ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 6, pp. 34-54
DOI:10.33168/JLISS.2024.0603

Prediction of Cumulonimbus Clouds in Airport Vicinity using NOAA Satellite Imagery and Random Forest Models

Yenniwarti Rafsyam[1,2], Eri Prasetyo Wibowo[1], Danang Surya Candra[3], Aini Suri Talita[1], Arief Rinaldi[2]

[1]Doctoral Program in Information Technology, Gunadarma University, Indonesia

[2]Department of Electrical Engineering, Politeknik Negeri Jakarta (PNJ), Jl. Prof. Dr. G.A. Siwabessy, Kampus UI Depok, Indonesia.

[3] Research Center for Remote Sensing, National Research and Innovation Agency (BRIN), Cibinong, 16911

yenniwarti.rafsyam@elektro.pnj.ac.id, eri@staff.gunadarma.ac.id, dana004@brin.go.id, ainisuri@staff.gunadarma.ac.id, arief.rinaldi.te1@mhsw.pnj.ac.id

Abstract. This paper demonstrates a machine learning approach using Random Forest models and NOAA satellite imagery to effectively predict cumulonimbus (Cb) cloud formation, achieving over 75% accuracy. Unlike previous studies that rely on ground-based data, our method leverages satellite images to provide enhanced coverage and perspective. Of the algorithms tested, Random Forest significantly outperforms other models like Support Vector Machines and Neural Networks. Our results indicate that employing NOAA satellite images and Random Forest models holds promise for improving Cb prediction to support aviation safety

Keywords: Cumulonimbus (Cb) Prediction, Machine Learning, Random Forest, NOAA Satellite Imagery.

1. Introduction

A cumulonimbus (Cb) cloud is a type of cloud that has a flat, wall-like shape that is very dark, dense, tall, and large. These clouds are also known as thunderclouds which can bring storms, hail, thunder, or lightning. Cumulonimbus clouds originate from cumulus clouds which are created as a result of air rising and condensing in the sky (Fahruddin Rais et al., 2020). These clouds develop through the process of condensation. After that, the air inside the cloud becomes heavier than the surroundings. This air instability triggers cumulus clouds to grow into vertical clouds. Air moving up the cloud keeps water droplets and ice crystals trapped in the cloud (Liu et al., 2014). After clouds form at high altitudes when the temperature is below freezing point, rainwater appears. Rainwater cannot be held back by the air moving upwards, then the waterfalls to the earth. Rainwater, evaporation, and cooling of the air near the cloud boundary create downward-moving air currents (Tuomola et al., 2021). In this stage, the clouds will get higher and touch the stratosphere. On the other hand, the top of the cloud will spread and form a base, this stage allows lightning, thunder, heavy rain, or hail to occur. It is large, dense, and the cause of rain or storms (Hyvärinen et al., 2015).

Aviation safety is profoundly influenced by weather conditions, playing a pivotal role in shaping operational decisions and procedures. Adverse weather, such as thunderstorms, turbulence, fog, or low visibility, poses significant challenges to flight safety (Rossow et al., 2022). Pilots and air traffic controllers must navigate through constantly changing weather patterns, making informed decisions to mitigate risks and ensure passenger and crew well-being. Weather-related factors impact various aspects of aviation, from takeoff and landing to en-route navigation (Temme & Tienes, 2018). Modern aircraft are equipped with advanced weather radar and forecasting systems, enabling pilots to receive real-time information about weather conditions and make timely adjustments to flight plans. Additionally, aviation authorities impose strict guidelines and regulations to ensure safe operations in different weather scenarios (Olaganathan & Ham, 2020).

Delays, diversions, or even cancellations may be implemented when adverse weather conditions compromise the safety of flight operations. Overall, a comprehensive understanding and effective management of weather-related challenges are integral components of aviation safety protocols, ensuring that the industry continues to prioritize the well-being of all those involved in air travel (Kleber et al., 2022). The existence of cumulonimbus clouds can have significant effects on aviation safety (L. Peck, 2015). Pilots and air traffic controllers closely monitor weather conditions to ensure safe flights and cumulonimbus clouds pose several hazards such as turbulence, icing in flight, in-flight electrical disturbances, lightning, wind shear, and reduced visibility.

