assumption of independence. A more descriptive term for the underlying probability model to be an independent feature model The advantage of the classification is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) are required for classification. Because of the independent variables is assumed, only the variation of the variables for each class must be determined and not all covariance matrix.

2.5 Feature Selection Techniques

2.5.1 Principle Component Analysis

Variable and feature selection have become the focus of much researches in areas of application for which data sets with tens or hundreds of thousands of variables are available **[11]**. Dimensionality reduction of a feature set is a common preprocessing step used for pattern recognition, classification applications and compression schemes. Principal component analysis (PCA) is one of the popular methods used, and can be shown to be optimal using different optimality criteria **[12]**. However, it has the disadvantage that measurements from all of the original features are used in the projection to the lower dimensional space. This paper proposes a feasible method for dimensionality reduction of a feature set by choosing a subset of the original features, containing most of the essential information, in the same criteria as the PCA. PCA procedure is basically aimed at simplifying the observed variables by means of shrinking (reducing) dimension. This is done by removing the correlations between the independent variables through the

transformation of the independent variables to the origin of new variables that are not correlated at all or commonly referred to as principal component. After several components of the PCA results obtained free multicollinearity, then these components are changed into a new independent variable to regress or analyze its influence on the dependent variable (Y) using regression analysis; with a bit of a factor, the greatest possible variance is X1.

2.5.2 Based on Correlation

The aim of feature selection is to select a subset of the original feature space that is more informative to target class in performing machine learning tasks but to ignore features that are irrelevant and redundant. In this paper, was developed a feature selection algorithm based on information-theoretic measures. Based on the entropy of the designated features, symmetric uncertainty is obtained as a measurement of the relevance of the features offered.

This algorithm mainly consists of two parts to achieve the goal of reducing the dimensions of the original feature space. At first the algorithm removes features that are not relevant to the poor predictive ability to target class. Then the algorithm eliminates redundant features that are correlated with one of the other features. Finally, another feature that is selected is a significant feature that contains the indispensable information about the original set of features.

Given a set of data with a number of input features and the target class, the first algorithm calculates mutual information between features and classes. The algorithm then ranked in order of features in accordance with their degree of association to the target class.Then begins by calculating the strength of the correlation between each pair of features. The total number of mutual information for each feature is obtained by adding all the measures of mutual information together are associated with that feature. To adjust the discriminatory power of the mutual information carried on features and feature-to-class to the same level, we introduce the factor w and is equal to the average of the information-to-class features divided by the average of

the sum of feature information -to-feature. By multiplying w for each size of the feature-to-class, well-to-class features and feature-to-feature ranking to reach the same importance. Finally, the difference of them counted and we just keep those features whose value is greater than zero, which means the feature is selected the most "significant features" that hold information which is indispensable original feature space.

2.5.3 Linear Discriminant Dimensionality Reduction

One method of learning a representative subspace is linear discriminant analysis (LDA) .LDA preferably in face recognition because ble stamp discriminant subspace derived from large scale Training data for classification. LDA seeks such sub-space in which the samples collected from the same class where as samples from different classes are separated, so that the sample well classified **[12]** .

As a preprocessing step of LDA, feature selection plays an important role in selecting the most informative features and complementary to the LDA learning. LDA can do selection and feature subspace learning simultaneously. Inherits the advantages of the value of Fisher and LDA. That is, he was able to find a part of the original features are useful, based on the yield of new features to the feature transformation. relationship between LDA and multivariate linear regression problems, which provides solutions based regressionLDA.

Although based on Fisher criterion, Fisher scores are not able to do a combination of features such as LDA. Features selected by Fisher scores are part of the original features. However, as mentioned previously, change may be more discriminatory features from the original feature. On the other hand, although the LDA recognizes the combination of features, the change all original features and not just those that are useful as the Fisher scores. Additionally, because the LDA to use all the features, the resulting transformation often difficult to interpret. It can be seen that the Fisher score and LDA are actually complementary to a certain extent. If we combine the Fisher score and LDA in systematic way, they can support each

other. One intuitive way is doing Fisher scored before the LDA as a two-stage approach. However, because This two-stage done individually, the entire process will tend suboptimal. This motivates us to integrate the value of Fisher and LDA in a principled how to complement each other.

And the last technique to implement the dimensionality reduction is by using Physical and dynamical theory related Atmoshpere system in the tropic, so exactly the variable who has been selected totally just only related equation. But mostly all variable that selected normally has using by majority scientis in Indonesia, so in the future we can select with the other parameter based on founding in each area.

CHAPTER 3
RESEARCH METODOLOGY

The weather has the characteristics of continuous, multidimensional, dynamic aspects and experiences a complex and chaotic process. In the space, time dimension of weather conditions will never be the same over time; even if conditions may be similar, the essence is a different formation complexity. This weather phenomenon is one of the factors that challenge separate methods to generate accurate weather forecasts. Methods in weather forecasting are developed rapidly. At first, the very conventional method was used, that is by studying the weather conditions earlier and then comparing them with the current conditions to see the trend in the future. Along with the development of science and knowledge in particular areas of mathematics, a new era of weather forecasting methods called numeric concept was introduced.

This process was designed to get accurate weather forecasts by using Neural Network models which are complex, starting from the initial condition data qualification then the data are treated with either dynamic model forecasting models or physical models up to resultant forecasts. Due to the nature of the complexity of weather existing in nature, ANN model assumptions were used in the equation, resulting the model equations of lower accuracy for some conditions. Other conditions that can affect the accuracy of ANN cumulonimbus forecasts are the presence of systematic errors caused by poor data quality initials, thus having a tendency to reduction of the skill. In projections of forecast period, the random error caused by the condition of weather is very dynamic and chaotic, so that at any time it can interfere with the quality of forecasts. In addition, the relatively low resolution cannot capture the parameters in the area that is less than the resolution. This also raises a new problem. To overcome the problem of imperfect numeric weather prediction more advanced post processing methods are required to improve all numeric forecast skill.

This section explains the research using a multi-variable obtained from the radiosonde observation. The process starts from the application designed for getting indices for neural Network based scheme to do a multivariate analysis for short term forecasting and then, the improve of occurrence of Cumulonimbus development is improved. The detail of each part is explained as follows.

3.1 Proposed Calculating Radiosonde Atmosphere Instability Index

Firstly we creates a program by Python 2.7, a process for identifing the data of radiosonde indices based on stability atmosphere was conducted in Jakarta following tropical dynamics theorem **[6][14] [23]**. Then, the nominal data were converted to the numerical values for possible computational analysis. Some index values obtained are very different, and an evaluating algorithm must be developed.

The valid sounding observation was in between from 1 January 2010 – September 2014. The observation was made by the Indonesian Meteorological Climatological Station (BMKG, WMO code 96749), and the operation was launched every 12 hours excluding the operational and analytic problem. All the indices were computed by the algorithm related with the equation based as below and calculated in three different layer, totally found 78 indices. In Table 3.1 shows the 48 based indices, and 30 other indices were used different layer, i.e. Convective Inhibition, Downdraft potential, Energy Helicity Index, Energy helicity, Equilibrium level, hail, lifting condencation level, lifted index, Melting level, Precipitable water in cloud, updraft and Vertical flux.

Table 3.1 Atmosphere Indices.

No	Indices	No	Indices
1	Boyden Index	25	Medium Level Wind: u comp
2	Bul Richardshon number	26	Medium Level Wind: v comp
3	Bulk SFC-850 Shear	27	Mean relative humidity in the first 500mb
4	CAP's Theta_es difference	28	Maximum Buoyancy
5	Convect. Availab. Pot. Energy	29	Most unstable parcel
6	Convective Inhibition	30	Planetary boundary layer estimation height
7	Temperature difference @ 500 hPa	31	Precipitable water in cloud
8	Temperature difference @ Tp=-15 C	32	Precipitable water in environment
9	Downdraft Potential	33	Relative Helicity of storm
10	Energy Helicity Index of storm	34	Severe WEAther Threat
11	Euqilibrium Level	35	Severe weather index local
12	Maximum Hail diameter	36	Wind Shear 12km
13	High Level Wind: u comp	37	Wind Shear 3km
14	High Level Wind: v comp	38	Showalter Index
15	High relatif humidiy mean	39	Cloud base LCL temperature
16	Helicity of first 3 km	40	MP equivalent temperatur
17	K Index"	41	core of updraft
18	Lifting Condensation Level	42	mean water vapor v horizontal flux
19	Level of Free Convection	43	wet buld zero height
20	Lifted Index	44	mean bouyancy accelaration of the firswt 250 hpa
21	Low Level Wind: u comp.	45	MUP height
22	Low Level Wind: v comp.	46	Boyden Index at Tv

No	Indices	No	Indices
23	Mean relative humidity in the first 250mb	47	bulk richardson at Tv
24	Meltinglevel (parcel at 0C)	48	convective available potential

3.2 Proposed Detection of Cumulonimbus Events

3.2.1 Using Cloud Type Classification Based on JMA Data [8]

The 12-hour period was associated with the sounding derived indices as predictor variables. The predicted variables were built using Cloud Type identification MTSAT and validated using the weather observation report (METAR and Synoptic report).

Manual observation in Cengkareng Meteorological Station was done for 24 hours per day, with some equipments and legal report exchanges to all meteorological stations in the world under the World Meteorological Organization (WMO). A METAR weather report is predominantly used by pilots in fulfillment of a part of a pre-flight weather briefing, and by meteorologists, who use aggregated METAR information to assist in weather forecasting. Raw METAR is the most popular format in the world for the transmission of observational weather data. It is highly standardized through the International Civil Aviation Organization, which allows it to be understood throughout most of the world.

The second application program was developed by using Python 2.7 to identify the Cb event using combination of data observation from MTSAT channel. The threshold for classifying the Cb is less than -45°C. The application was developed to identify cumulonimbus in Jakarta-Indonesia area with the high convectivty, and calculate range area starting from 4 km, 8 km, 12 km, 20km, 40 km, 80 km, and 200 km.

Since the top of a well developed Cb structure reaches top layer of atmospher, the amount of water vapor in the air column between the cloud top an a satellite is

negligibly small **[6].** Thus, it is considered that the radiance observed with a thermal window channel (11μm band, denoted as IR1) and water vapor channel (6.7 μm band, denoted as WV) is nearly equal. To theoretically confirm, the equivalent blackbody temperature to be observed from satellite in each channel was computed for the atmosphere model of the tropical latitude at rainy season, computed by placing a black body cloud starting from 7 km at 500 meter increments up to 13 km (tropopause). At the level of 1 km below the tropopause, the difference becomes 1.5K. Thus, 1.5K was adopted as a threshold for Cb.

3.2.2 Proposed Cloud Type Classification using Neural Networks

Besides trying to find optimal sounding indices to predict the Cb in short term forecast, this part also tries to estimate the cloud type by using MTSAT channel in tropical area. We used 7 features to develop this process, like shown in Table 3.2.

Table 3.2 Features description for Cloud Type MTSAT ANN.

No	Parameters	Description
1	IR1	Top brightness temperature of IR1
2	IR2	Top brightness temperature of IR2
3	WV	Top brightness temperature of WV
4	IR4	Top brightness temperature of IR4
5	IR1-IR2	Split window technique (current method)
6	IR1-WV	Split window technique
7	IR2-WV	Split window technique

According to the cloud type classification by JMA **[16]** there are eight classes, i.e. Clear, Cu, Sc, St, Cb, Dense, Midle, High, commonly defined. However, in this study we only employed six classes ,i.e. Clear, Low, Dense, Cb, Midle, and High classification. Based on the observation data represented by synoptic report in 20 locations around Indonesia, we found some cloud types were recorded in every 3

hours, within 6 months in between Januari 2013 – Juni 2013. During the observation, some places experienced changing seasons, rainy season and also dry season.

Each observation data represents one pixel on MTSAT data around 5 km. The data were splitted 60 % into training, 20 % tested and 20 for validation using neural network classifier. The problem of predicting the occurences of Cumulonimbus cloud was transformed the problem of classify the indices into two classes relevant and irrelevant to the occurrence of Cb cloud. Due to different amount of data in each class, the problem of data imbalance was indeed. The methods of Bootstrap and Adaboost Algorithm was adopted with a minor modification to cope with imbalanced data.

3.3 Imbalanced Data Problem

In recent years, there are some techniques applied to overcome those problems such as Adaboost, SMOTEBoost, Borderline SMOTE, RAMO Boost, bootstrap resampling with adaboost technique **[17].**

After making some experiments, we found the most effective technique to overcome this research problem is Bootsrap re-sampling with Adaboost technique **[18]**, so we used this technique eventually.

These are the details of step for following actions:

a. Locating each sub-cluster

By using a technique from self organizing map, we calculated each cluster from minority and majority class. Because the number of sub-cluster in high dimensional data was unknowm, so the neurons number in SOM was employed to initialate from the experiment.

b. Computing direction and width of each sub-cluster

Calculating the outskrits from each sub-cluster was done to represent the syntetic data, by finding the size dan shape. With eigenvectors, we identified the shape and egienvalues in each dimension for representing the size.

c. Identifying boundary data of each sub-cluster

To calculate the distance beetwen two sub-clusters of different classes, we used the Hausdorff distance. In specified dimension, each sub cluster is viewed as a set of vectors, by euclidean distance measure spliting from the shortest distance between two majority and minimum classes in each boundary layer.

3.4 Data normalizaiton

Data normalization is often performed before the training process begins. When nonlinear transfer functions are used at the output nodes, the desired output values must be transformed to the range of the actual output of the network **[19]**. Even if a linear output transfer function is used, it may still be advantageous to standardize the outputs as well as the input to avoid computational problems, to meet algorithm requirement and to facilitate network learning.

The following is the normalization used in this work:

$$x_n = (x_0 - x_{min})/(x_{max} - x_{min})$$

For input normalization of sounding indices data, a channel is defined as a set of elements in the same position over all input vectors in the training or test set. That is, each channel can be thought of as an independent input variable.

3.5 Feature selection process

This research purposed some methods to reduce the features sizes. We try to implemented Principle component analysis to create new variable that represented 85 % variables of dataset. To compare the accuracy for classifying process then the features also evaluated with Liniear discriminant analysis, Correlation and selection based on dynamics theory.

3.6 Classifier Techniques

This research separated in two majority problems for classification:

1. The first experiment Implemented some technique to select best neural network to generate cloud type from MTSAT data, later we will use the best algorithm to predict the Cb occurences by using radiosonde indices. In the experiments many classifier have been implemented to find the best one. Finally in this report we put highest classifier with good result, such as; support vector machine, Multi Layer Percepton, Radial Basis network, Simple logistic, random tree and naïve bayes classifiers. After measure the performance so the algorithm will implemented in big data to reach the analysis problem related with the cumulonimbus and atmosphere indices.
2. The second experiment try to classify the Cumulonimbus occurences by using Radiosonde Indices, same as before some classifier technique **[6][20][21][24]** also implemented in this research. But finally we just presented the most majority result with the good accuracy. These are Radial Basis function network, Supper Vector Machines and Random Forest classifier. Mostly all thus technique has been implemented in othe research before, so can be good compariton during the result and analysis.

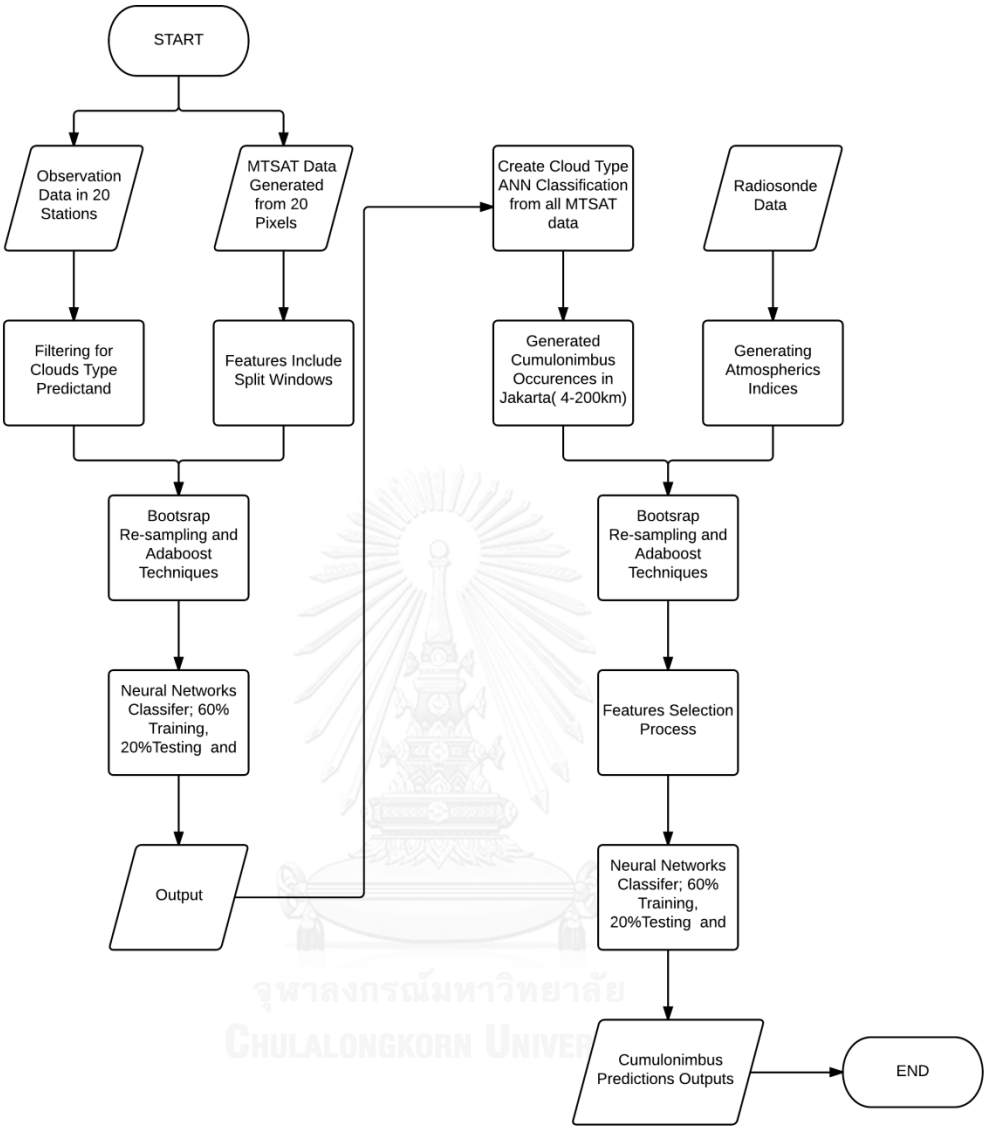

Figure 3. 1 Cumulonimbus prediction research flowchart

From the Figure 3.1 shows the experiment were splited in two different parts, with maximizing the generated predictand by MTSAT channels we can found better Cumulonimbus indentification that later we developed the prediction model in the ranges of Jakarta area.

3.7 Assesment Metrics

In this research, the performance of the classifier evaluated by the following measures, i.e. Overall accuracy, F-measures, True Positive Rate, True negative rate, Precision, and Recall.

$$\text{Overall accuracy (OA)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FN} + \text{TN}}$$

$$\text{TP Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FP Rate} = \frac{\text{FP}}{\text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{FP}}{\text{FP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F} - \text{Measure} = \frac{(1 + \beta^2) \times \text{Recall} \times \text{Precision}}{\beta^2 \times \text{Recall} + \text{Precision}}$$

CHAPTER 4
EXPERIMENTS

At this chapter we setup three different sections for measure the best classifer to predict the Cumulonimbus occurences in some classes,i.e. 4km, 8km, 12km, 20km, 40 km, 80km and 200km sizes area with times accumulated predictand in every 3 hours,i.e. 0 UTC, 3 UTC, 6, UTC, 9 UTC and 12 UTC. But before select the best predictand data, that generated from MTSAT satellite **[22]**. We will choose the best algorithm that produced Cumulonimbus data from cloud type MTSAT classification in Jakarta. So the experiments flow was like below:

- Create the new algorithm for MTSAT Cloud type classification to generate Cumulonimbus data by using Neural Network
- Develop a Cumulonimbus prediction by using selected Radiosonde indices in Jakarta, Indonesia
- Radiosonde Indices analysis related the Cumulonimbus development

4.1 MTSAT cloud type classification using Artificial Neural Network

The challenge in this research is the use of remote sensing to determine the existence of cumulonimbus cloud activities. In the first step, we used the data from the Japan Meterology Agency (JMA) that has been developed for a long time **[8]** and still used up to now. This algorithm only employs the difference among few channels on MTSAT to determine the clouds types as well as the presence of Cumulonimbus.

MTSAT satellite has been practically used in Indonesia because it orbits in the surrounding eastern part of Indonesia. Then, our hypothesis was used to match the Cumulonimbus events with Radiosonde data in observation station in Jakarta. A good validation value was obtained by doing random checking on the sample of

cumulonimbus clouds from JMA cloud types. The result showed about 75% of the sample were in accordance with cumulonimbus clouds as recorded by the METAR Soekarno-hatta airport in Jakarta. However, there was a weakness in this validation value because only recorded Cb parameters were used to estimate the existence of the cloud around observation areas.

In this research, we also studied wheter the surface observation data can be used to form a new classification for determining the types of clouds. By using the data from 20 different observation stations around the country, to a new technique was developed by using a combination of Artificial Neural Network to replace the old ones namely the Split window and threshold based.

There are eight cloud type models issued by JMA which are clear, Cb, Sc, Cu, Dense, CH, CM, ST like in Figure 4.1. At previous discussion on the offect of Cb cloud, our study focused on predictiong the possible formation of Cb cloud in radiosonde experiment. To easily obtain the sample from observation data at the station, classes Sc, St and Cu clouds were merged into low clouds category, while Dense remained separated from the Cb cloud to see the possibility of clouds growth that can trigger thunderstrom. And keep three others types Clear, Middle and High. Data from synoptic observation were gathered in every 3 hours and the filtering with 4 oktas boundary was done for each cloud type. Therefore, in those hours, one type of clouds, or sometimes in conjunction with other cloud types would be determined.

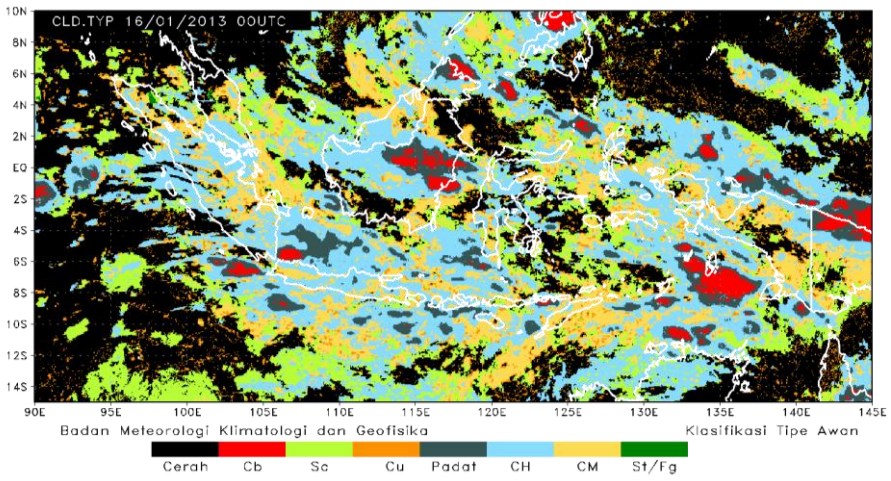

Figure 4. 1 An example of cloud type image **[22]**.

Data used in this experiment were taken once from from synoptic database in BMKG during two seasons of the same year, during January-June 2013. Data selection were used to separate those that are eligible from those that are not eligible.

4.1.1 Features and Samples

The data samples generated from meteorological station under BMKG with 24 hours operation are as follows.

Table 4. 1 Observation stations list.

No	Station	No	Station
1	Banda Aceh	11	Pontianak
2	Medan	12	Palangkaraya
3	Padang	13	Balikpapan
4	Palembang	14	Makassar
5	Batam	15	Manado
6	Lampung	16	Ambon
7	Jakarta	17	Ternate
8	Semarang	18	Biak
9	Surabaya	19	Jayapura
10	Denpasar	20	Kupang

We collected 16400 samples without a blank from all above stations. Presumably representing all characteritics of Indonesian climate., With three different conditions based on topografy reason between MTSAT channel and surface. By excluding visible channel which only can be used in the daytime, all samples showed that different length of temperature between day and night cannot be calculated.