To mitigate these risks, pilots rely on weather information, radar systems, and communication with air traffic control to make informed decisions about their flight paths. Modern aircraft are equipped with weather radar, lightning detection systems, and other technologies to enhance safety when flying through or around cumulonimbus clouds. Additionally, aviation authorities provide weather forecasts and warnings to help pilots plan their routes and avoid hazardous weather conditions (Kim, 2011).

Cumulonimbus existence prediction has been reported in various types of research. This can help the air traffic control to determine whether it is safe for the aircraft to take off or land. Research conducted by (Rais, et al., 2020) used the Integrated Forecast System (IFS) European Medium-Range Weather Forecast (ECMWF) model in the Flight Information Region (FIRs) to predict the Cb cloud in Jakarta and Ujung Pandang flight route. Another research also conducted by Rais et al compared various numerical weather models to predict Cb cloud existence (Rais et al., 2022).

Artificial intelligence has also been reported in predicting Cb clouds. (Zhang et al., 2023) use deep learning to process the data collected by radar to predict the Cb cloud presence. (Putra et al., 2021) use machine learning to predict the Cb cloud from radiosonde data. These researchers use data collected from the ground to predict the cloud. In this research, data collected from NOAA satellite images is used to predict the presence of Cb clouds in the Indonesia airport area. NOAA satellite is operated by the National Oceanic and Atmospheric Administration (NOAA).

NOAA operates a fleet of satellites that are part of the National Environmental Satellite, Data, and Information Service (NESDIS). These satellites are designed to observe and monitor various aspects of the Earth's atmosphere, oceans, and land surfaces, providing critical data for weather forecasting, climate monitoring, and environmental research (Davis, 2007). In this research, a satellite image captured by NOAA is used as the dataset.

The data set used is the result of georeferencing from NOAA L1B Lapan satellite imagery. This data set is the novelty of this research because there is no similar dataset yet. This research uses the machine learning method with Random Forest, Support Vector Machine (SVM), and Neural Network (NN) algorithms to detect the presence of Cb clouds in the Indonesian airport area.

2. Literature Review

2.1.Cumulonimbus Clouds

Cumulonimbus clouds (Cb) are one of the most dramatic types of clouds and can have significant impacts on local weather. They form from strong convective processes in the atmosphere, where warm and moist air rises rapidly to high altitudes, meeting colder air above and causing rapid condensation of water vapor. Cumulonimbus clouds can reach heights of more than 10 kilometers above the Earth's surface, making them one of the most monumental and easily recognizable types of clouds in the sky (Fahruddin Rais et al., 2020).

The main characteristic of cumulonimbus clouds is their ability to produce extreme weather. Thunderstorms, heavy rain, strong winds, and even tornadoes are often associated with the presence of these clouds. Strong electrical activity within them often results in thunder that reverberates through the atmosphere, while heavy rain and strong winds frequently accompany their presence. Cumulonimbus clouds often resemble tall towers with tops resembling mountains or cauliflower florets, providing a frightening visual for observers on the ground (Hyvärinen et al., 2015).

In addition to causing bad weather, cumulonimbus clouds can also pose serious threats to safety and infrastructure. In some cases, they can produce tornadoes, which can cause severe damage to buildings and even threaten human lives. Therefore, good monitoring and understanding of cumulonimbus clouds are crucial in efforts to improve weather forecasting and enhance disaster management capabilities (Tuomola et al., 2021).

2.2.National Oceanic and Atmospheric Administration (NOAA)

The NOAA satellite, comprising the GOES (Geostationary Operational Environmental Satellites) and JPSS (Joint Polar Satellite System) series, plays a crucial role in monitoring and understanding atmospheric dynamics, weather conditions, and climate change worldwide. Operational in microwave frequencies for remote sensing and data transmission, and utilizing radio frequencies for communication with ground control stations, these satellites consist of several main components that work synergistically to gather the necessary data for more accurate weather forecasting, improved climate modeling, and deeper understanding of environmental phenomena (Kopacz et al., 2023).

NOAA satellites rely on a robust satellite platform to support crucial instruments like optical, thermal, and microwave sensors, vital for observing atmospheric conditions and the environment. The power system, comprising solar panels and batteries, generates the necessary power for all operations. A sophisticated communication system facilitates communication with ground control stations and transmits collected data for analysis on Earth. (Saidah et al., 2020). This facilitates weather monitoring, climate modeling, and environmental observation by providing essential data on surface temperature, humidity, and oceanic and terrestrial conditions. Its superior temporal resolution allows for frequent monitoring of weather and environmental changes, enhancing our understanding of atmospheric dynamics and environmental changes over time. (Rossow et al., 2022).