As a predictor, we used 4 MTSAT channels (IR1, IR2, WV, IR4) and the combination in split window technique. Since the treshold **[15]** is unclear to be used

in Indonesia, the optimizing value is needed to select best feature to be used as input variable for artificial neural network model. After conducting feature test, some variables having similar direction were not included. Finally, 7 features were selected as input for training data.

The connection between split window technique and IR1 shows a strong interaction to detect some cloud types in terms of convectivity. We can check the scoring for cumulonimbus that has the special character. Expectedly this will be helpfull when the training data are applied together with the machine learning, also for some other cloud types. Even in this research, some low clouds (Cu, Sc and St) were merged to be one type. The cumulonimbus cloud is actually categorized as low level cloud, but because it can reach high cloud level, so we make an exception for this cloud type.

Simple method of split window does yield a deeper yet easier understanding on the Cb cloud characteristics. In fact, this condition sometimes is confusing and complicated for cloud events in Indonesian regions because there are a number of error values or wrong estimation for the Cb cloud criteria due to influencing topographic and meteorological factors. Thus, optimization using neural network is suggested to obtain more optimal estimation results.

4.1.2 Cloud Type Classification by Neural network

All these training data for this experiment were averaged by five standards of 10-fold cross validation. At the first part of the experiment, the data set are from the original data sampled. It was used without any modification, only trained by neural network classifier and some different techniques to measure the accuracy.

By using some different techniques like SVM, RBF, multi layer percepton, FLDA, simple logistic, and random tree classifiers, the experiment was carried out to know which one is the best . like shown in Table 4 the resul of implementation classifier technique with the original data. So the patterns from the cloud occurences and others was so large different. But for some type of cloud these is still work very well.

Table 4. 2 Result of data set classification with original data.

Cb								
Method	**Overall Accuracy**	**TP Rate Min**	**TP Rate Max**	**Recall**	**F-Measure Min**	**F - Measure Max**	**MCC**	**ROC area**
Fisher's Linear Discriminant function.	0.987	1.000	0.977	0.978	0.735	0.988	0.753	0.997
Supper Vector Machine	0.991	0.924	0.993	0.990	0.856	0.995	0.853	0.958
Multi Layer Percepton	0.992	0.905	0.995	0.992	0.874	0.996	0.870	0.998
Radial Basis Function Network	0.992	0.908	0.994	0.991	0.865	0.996	0.862	0.998
Simple Logistic	**0.993**	0.908	0.995	0.993	0.881	0.996	0.878	0.998
RandomTree	0.989	0.923	0.995	0.989	0.827	0.989	0.822	0.909
Naivebayes	0.987	1.000	0.987	0.979	0.743	0.989	0.760	0.998
Clear								
Method	**Overall Accuracy**	**TP Rate Min**	**TP Rate Max**	**Recall**	**F-Measure Min**	**F - Measure Max**	**MCC**	**ROC area**
Fisher's Linear Discriminant function.	0.861	0.981	0.691	0.794	0.772	0.812	0.647	0.936
Supper Vector Machine	0.926	0.925	0.923	0.924	0.896	0.940	0.837	0.924
Multi Layer Percepton	0.928	0.929	0.925	0.926	0.900	0.925	0.842	0.974
Radial Basis Function Network	0.925	0.921	0.924	0.923	0.895	0.939	0.835	0.972
Simple Logistic	0.925	0.908	0.933	0.924	0.895	0.941	0.836	0.973
RandomTree	0.903	0.863	0.924	0.903	0.863	0.924	0.788	0.894
Naivebayes	0.905	0.979	0.826	0.881	0.854	0.899	0.773	0.960
Dense								
Method	**Overall Accuracy**	**TP Rate Min**	**TP Rate Max**	**Recall**	**F-Measure Min**	**F - Measure Max**	**MCC**	**ROC area**
Fisher's Linear Discriminant function.	0.983	1.000	0.960	0.961	0.600	0.979	0.641	0.971
Supper Vector Machine	0.974	0.470	0.993	0.977	0.547	0.988	0.544	0.731
Multi Layer Percepton	0.971	0.000	1.000	0.971	0.000	0.985	0.000	0.977
Radial Basis Function Network	0.897	0.076	1.000	0.973	0.140	0.986	0.258	0.986
Simple Logistic	0.975	0.486	0.992	0.977	0.554	0.988	0.548	0.988
RandomTree	0.985	0.756	0.992	0.985	0.749	0.992	0.742	0.874
NaiveBayes	0.984	0.996	0.963	0.964	0.615	0.981	0.653	0.977
High								

Method	Overall Accuracy	TP Rate Min	TP Rate Max	Recall	F-Measure Min	F - Measure Max	MCC	ROC area
Fisher's Linear Discriminant function.	0.824	0.745	0.786	0.786	0.615	0.852	0.487	0.848
Supper Vector Machine	0.827	0.426	0.959	0.836	0.544	0.819	0.481	0.692
Multi Layer Percepton	0.853	0.645	0.919	0.856	0.673	0.908	0.582	0.908
Radial Basis Function Network	0.851	0.603	0.933	0.857	0.659	0.910	0.574	0.907
Simple Logistic	0.795	0.339	0.953	0.812	0.453	0.887	0.386	0.849
RandomTree	0.831	0.634	0.890	0.831	0.633	0.890	0.523	0.762
NaiveBayes	0.794	0.672	0.779	0.754	0.556	0.830	0.404	0.793

Low

Method	Overall Accuracy	TP Rate Min	TP Rate Max	Recall	F-Measure Min	F - Measure Max	MCC	ROC area
Fisher's Linear Discriminant function.	0.687	0.711	0.406	0.479	0.396	0.543	0.103	0.579
Supper Vector Machine	0.578	0.000	1.000	0.760	0.000	0.864	0.000	0.500
Multi Layer Percepton	0.791	0.339	0.953	0.806	0.456	0.882	0.388	0.821
Radial Basis Function Network	0.795	0.407	0.938	0.810	0.507	0.883	0.418	0.846
Simple Logistic	0.578	0.000	1.000	0.760	0.000	0.864	0.000	0.494
RandomTree	0.798	0.572	0.871	0.799	0.577	0.869	0.446	0.722
NaiveBayes	0.798	0.932	0.385	0.516	0.480	0.547	0.292	0.771

Middle

Method	Overall Accuracy	TP Rate Min	TP Rate Max	Recall	F-Measure Min	F - Measure Max	MCC	ROC area
Fisher's Linear Discriminant function.	0.687	0.711	0.406	0.479	0.396	0.543	0.103	0.579
Supper Vector Machine	0.578	0.000	1.000	0.760	0.000	0.864	0.000	0.500
Multi Layer Percepton	0.791	0.339	0.953	0.806	0.456	0.882	0.388	0.821
Radial Basis Function Network	0.795	0.407	0.938	0.810	0.507	0.883	0.418	0.846
Simple Logistic	0.578	0.000	1.000	0.760	0.000	0.864	0.000	0.494
RandomTree	0.798	0.572	0.871	0.799	0.577	0.869	0.446	0.722
NaiveBayes	0.798	0.932	0.385	0.516	0.480	0.547	0.292	0.771

Like in Cb and dense classes, the accuracy shown more than 0.9 its mean the original features can perform good classification in that classes. It should be noted

that these two classes have specific characters that has a very small differences between top bright temperatures and split window technique

But finally to improve the accuracy we also tried the other experiment with using bootsrap resampling adaboost technique. In the Table 5 we can found the result by implemented those technique using SVM classifier selected as most efficient to this problem.

Table 4. 3 Classification result with bootsrap re-ssampling and adaboost technique for imbalance data sets.

Method	Overall Accuracy	TP Rate Min	TP Rate Max	Recall	F-Measure Min	F - Measure Max	MCC	ROC area
CB								
AdaBoostM1	0.992	0.958	0.992	0.991	0.870	0.995	0.869	0.997
Bootsrap and AdaBoost	**0.994**	0.960	0.994	0.994	0.942	0.995	0.937	0.997
Clear								
AdaBoostM1	0.928	0.964	0.896	0.920	0.895	0.935	0.837	0.956
Bootsrap and AdaBoost	**0.941**	0.962	0.919	0.941	0.945	0.937	0.883	0.979
Dense								
AdaBoostM1	0.988	0.772	0.995	0.988	0.792	0.994	0.786	0.991
Bootsrap and AdaBoost	**0.990**	0.874	0.994	0.990	0.787	0.994	0.781	0.970
High								
AdaBoostM1	0.852	0.666	0.909	0.853	0.675	0.905	0.581	0.906
Bootsrap and AdaBoost	**0.920**	0.815	0.962	0.922	0.642	0.956	0.599	0.932

Method	Overall Accuracy	TP Rate Min	TP Rate Max	Recall	F-Measure Min	F - Measure Max	MCC	ROC area

Low								
AdaBoostM1	0.812	0.583	0.890	0.816	0.603	0.881	0.484	0.863
Bootsrap and AdaBoost	**0.938**	0.845	0.971	0.940	0.671	0.967	0.638	0.922
Middle								
AdaBoostM1	0.812	0.583	0.890	0.816	0.603	0.814	0.484	0.863
Bootsrap and AdaBoost	**0.963**	0.742	0.982	0.964	0.756	0.980	0.736	0.957

By using this optimizion, we could find the best result with overall accuary is more than 90 % and the highest value is Cb class with bootstrap and adaboost technique of SVM 0.994 for the overall accuracy. Thus, we tried to implement this method to overall data in this research during Januari 2010 – September 2014 and then compared it to the current algorithm presented by JMA. The model must be validated by using other manual observation in different locations. We also tried to validate it in higher latitude because the sample was only generated from the tropical area.

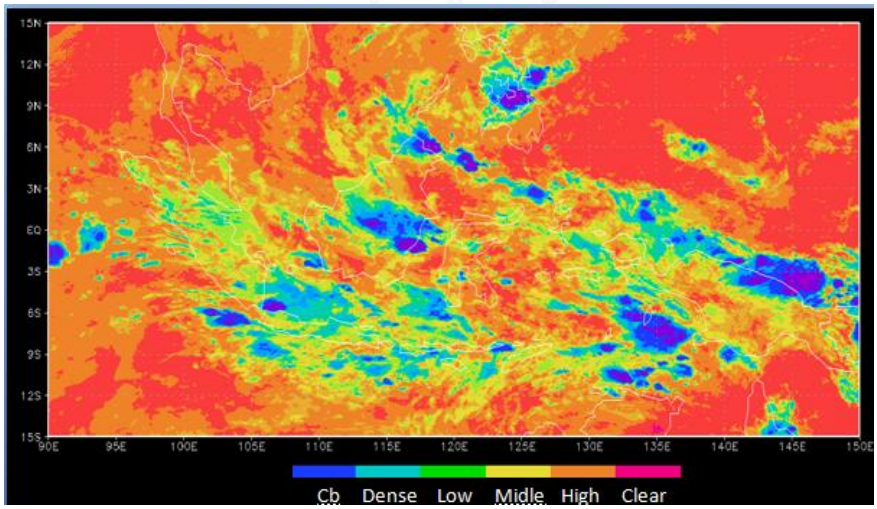

Figure 4. 2 An example of new cloud type tlassification using ANN.

In Figure 4.3 the new algorithm shown that this method can be applied on the large grid of data. With more specific on detection in the low cloud level this application will usefull for early warning information

4.1.3 Comparison between JMA Cloud Type and New Proposed Cloud Type

To see whether the experiment of determining the new type of cloud can give better results, a comparison was done between a result obtained using a method of JMA MTSAT-2 products and algorithms developed in this research. The model used to generate the latest cloud classification was experimented by applying the imbalanced data resolution, using resampling techniques, adaboost technique, as well as using SVM classifier as the results of the experiment showed the best values. At first, we used 7 different classifiers. But in the implementation, those seven neural nets ran very slow. Finally, it was decided to use the SVM to classify the types of clouds, so that history data could be run and then generated for comparison. In this case, only 5 classes (Cb, Low, Middle, High and Clear) were compared according to the surface observations on Cengkareng meteorological station during July 2013 – July 2014. Since the running process was very low because using big data, then for comparison used the limited data.

Table 4. 4 Accuracy rates comparison of JMA algorithm and New purposed.

Type	JMA	New
Cb	86.49	98.80
Low	75.06	94.62
Middle	82.72	91.37
High	76.82	89.49
Clear	82.15	95.70

In Table 4.4 we know comparison of the accuracy in Jakarta between old algorithm and the new purposed. The JMA algorithm is actually quite good, but it still ranges below 90%. Therefore, an experiment was conducted to improve the

value to get more optimum accuracy. Then, a better value with even accuracy was obtained, reaching up to 98.80 for the Cb cloud type. For this study to classify Cumulonimbus occurrence finally we used the new cloud type algorithm.

4.2 Cumulonimbus Classification

By limiting the number of input in the neural network classifier, we built the model with the hidden neurons. Then, we selected the member of input in the training data to optimize the result of the model. We also implemented the feature selection technique to choose the best variable to the neural network. In addition, we also ran the feature extraction by principle component analysis technique to improve the accuracy score. Since the model found the imbalanced data problem, similar to the first part, we also implemented bootsrap resampling and adaboost technique to solve the problem.

The classification was too large with 5 different time observations (1-3 hours, 4- 6 hours, 7- 9 hours and 10-12 hours observations) and 7 different target areas to prove the size of the model can be workable, starting from 4 km, 8 km, 12 km, 20 km, 40 km, 80 km and 200 km. These areas also represent the topographyc condition around Jakarta. When needed for early warning system, the model can be applied not only for aerodrome forecast.

In the experiment we did previously when the application of radial basis function used training data limited in one season and the result indicated lower accuracy. This is due to the use of classification limited to RBF with Gaussian function without considering the differences in the data, in which there was an imbalance between variables with real value and wrong variable. Though feature selection has been applied with PCA to form a new variable with a good representation, but the result is not optimal yet.

Then in the next experiment, besides using feature selection technique with the PCA, we also applied the selection data using the optimum value through the Pearson Correlation, LDA and also feature that has been selected based on dynamic

reasons in accordance with the existing air lability theory. However, the result obtained using either the whole data or selected variable remains insufficient.

The solution of imbalanced data problem has been extensively performed in numerous previous studies. Because in this study a fairly important problem was found in imbalanced data, then some techniques were explored to obtain optimum classification result. Based on the literature study, botstrap and adaboostM1 techniques were chosen, along with the method applied with a combination of resampling and adaboost technique.

The result obtained shows a significant change in the classification model using training data directly after going through resampling and then was compared to the value through the first feature selection. Then, it was obtained that resampling and adaboost techniques demonstrated value and speed which are much more optimal than other techniques. Thus, those techniques were selected to be applied to the overall data.

The main feature selection in this study is the use of principal component analysis (PCA). The interpretation of the principal components is that it is a new axis system in vector spaces with many dimensions where many variables are observed. Through the principal components, cross axis has experienced modification in the scale and has been rotated to get variance ordered from the smaller one and orthogonal.

If the variance of the observed variables affects the weights or the principal component coefficients, then the main components analysis can be performed using covariance matrix. Variance is information from the observed variable which means when a variable has the same value from all the observations done, then the variable has no information which can distinguish it among all observations.

The principal component is a set of new variables which are a linear combination of the observed variables. It has a variance that is increasingly smaller. Most of the variance (diversity or information) in a set of observed variables tend to converge on a few first main components, leaving only a little information from the

original variables on the last major component. This means that the principal components on the last sequence can be omitted without losing much information. In this way, the principal component analysis can be used to reduce variables.

For the purposes of reducing variables, it is necessarily to determine how many major components that must be taken. There are several ways to determine how many primary components that should be taken. By using the scree plot that many components taken are on the point curve that is no longer sharply declining. But at this research we were used a cumulative proportion of variance to the total variance.

It has been explained above that among the main components are orthogonal which means that every principal component is representative of all the original variables, so that the principal components can be a replacement of original variables as long as an analysis toward the variable requires an orthogonality. In the multiple linear regression analysis, it requires an absence of multicollinierity among its free variables. If multicollinierity is found in the relevant data, thus the principal components can be used as a substitute for the free variables in the regression model.

In the principal components analysis, some following measurements are obtainen. The value of the total variance is the information from the original variables that can be explained from its principal components. The variance proportion of principle component to k to the total variance indicates the percentage of original variable information contained within the principal components to-k. And correlation coefficient value between the principal components and the related variable

In addition to the PCA, an experiment was also carried out using corelation technique by considering the most optimal relationship on each variable. After that, the limit was specified with correlation was so small. Furthermore, a new variable with Liniear projection Discriminant analysis (LDA) was also chosen, but the result has not yet passed the value for a better accuracy of the PCA. The variable selection based on phisical theory.

4.2.1 Classification Results

The experiment employed some neural network techniques as the best classifier measured by RBF network, SVM and random forest technique. They were combined with feature selection technique by using corelation, dynamics selection, as well as feature extraction through principle component analysis and linear discriminant analysis. In addition, we also tried the adaboostM1 technique for imbalanced data and compared the data with bootstrap resampling and adaboost. As a comparison to test the maximum value between some feature selection techniques and as a problem solving of imbalanced data. From a dynamic selection technique, we figured out that this technique seems appropriate when used with a random forest, while LDA is also able to produce up to 90 percent accuracy with the random forest. This is almost the same as the selection using LDA and correlation techniques, but these values are not stable over time in each class.

Here is the experiment that obtained the most optimum result in general. It employed SVM classification and applied PCA that was previously processed using the bootstrap resampling with a combination of adaboost technique. Because the display of the results is very long, then every class in this chapter is indicated by the above method. In the future, if the application of each neural net shows better results, then every developed model may have different neural net in every classes.

Table 4. 5 Result of SVM with bootstrap re-sampling and adaboost technique **[18]** for imbalanced data, using PCA as feature selection processed.

Class	Overall Accuracy	TP Rate Min	TP Rate Max	Recall Min	Recall Max	F-Measure Min	F - Measure Max	ROC area
Time0Size4	0.972	0.984	0.963	0.984	0.963	0.968	0.963	0.994
Time0Size8	0.943	0.922	0.952	0.922	0.952	0.906	0.952	0.976
Time0Size12	0.943	0.922	0.952	0.922	0.952	0.906	0.952	0.976
Time0Size20	0.887	0.857	0.900	0.857	0.900	0.834	0.900	0.932
Time0Size40	0.895	0.889	0.893	0.889	0.893	0.853	0.893	0.930

Class	Overall Accuracy	TP Rate Min	TP Rate Max	Recall Min	Recall Max	F-Measure Min	F - Measure Max	ROC area
Time0Size80	0.885	0.868	0.889	0.868	0.889	0.835	0.889	0.932
Time0Size200	0.845	0.805	0.868	0.805	0.868	0.797	0.868	0.891
Time3_Size4	0.935	0.941	0.928	0.941	0.928	0.898	0.949	0.976
Time3_Size8	0.930	0.935	0.923	0.935	0.923	0.895	0.944	0.967
Time3_Size12	0.892	0.882	0.893	0.882	0.893	0.854	0.911	0.942
Time3_Size20	0.867	0.816	0.889	0.816	0.889	0.797	0.899	0.917
Time3_Size40	0.824	0.748	0.856	0.748	0.856	0.729	0.867	0.879
Time3_Size80	0.912	0.898	0.920	0.898	0.920	0.884	0.929	0.959
Time3_Size200	0.847	0.813	0.858	0.813	0.858	0.781	0.877	0.909
Time6_Size4	0.957	0.974	0.945	0.974	0.945	0.933	0.965	0.979
Time6_Size8	0.960	0.973	0.950	0.973	0.950	0.942	0.967	0.986
Time6_Size12	0.918	0.917	0.912	0.917	0.912	0.871	0.935	0.955
Time6_Size20	0.926	0.929	0.920	0.929	0.920	0.887	0.941	0.965
Time6_Size40	0.927	0.934	0.916	0.934	0.916	0.882	0.941	0.965
Time6_Size80	0.932	0.936	0.924	0.936	0.924	0.891	0.946	0.966
Time6_Size200	0.929	0.931	0.924	0.931	0.924	0.894	0.943	0.968
Time9_Size4	0.900	0.889	0.903	0.889	0.903	0.863	0.918	0.946
Time9_Size8	0.884	0.876	0.884	0.876	0.884	0.845	0.904	0.942
Time9_Size12	0.870	0.914	0.822	0.914	0.822	0.861	0.867	0.918
Time9_Size20	0.867	0.911	0.818	0.911	0.818	0.861	0.862	0.927
Time9_Size40	0.882	0.923	0.837	0.923	0.837	0.878	0.878	0.943
Time9_Size80	0.859	0.879	0.838	0.879	0.838	0.841	0.868	0.928
Time9_Size200	0.869	0.921	0.808	0.921	0.808	0.871	0.856	0.931
Time12_Size4	0.898	0.908	0.888	0.908	0.888	0.883	0.908	0.953
Time12_Size8	0.891	0.918	0.863	0.918	0.863	0.878	0.895	0.939
Time12_Size12	0.887	0.944	0.812	0.944	0.812	0.898	0.864	0.938
Time12_Size20	0.887	0.933	0.825	0.933	0.825	0.900	0.864	0.934
Time12_Size40	0.869	0.938	0.763	0.938	0.763	0.891	0.825	0.913
Time12_Size80	0.890	0.949	0.805	0.949	0.805	0.906	0.859	0.926
Time12_Size200	0.871	0.963	0.680	0.963	0.680	0.905	0.776	0.893

From the result above in Table 4.5, we know the most effective combination technique for the model gives best prediction in near size and time, while larger area shows the weakness point. However, the longest time prediction is still strong with more than 85% accuracy. This means the model can be improved to predict the cumulonimbus cloud.

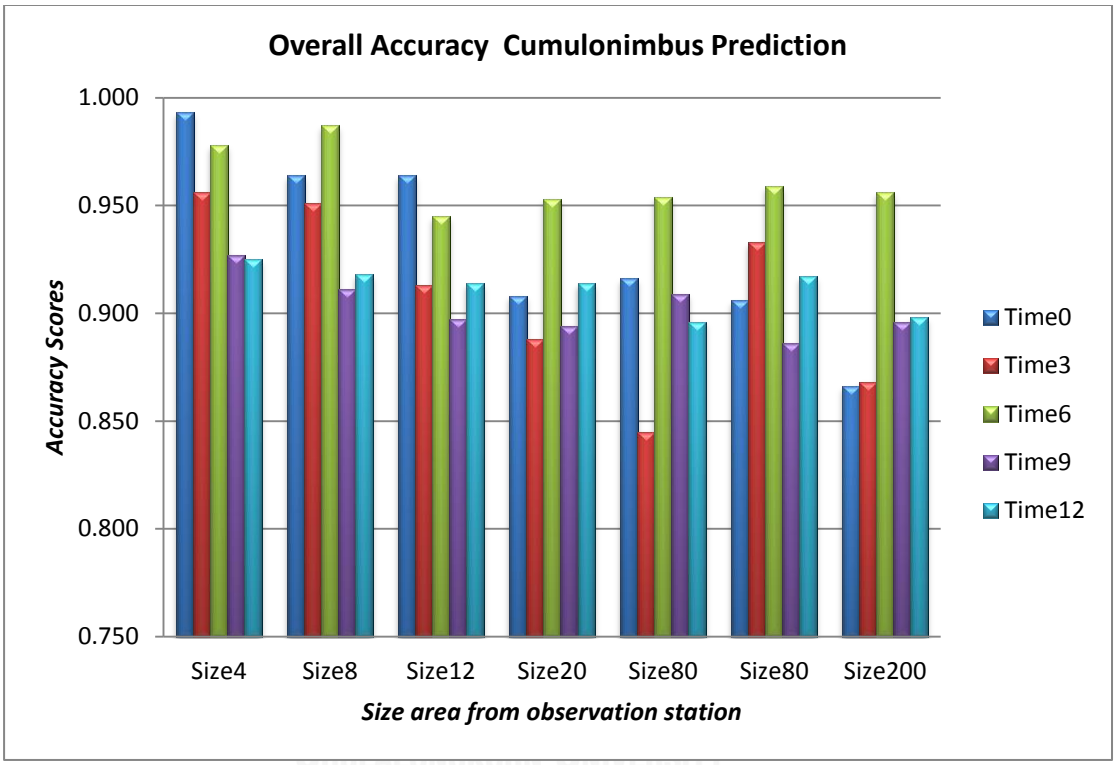

Figure 4. 3 Accuracy during in time prediction and size area.