2.3.Machine Learning

Machine learning is a branch of artificial intelligence aimed at developing algorithms and statistical

models that enable computer systems to learn from data and perform specific tasks without the need for explicit instructions. Unlike conventional approaches where programmers have to determine precise rules and steps, in machine learning, computer systems use patterns and inference to understand data and generate decisions or predictions (Mishra & Gupta, 2017).

Machine learning algorithms allow systems to process large volumes of historical data, identify complex patterns, and extract useful information from that data. In other words, these systems can learn from past experiences to improve performance in the future. This learning process occurs through repeated iterations of learning, where the system gradually enhances its understanding and capability to handle various situations (Seth, 2024).

2.4.Random Forest

The Random Forest (RF) algorithm, commonly used for classification tasks, combines outputs from multiple decision trees to enhance prediction accuracy. Initially, random samples are selected from the training data via bootstrap sampling. Decision trees are constructed using these samples, with splits based on criteria like Gini impurity or Entropy. This iterative process continues until further splitting becomes impractical. RF, a supervised learning method, excels in handling complex datasets and delivering robust predictions.

Once all trees are built, predictions from each tree are collected and combined, either by majority vote (for classification) or averaging (for regression), to produce the final prediction. Model performance evaluation is conducted using appropriate evaluation metrics, such as accuracy or mean squared error. Random Forest has advantages in handling complex data, and the diversity of trees built from bootstrap samples helps prevent overfitting (Wan et al., 2021).

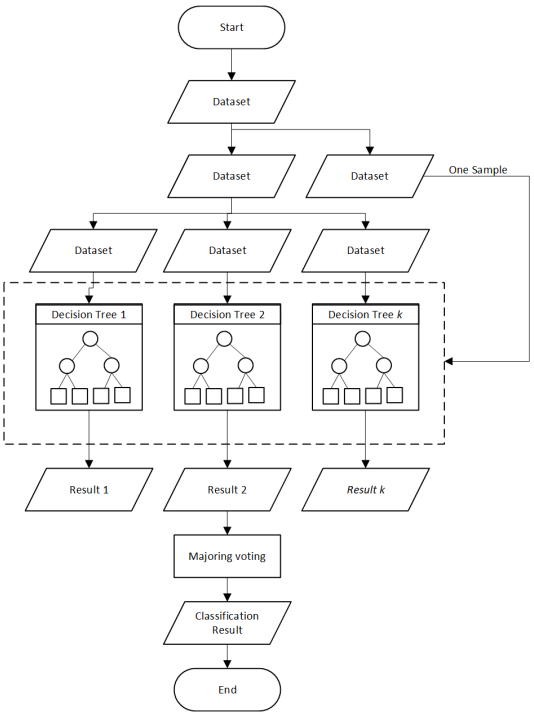

Fig. 1: Random Forest Algorithm Flowchart.

Fig. 1 illustrates the process starting with the selection of k subsets from the training set, denoted as D = {D1, D2, … Dk}. Each subset maintains the same size as the original training set, selected using a sampling method allowing repeated selections. The subsequent step involves training decision trees. This includes taking a sample with n features from each training subset and randomly selecting m features to construct a feature subspace denoted as S, where m < n. Then, the optimal splitting point for the decision tree node is computed based on the formed feature subspace S (Wan et al., 2021).

2.5.Support Vector Machine (SVM)

SVM is a popular algorithm for classification and regression tasks in machine learning. It constructs a hyperplane in a high-dimensional space to maximize the margin between classes, using a subset of training data called support vectors. Its goal is to minimize classification errors while maximizing the margin, representing the closest distance between support vectors and the hyperplane. (Alita et al., 2020).

SVM's key advantage lies in its capacity to handle non-linear data using kernel functions, which map data into higher-dimensional spaces for linear separation of classes. It effectively addresses overfitting and offers a strong geometric interpretation. Initially designed for binary classification, SVM now supports multi-class classification, regression, and outlier detection. Its benefits include high performance on small and large datasets, robustness with high-dimensional data, and ease of implementation. SVM finds widespread use in applications including pattern recognition, text classification, and bioinformatics, among others. (Abdusyukur, 2023).