This experiment is as the first step in gaining a new understanding of radiosonde utilization to predict Cb. Therefore, its usage is not limited to the nowcasting only, but also for short-term weather forecasting. As shown in Figure 4.4 the predicting accuracy became lower depend on the time scale and lost when the size also larger. But considering the high accuracy which is above 85%, the result of this model is quite promising. After going through various, long-time experiments, then the result obtained is quite satisfying.

4.2.2 Best indices for Predicting Cumulonimbus

Though the feature selection in each class possesses each distinguished character of which 78 variables do not have the same aggregations to predict weather in each class, in general it was figured out some 15 variables that most frequently appear and have the largest distribution in the PCA to predict Cb. The indices are namely:

Table 4. 6 Mostly variables shown up on PCA selection in all classes.

No	Indices
1	CAP's Theta_es difference
2	Convect. Availab. Pot. Energy
3	Convective INhibition
4	Downdraft Potential
5	Energy Helicity Index of storm
6	Euqilibrium Level
7	High Level Wind: u comp
8	K Index"
9	Lifting Condensation Level
10	Level of Free Convection
11	Lifted Index
12	Precipitable water in cloud
13	Relative Helicity of storm
14	Severe WEAther Threat
15	Showalter Index

This model used the selected variable from the Cb predictand generated after PCA, it is assumed that the results in Table 4.6 is the best indices that can used for Cb prediction. Every classes has different variable to make their classifier model so this variables selecty was pretty complex. From the result in Jakarta we found that

the development of a new classification of clouds has a good result, then it is likely the model will have a more optimal degree of accuracy.

4.3 Statitistic Analytics of Predictors

In this part, we will introduce the characteristics as well as some important indices related to atmosphere dynamics theory. By performing the threshold during cumulonimbus activity and analysing it using box and whiskers diagrams, the most parameters for the artificial neural network would be relevant to predict this event.

The distribution in Figure 4.5 was separated by considering differences of time, and accumulating the range of the area which was noted at 4 – 200 km. The indices variable will give the reasonable scoring to each case of the cumulonimbus. The data observation will be separated into two different times, namely 00.00 UTC and 12.00 UTC. With two different characters of atmosphere from morning and evening, this research will be very useful to forecast other severe weather in the future.

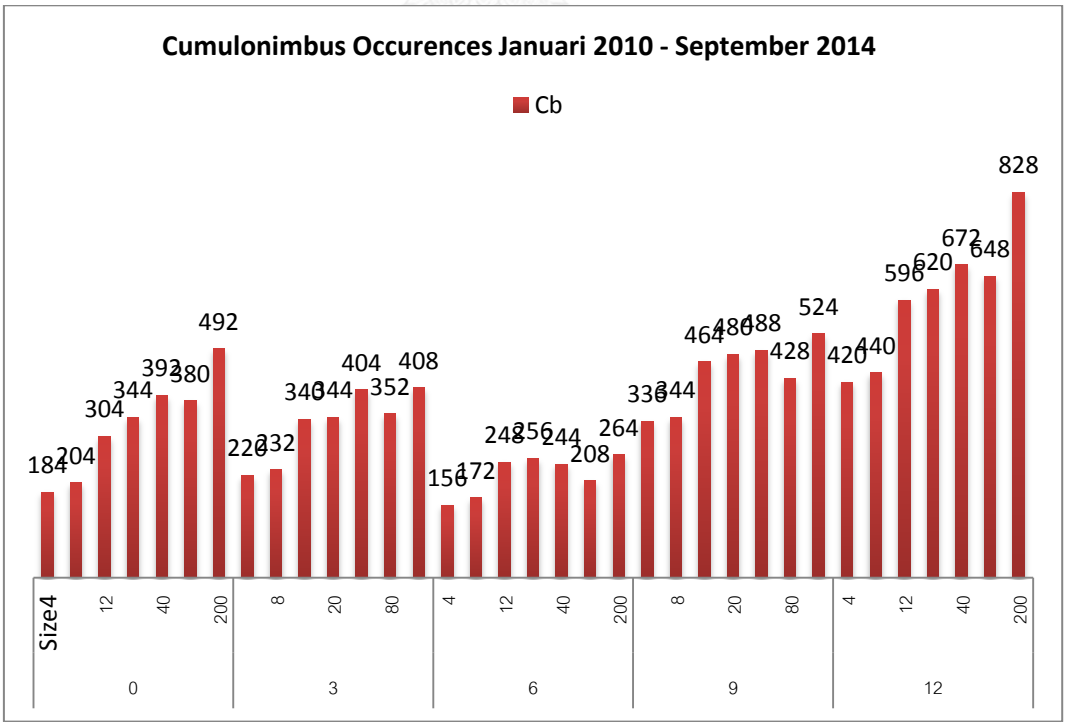

Figure 4. 4 Cumulonimbus Occurence during Januari 2010 - September 2014 presented by MTSAT data.

4.3.1 Critical Level on Thermodynamics Diagrams

The main result obtained from radiosonde data processing in long period indicates that we can understand the characteristics of the air vertically and study its effects on the formation of extreme weather such as cumulonimbus cloud events as cited in this study. To facilitate our understanding, box plot diagram was utilized, representing the events when whether is fair, the sky is clear, and Cb is present by describing the main index value in describing the lability based on dynamic meteorology theory.

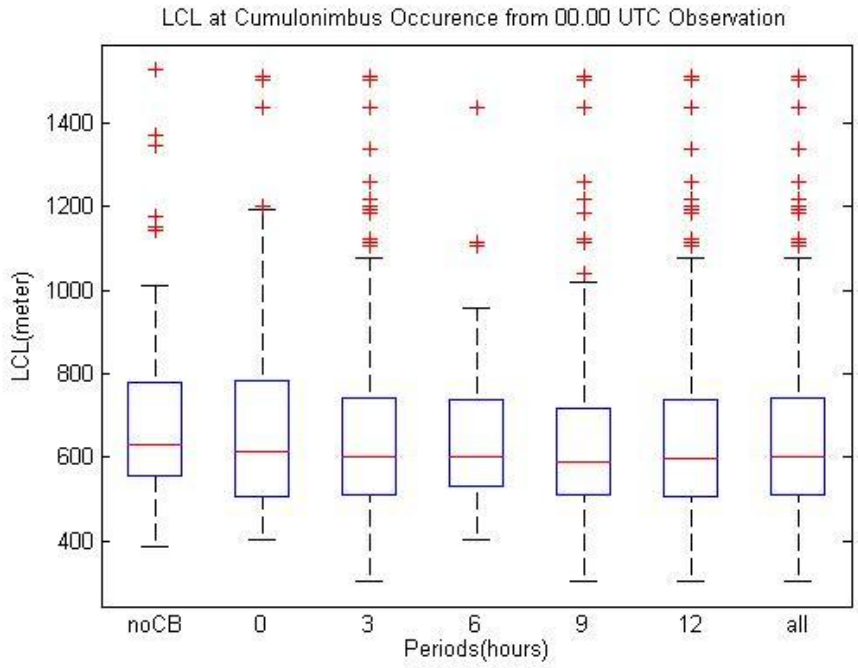

Figure 4. 5 LCL at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

In the first stage, we measured the level of the parcel with lifted dry adiabatic and changed them to become saturated. Using lifting condesation level (LCL), we measured the temperature and dew point temperature following the mixing ratio

until intersecting to the LCL. This would show us the level of which cloud base will begin the vertical lifting for the air parcel.

In the morning period shown in Figure 4.6, the level has different variations, since most level lies between 500 – 800 meters. This means when the level is near to surface, the process of lifting will start imediately. Some high LCL heights indicate that the level of the origin might be unstable. During the night period, the parcel was higher than the data obtained in the morning. The most active cloud convections happen during afternoon until evening periods, so the parcel possibly reach other level at night. By monitoring the level of free convections, we can estimate the level above the parcel that will be able to trigger convection from any force in the surroundings.

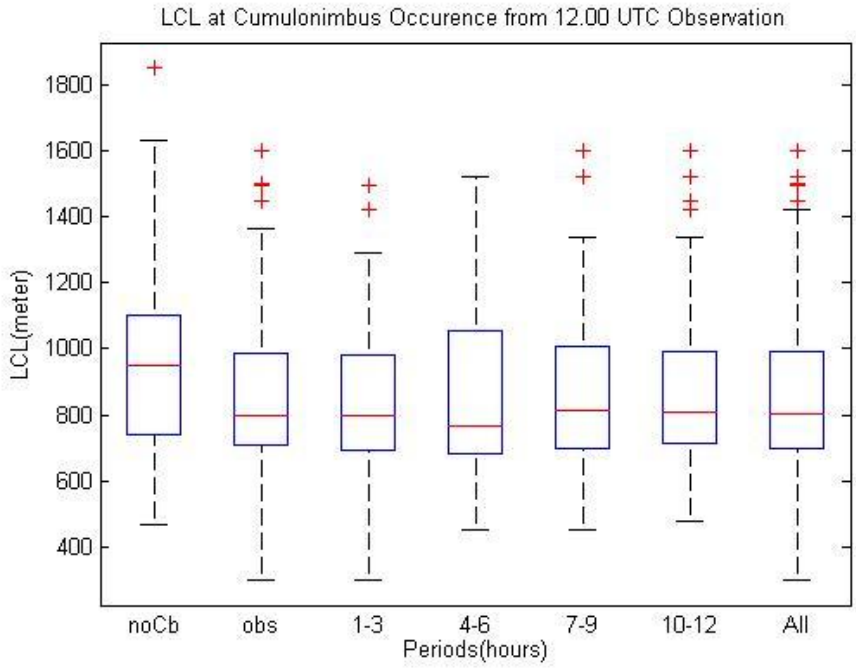

Figure 4. 6 LCL at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

When there is no Cb event, LCL value is higher compared to when Cb event exist. In general the value of LCL in daytime period like in Figure 4.7 has a lower height; it shows the potential formation of convective clouds is greater at daytime

than at night. The difference between morning and evening periods is quite striking, in which air masses push activity at night will be heavier because of the lack of heat at the surface, so that the convectivity process will become more difficult. This remarks as the unique characteristic of tropical regions; that is why, the radiosonde data usage is very beneficial.

We also used the level of free of convection. LFC is the height where air parcel increases to be warmer than the surrounding, so it elevates convective clouds. Air parcel lifts dry adiabatic until it reaches a saturation point (altitude LCL) and then elevates saturated adiabatic.

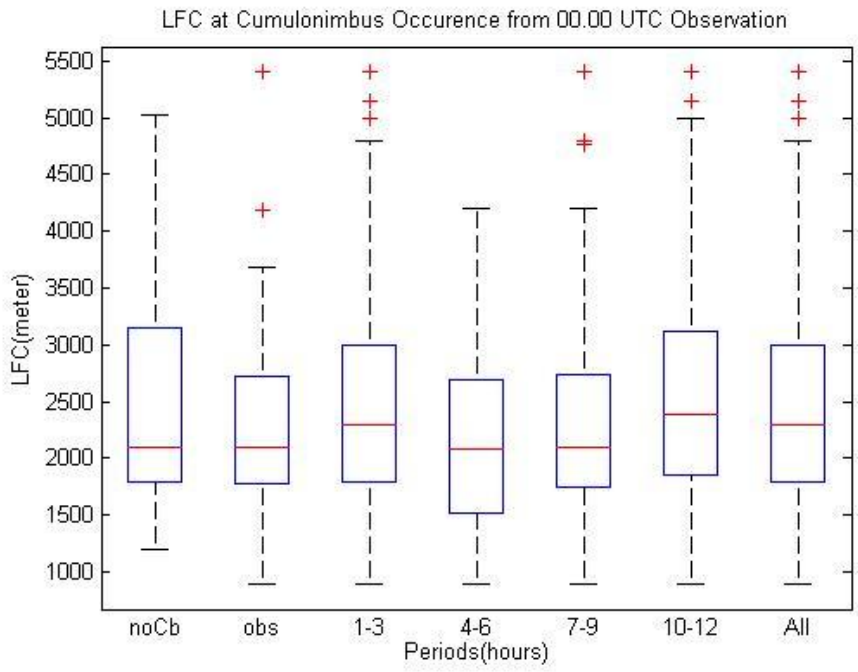

Figure 4. 7 LFC at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

From the 00.00 UTC observation in figure 4.8, the level of free convection became higher during the day; this means the force was mostly due in the last periods on the day time. With strong radiation during middle time, this would be quite enough to push the parcel to become convective. Compared to the night period, the LFC is lower than in the morning. This make senses and is quite

reasonable, since the efficient forcing may be less than in the day time, except when the external factor with larger scale forcing affects the parcel and produces the severe weather.

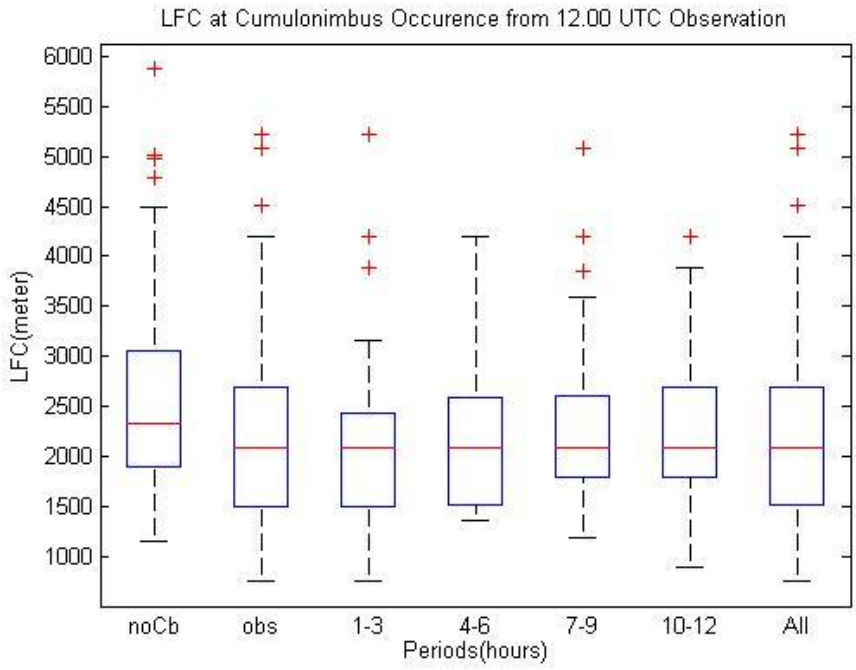

Figure 4. 8 LFC at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

LFC values tend to decrease at night than during the day, including at events that indicate the presence of Cb in Figure 4.9. In addition, the peak value is also quite high and it means the convective cloud formation is able to achieve an optimum height either at night or during the day, though the main source of its formation can be from different things due to regional factors or local convectivity.

4.3.2 Severe Weather Characteristics

The cumulonimbus activity is always related to strong severe weather impact. Many methods have been developed to measure the indices that can represent the best figure for deep convectivity. The resources to be measured in high latitude are gathered by using convective available potential energy (CAPE). When this is implemented to tropical areas with high radiation, the score will show a balance with more than 1000 J/Kg. In the day times like Figure 4.9 and night phase, the result mostly shows the strong energy for convectivity.

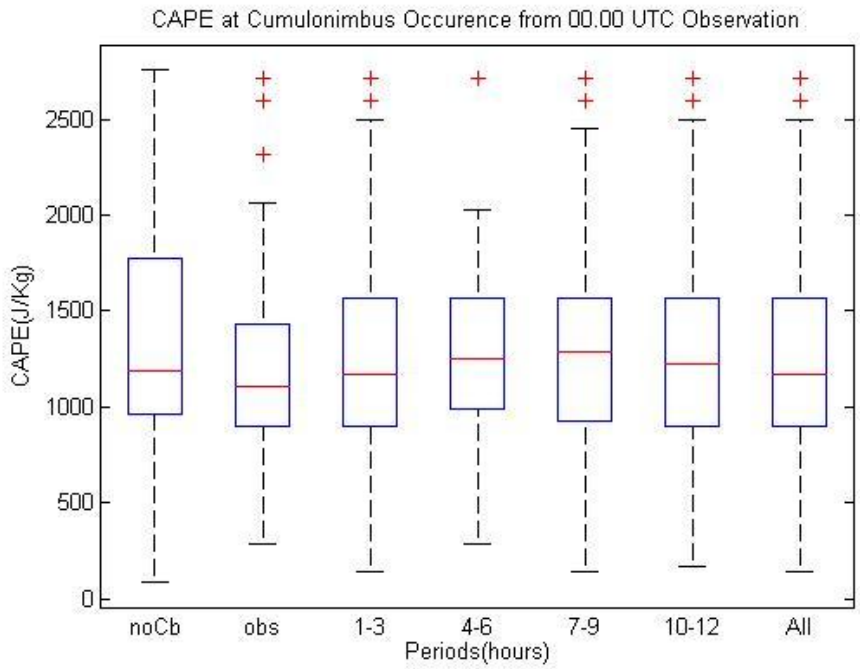

Figure 4. 9 CAPE at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

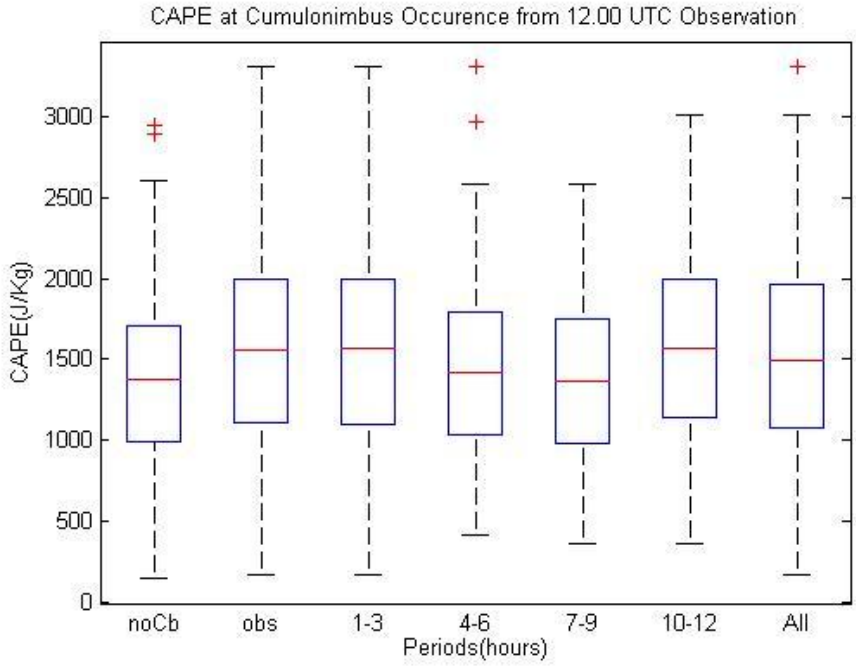

Figure 4. 10 CAPE at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

Observation at 12.00 UTC in Figure 4.10 showed the big value of CAPE; this is possibly connected with most storms that occur during this period. Related with the cumulonimbus forecasting, the value is still high compared to the observation time. From the parcel theory, we can measure the updraft by using lifted index and combine it with CAPE.

The value of convective available potential energy (CAPE) is a representation of the value of the air parcel lifted after certain altitude. This indicates the presence of potential instability in atmosfere. CAPE is expressed in J/Kg which shows how much energy to release the potential of thunderstorms. The value is generally positive and the greater value indicates bigger potential for severe weather. According to literature about high latitudes issued by the MET office, the value of CAPE > 5000 means thunderstorm will potentially appear in the next period. In this research, it shows that the maximum value in the CB cloud formation ranges from 1000 – 2000 J/Kg.

Some atmosphere indices were developed especially to measure the convective storm activity, like on thunderstorm and lightning effect. Almost all of those parameters were derived from radiosonde data as the representations of atmosphere condition. The examples of parameters that calculate the potensial of strong convectivity are lifted index, K-Index, boyden index . Those indices measure the differences of moisture value and temperature from some altitude levels in troposphere layer. Right now, meteorologists develop the indices by numerical weather prediction and use them in a larger area of research, not only for convectivity measurement.

Indices for measurement are not only for convective process, but they are also run out in energy process as a part of supporting air mass index. This is very important to the area with strong radiation like in the tropics. Miller in 1972 developed SWEAT index to prove the severe weather activity not just on storm. This is quite similar with Craven significant severe index. In addition, there is also Energy Helicity index which measures the energy addresed to potential of hail.

The main idea of this research is to maximize radiosonde observation as the observation tool in weather and climate, as it involves a wider scope than other techniques like remote sensing or even numerical weather prediction. Keeping the data can be optimized especially to measure the potential of convectivity and everything associated to it like thunderstorm, lightning, hail, small tornadoes or other weather effects which bring hazards.

To make the research more reasonable for implementation in the operation used for meteorological agency, then some tresholds for indices were also developed to measure the characteristics of each variable in the real condition in tropical atmophere. The physical and dynamical laws are always included when talking about weather phenomena, combined with mathematics especially neural network for making classifier. This research can answer a lot of questions and respond the challenge on this field. Jakarta as the sample of this proposed research with its special characteristics in the climate and weather activities hopely can become the reference for research in the future. The research does not only

mention that certain activity like some indices are not workable, but also facilitates it with physical and dynamic evidence.

There are several main components to trigger the strom occurrence. The pole of the cloud develepment is the process of deep moist convectivity or usually call as DMC. Many theorems are presented to explain this activity, but the easiest one to interpretate is the convectivity process in the troposphere level that makes the changing of the storm effect to their environment. The idea is instability process that makes and develops the cloud system, as in Jakarta with the tropical system and strong radiation. Deep convection becomes the important part for measuring the process of Cumulonimbus cloud growth, followed by some storms and other hazards. The perturbation flow always brings two impacts to the atmopshere itself. This phenomenon can be represented by the availability of convective available potential energy (CAPE) that turns in the dynamics variable to become energy. The forcing of this energy makes the instability situation become stronger and produce torential storm with real effect on the surface.

The principle of CAPE is how the energy forces the parcel in the air to the next layer level, since the lifting process always equals to the adiabatic. Thus, when the parcel becomes less dense to the environment, it will release the latent heat to other side. Then, the process to force up becomes so fast and accelerates it to push the air in the unstable atmosphere. We noted some activities of CAPE are not always be the source of convective activity, for instance, several places in Indonesia with large number of CAPE do not always affect to the real convection with strom event. The characteristics of local area also give some impacts to the local atmosphere system. With the strong vertical wind and less moisture to move up the instability process, this makes the occurrence of Cumulonimbus event does not grow. Nonetheless, in other case, the result can be the opposite especially related to storm, with low value in CAPE but the troposphere layer is very unstable when dealing with high frequency of strom activities.

Since vertical moisture and temperature are always related to atmospheric instability, this profile is also connected to the CAPE index as the energy source to

the parcel. When we noted the positive bouyancy, then the parcel shows lifting to other level. Vertical momentum as perturbation shorted to balancing the principle of hydrostatic when buoyancy indices raise the quantity.

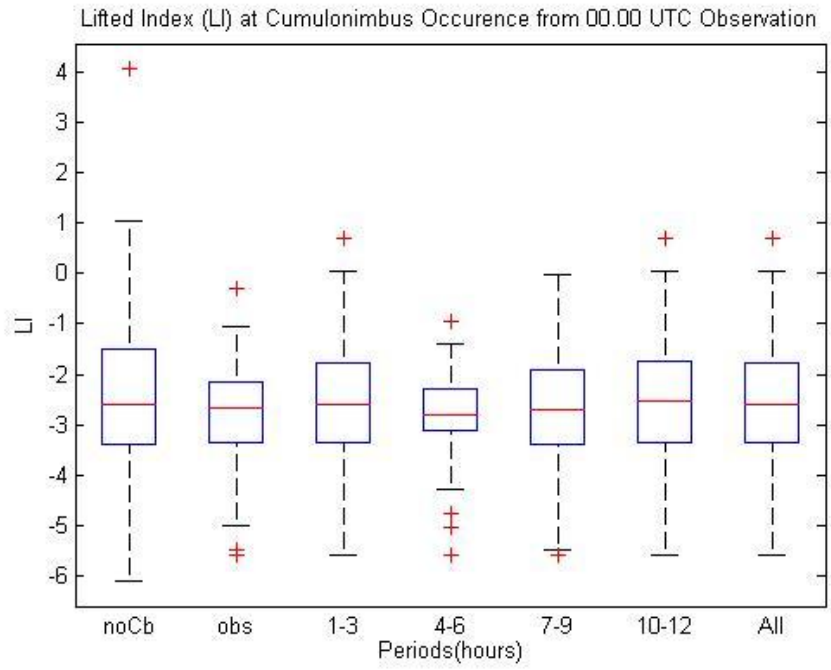

Figure 4. 11 Lifted Index at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

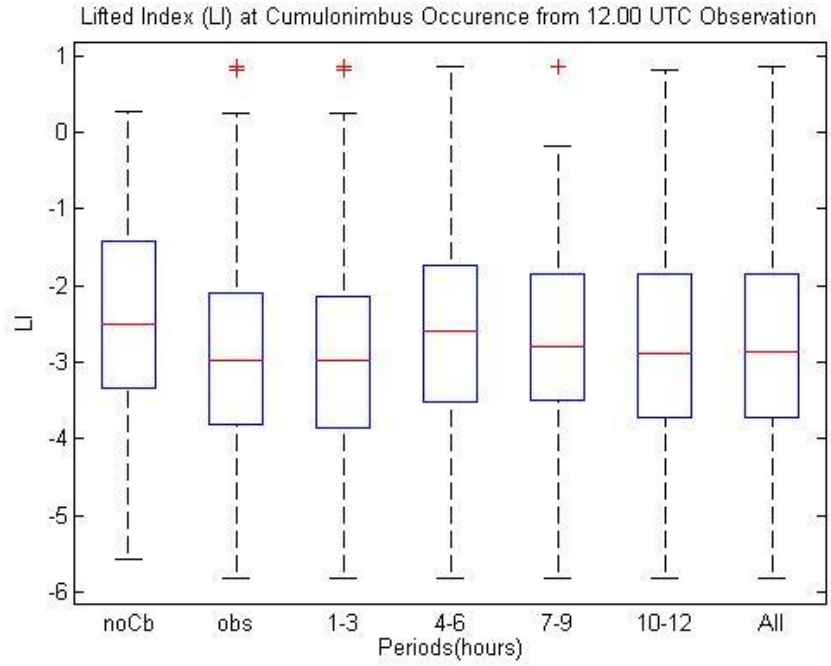

Figure 4. 12 Lifted Index at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

Lifted (LI) is the index value determined from the temperature difference between the air parcel raised in adiabatic (a process by keeping no heat out or into the parcel of air), and the environmental temperature in an altitude of the p pressure in the troposphere. In contrast to the SI which is used to estimate the unsteady layer between 850 hPa and 500 hPa, LI is used to estimate each unsteady layer that is generally for determining planetary boundary layer below 850 hPa.

When LI value is ≥ 0, the atmosphere under the layer referred to is in steady condition. In the situation, turbulence will not easily occur. When LI is between 1 and 6, the atmosphere under layer is categorized as a steady state; in such circumstance, turbulence is not easy to occur. Then, when LI is between 0 and-2, the atmosphere under planetary boundary layer is categorized in a light unsteady state. In such circumstance, turbulence can occur, next, the stormy and thunder clouds with lighting may arise, particularly at the time of cold front or warming during the day. When LI is between-2 and-6, the atmosphere is categorized as quite unsteady; in such condition, turbulence is easy to occur, so strong thunderstorm with the lightning

can occur. However, when LI is smaller than-6, the atmosphere is categorized as very unsteady. In such situation, turbulence is very easy to occur, so thunderstorm with very powerful lightning possibly occurs.

Of the experiment results shown at Figures 4.11-4.12 it was obtained that the value -1 to -3 has an indication of the presence of the Cumulonimbus event, either during the day or night. Though at night the value appears to be lower, in general it is roughly the same. At the limit value, the process of air parcel lifing to the Cb cloud formation can be interpreted by LI that the atmosphere is not in steady state.

Then, the K Index (KI) was used, which is measurement to estimate the potential thunderstorm cloud events based on vertical temperature shrinkage rate of the region containing water vapor in the lower layer, and the vertical expansion of the air layer containing water vapor.

Index K is good enough to be used to mark potential event of air-mass thunderstorm, but it is less suitable for thermal thunderstorm or thunderstorm which comes from the warming process. Besides being used to signify thunderstorm, K index is also used to mark the impact of thunderstorm, for example floods.

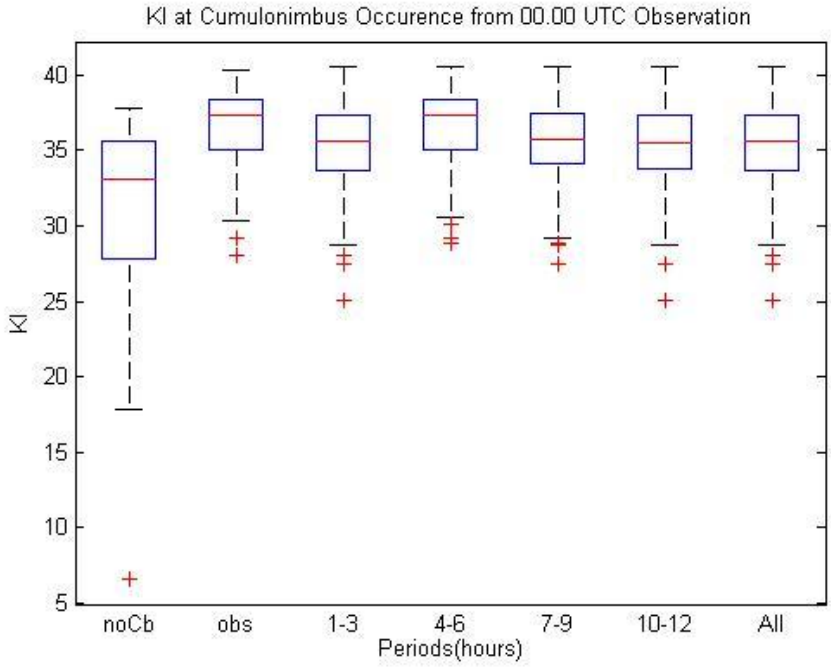

Figure 4. 13 KI at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

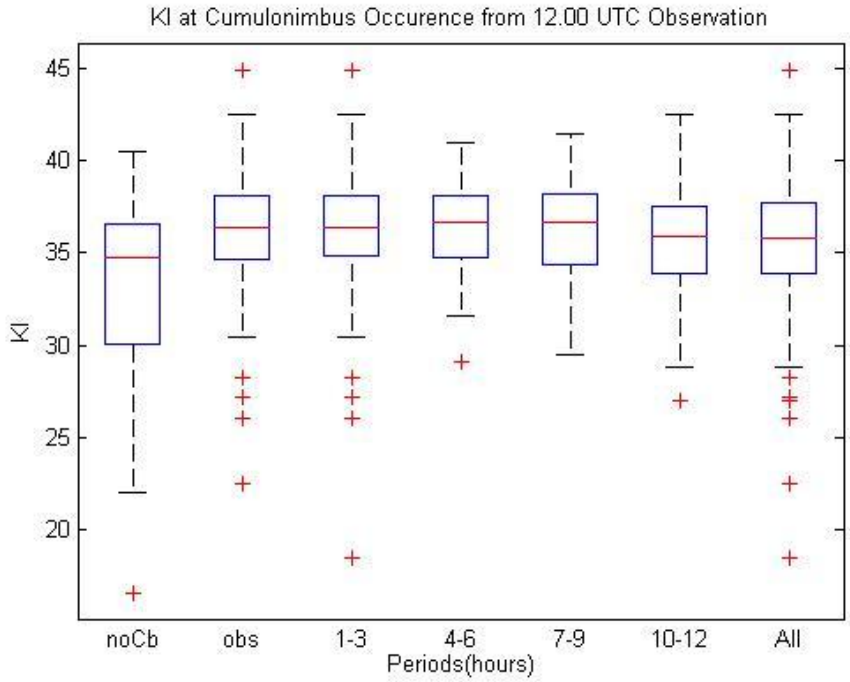

Figure 4. 14 KI at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

From statistic calculation, an indication was obtained that Cumulonimbus events in Jakarta are characterized by the KI value above 30, and in general the value ranges from 35-40 for the maximum. The condition is relatively same at daytime or night that shown in Figures 4.13 and 4.14, that is, above 35. This value will be very useful when used to construct a forecast modelling to predict the Cb value.

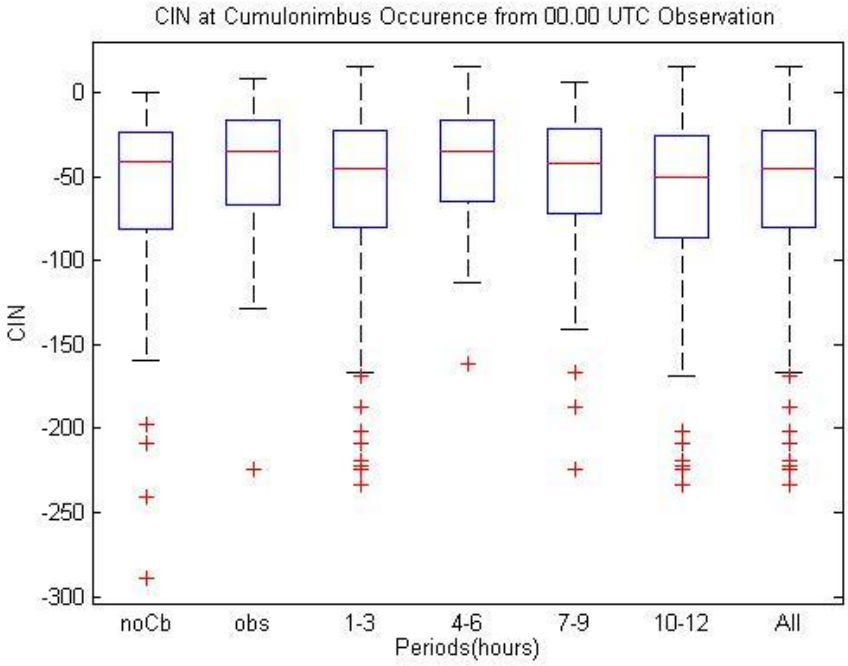

Figure 4. 15 CIN at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

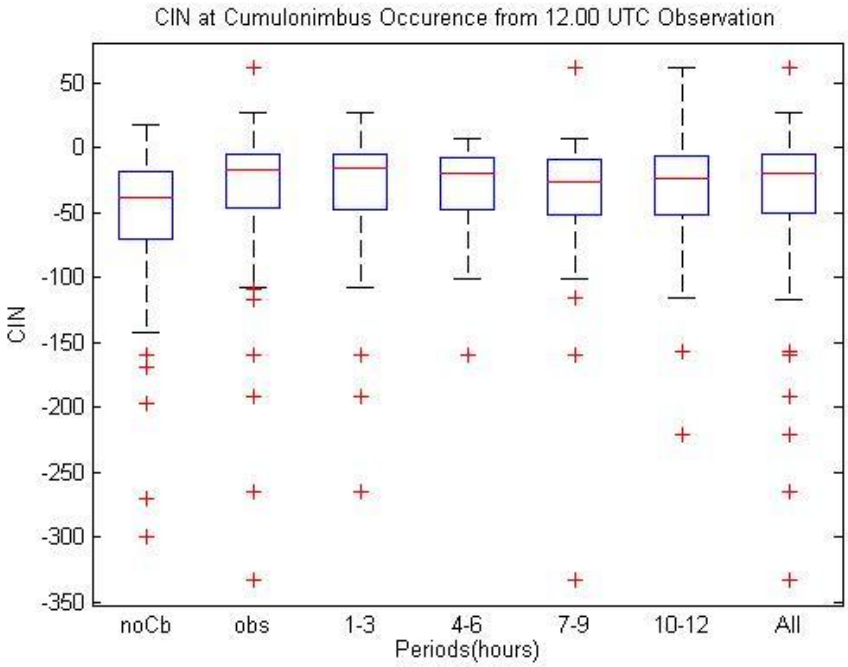

Figure 4. 16 CIN at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

Convective inhibition (CIN) is a numerical value in meteorology that is used to estimate the amount of energy required for the air parcel to rise from the surface up to the level of free convection. The index is based on the understanding that when the lower layer contains cooler air than the above layer, then for the cold air, turbulence may occur if there is a force or energy that can lift up until it reaches the level of free convection. The force or energy can be from external factors, such as the cold fronts, warming, meso-scale wind (land and sea breezes), and orographic lift.

Factors which do not allow CIN to exist is the state where the atmosphere in a certain region with hotter air layer that is above the cooler layer. The presence of hot air layer above the cold air layer enables air parcel to be always cooler than the outside air, until steady state is formed.

From the analysis on one dimension modelling, turbulence may occur if CIN is minimal. Of this model like in Figure 4.16, an understanding was gained that CIN in boundary layer almost always equals to zero. The data obtained also show that to lift the air parcel, the maximum value of CIN required ranges from -20 to -70.

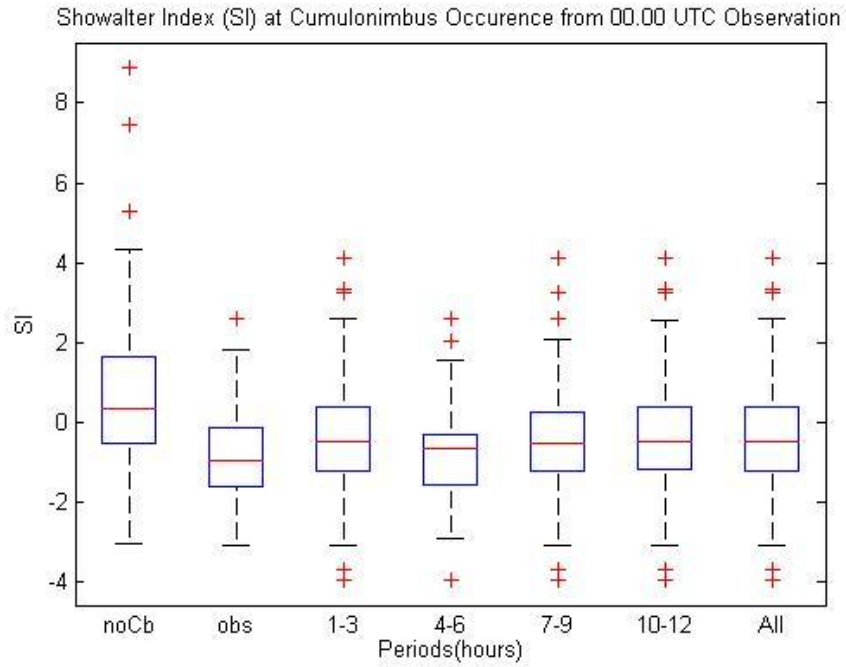

Figure 4. 17 SI at Cumulonimbus Occurences from 00.00 UTC Observation in Jakarta.

Showalter index (SI) is a value used to characterize the unsteady atmosphere in the middle troposphere (above the planetary boundary layer between 850 hPa and 500 hPa). The index formula is written as:

SI = (T500 - TX), in which the T500 is air temperature at the 500 hPa level, and TX is temperature of air parcel is on the 500 mb layer after experiencing saturated adiabatic process starting from the lifted condensation level at about 850 hPa layer. The determination of the index is based on many experiments based on the principle that the length of atmosphere column where saturated adiabatic process occurs is also the place for cloud formation. Positive Showalter index shows air rises until the middle troposphere in a steady state, while negative index shows unsteady state.

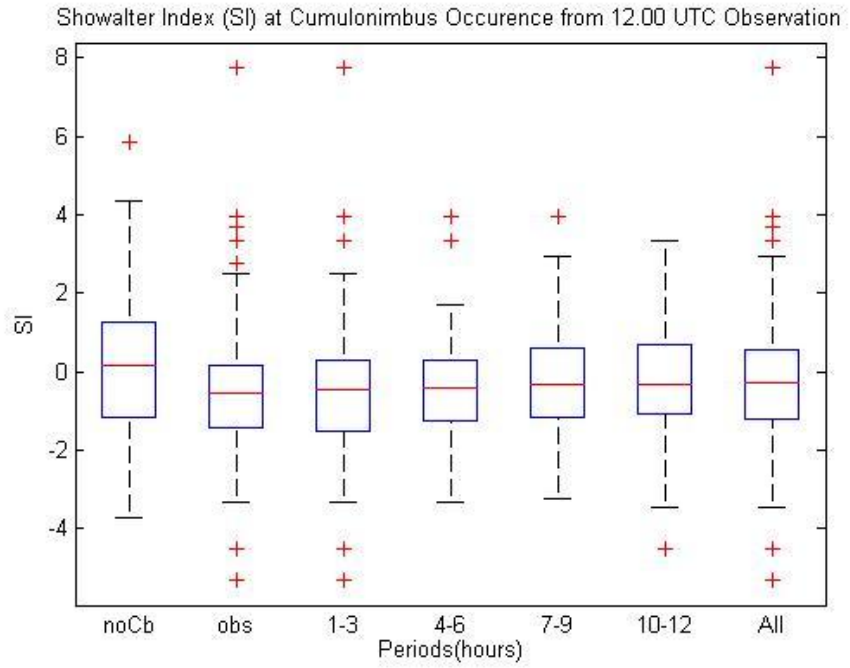

Figure 4. 18 SI at Cumulonimbus Occurences from 12.00 UTC Observation in Jakarta.

Outside tropical areas, generally the absolute value of SI is greater than in the tropics. SI is more appropriate to be used when planetary boundary layer is thin. The criteria are not the same in different places. Furthermore, in practice the air condition

is related to the air condition in planetary boundary layer (see Lifting Index). SI is better when used for air higher than air mass. Therefore, in using SI, we need to see the air mass thickness. Then, when used in tropical areas like the result in Figure 4.17 and 4.18, the unsteady state of the lower layer must also be considered. In this research, it was obtained that when Cb event occurs, the index value ranges from 0 to -2, and this was seen both at daytime and night.

CHAPTER 5
CONCLUSION

In this thesis, we used some neural networks as classfiers and tested two different problems with MTSAT data cloud type algorithm and radiosonde indices to detect the presence of cumulonimbus. To run these two methods, we cannot work independently because of imbalanced data and analytical problems to select the best features in detecting short term cumulonimbust event.

We adopted bootstrap re-sampling and adaboost tehcnique to solve imbalanced data problems. Compared to AdaboostM1 technique, this new purposed technique produces better result as the input features before being processed in neural networks.

As the new improvement of creating cloud type classification in MTSAT, the classifier between RBF and SVM is used but does not give much different accuracy. It means both models work properly. Even by comparing some results from other techniques such as Logistic linier and Random Forest classifier, the result still shows the effective algorithm. This research concludes that, with a combination from fourth channel in MTSAT, improvement by using cloud observation data in Indonesia using neural network technique proves that new MTSAT cloud type classification by neural network classifier is better than by split window techniques.

In the short term cumulonimbus forecasting, by using 78 different indices, we conclude some indices work very well to detect the predictand. Furthermore, the dynamic processes can go trough and be implemented together with artificial neural network as processing. We found that SVM after PCA get the most powerful result in all targets within the time and size constraint. In addition, the model proves that weather model always loses by time and size. The 4 km distance of predicting is the best value, but it also still can be used in the term of 200 km. Since the limitation of time prediction is 12 hours, the most efficient time prediction series is 3 hours in advance. Besides solving imbalanced data problem, the most impressive finding of

this research is feature extraction by principle component analysis which brings effectiveness to select the best and the appropriate variable as an input to the neural networks model. This is understandable since the wrapper, filtering and linear discriminant analysis technique cannot give a better result compared to PCA.

We found some indices mostly appeared in all features in predicting the Cumulonimbus occurences. Some indices are commonly used by previous researchers in relation to the theorem of atmosphere instability, such as K Index Lifted Index, SWEAT, SI, CIN, CAPE, LCL, LFC, SWI. Some other parameters have also been found and interacted very well like Downdraft, CAP, Helicity, precipitable water and Wind component. Since the characteristics during time and size of prediction area are too complex, this model must be set up on special indices in each class.

The methods of this research have been compared with feed forward algorithm**[6]** and backpropagation**[24]** with limited variable. The result shows that this model still has better methods to classify the Cumulonimbus in Jakarta, Indonesia. For the future work, it will be interesting to test neural network classifier by splitting predictand to other problems such as precipitable water or regression problem like rainfall. Furthermore, by combining together remote sensing and radisonde techniques, we can try to initialize numerical weather prediction model.

APPENDIX

REFERENCES

1. Tjasyono, B., *Meteorologi Indonesia I*. 2006, Jakarta: Badan Meteorologi dan Geofisika.
2. Stull, R., *Practical meteorology, an algebra based survey of atmospheric science*. 2012: The University British of Colombia.
3. BMKG. *Indonesia weather and climate information*. 2014 [cited 2014 17 March]; Available from: http.//www.bmkg.go.id/BMKG_Pusat/Meteorology.
4. ICAO, *Guidance on the harmonized WAFS grids for cumulonimbus cloud, icing and turbulence forecast*, in *version 2.5*. 2012, Met office - NOAA.
5. Weng, L.Y., et al. *Lightning forecasting using ANN-BOP & Radiosonde*. in *International Conference on Intelligent Computing and Cognitive Informatics*. 2010. Malaysia: IEEE.
6. Manzato, A., *Sounding-derived indices for neural network based short-term thunderstorm and rainfall forecast.* Atmospheric research, 2005. **83**(3 October 2005): p. 349-365.
7. Kuligowski, R. and A.p. Barros, *Experiments in short therm precipitation forecasting using artificaial neural network.* Mon Weather Reviews, 1998. **126**: p. 407-482.
8. Tokuno, M. and K. Tsuchiya, *Classification of cloud types based on data of multiple satellite sensors.* Advance Space Res, 1994. **14**(1994): p. 299-206.
9. Aldrian, E., *Division of climate type in Indonesia based on rainfall pattern Oceania.* Journal of Marine Science and Technology, 1999. **5**: p. 165-171.
10. Haykin, S., *Neural networks- a comprehensive foundation*. Vol. 2nd edition. 1999, New jersey: Prentice Hall.
11. Guyon, I. and A. Elissef, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**: p. 1157-1182.
12. Wolf, L. and S. Bileschi, *Combining Variable selection with dimensionality reduction*. 2005, Massachusetts institute of technology-Computer science and artificial intelligence laboratory: Cambridge.

13. Liberto, T.D. *The Walker Circulation: ENSO's atmospheric buddy*. 2014 [cited 2014 30 November]; Available from: http://www.climate.gov/news-features/blogs/enso/walker-circulation-ensos-atmospheric-buddy.
14. Groenemeijer, P.H. and A.c. Delden, *Sounding-derived parameters associated with large hail and tornadoes in the Netherlands.* Atmospheric Research, 2006.
15. Suseno, D.P.Y. and T.J. Yamada, *Two dimensional threshold based cloud type classification using MTSAT data.* Remote sensing letters, 2012. **3**: p. 737-746.
16. Innoue, T., M. Satoh, and H. Mapes, *Characteristics of cloud size of deep convection simulated by a global cloud resolving model over the western tropical pacific.* Journal of meteorological society of Japan, 2008. **126**: p. 1-15.
17. Ramyachitra, D. and P. Manikandan, *Imbalanced dataset classification and solutions: A review.* International Journal of Computing and Business Research, 2014. **5**.
18. Thanathamathee, P. and C. Lursinsap, *Handling imbalanced data sets with synthetics boundary data generation using bootstrap re-sampling and Adaboost techniques.* Pattern Recognition Letters, 2013. **34**(2013): p. 1339-1347.
19. Wilks, D., *Statistical methods in the atmospheric sciences*. 1995: Academic press. 467.
20. Litta, A.J., S.M. Idicula, and C.N. Francis, *Artificial neural network model for prediction of thunderstorms over Kolkata.* International Journal of Computer Application, 2012. **50**(July 2012): p. 0975-8887.
21. Reddy, K., K. Raviu, and V.G. Reddy, *Development of new artificial neural network algorithm for prediction of thunderstorm activity.* International journal of systems and technologies, 2009. **2 No 1**: p. 103-112.
22. JMA. *MTSAT Channels*. 2014 [cited 2014 15 October]; Available from: http://www.jma-net.go.jp/msc/en/.
23. Nash, et al., *Introduction to upper air measurement with radiosondes and other in situ observing system*. 2007, Met Office: United Kingdom.

24. Ali, A.F., et al. *Thunderstorm forecasting by using artificial neural network*. in *The 5th International Power Engineering and Optimization Conferences*. 2011. Selangor Malaysia: IEEE.