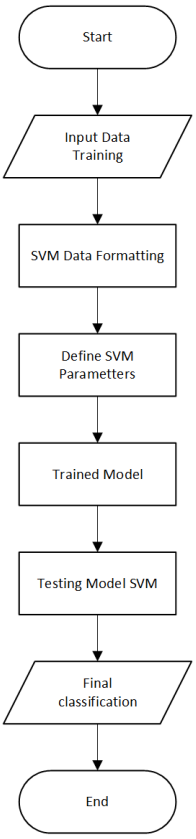

Fig. 2: Support Vector Machine (SVM) algorithm flowchart.

The SVM algorithm begins with preparing training and testing data. Fig. 2 outlines the process: selecting a kernel type (e.g., linear, polynomial, or radial basis function) based on the data, initializing SVM parameters like C, gamma, and degree, and training the model using optimization algorithms such as stochastic gradient descent or Quadratic Programming (QP). The trained model's performance is evaluated using testing data, employing metrics like accuracy or cross-validation. Finally, the SVM model is ready for predicting new classes using unseen input data. Thus, the SVM algorithm provides a structured framework for training classification or regression models by leveraging geometric principles to find the optimal hyperplane or separating function (Adawiyah et al, 2022.).

2.6.Neural Network

Artificial Neural Network (NN) is a computational model inspired by the structure and function of the human brain. It consists of a collection of simple processing units called neurons, which are connected

through a series of connections or weights. Each neuron receives input from other neurons, multiplies it by its weight, and runs an activation function to produce output. Through the training process, these weights are iteratively adjusted to optimize the network's performance in learning patterns or implementing specific functions. There are various types of neural network architectures, including feedforward, recurrent, and convolutional, each with its applications and strengths in processing data (Kumar Verma et al., 2013).

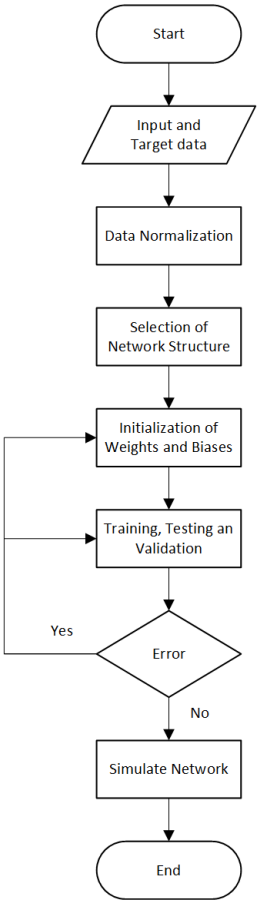

Fig. 3: Neural Network algorithm flowchart.

The neural network flowchart begins with initialization, where the number of layers and neurons is determined, and initial weights are randomly set as shown in Fig. 3. Input data is then fed into the input layer, followed by feedforward propagation through hidden layers to the output layer. Each neuron employs an activation function to modify its output. Subsequently, error computation involves comparing predictions with actual values using a loss function, and the backpropagation algorithm updates weights and biases by calculating the gradient of the loss function. This process iterates through several epochs using the training dataset to enhance network performance. Upon completing training, evaluation occurs on the validation or test dataset to gauge the model's performance. Finally, the neural network can be utilized for predictions on new data by inputting it into the input layer and obtaining output from the output layer. (Kumar Verma et al., 2013).

3. Methodology

This research began with the process of creating a dataset sourced from NOAA Level 1b (L1B) satellite images. Each image in the dataset has undergone calibration and geolocation, facilitating further analysis. Since NOAA L1b datasets are not available in public repositories such as Kaggle, UCI, or Figshare, we created this dataset independently. Although there are publicly accessible cloud datasets, these images are generated from ground-based camera image captures, unlike satellite images which are our focus. Our main objective is to develop a model to predict the presence of Cumulonimbus (Cb)

clouds around airport areas based on satellite images. After obtaining the dataset, the next step is to apply machine learning techniques to predict the presence of Cb clouds with high accuracy.

3.1.Dataset creation using the geometry point method.

The dataset was built using NOAA L1B-BRIN LAPAN satellite image data manually processed through QGIS software version 3.14 by applying point geometry method. This process involves several careful steps. The data is organized by obtaining 9 main pieces of information, including attributes such as id, class_id, as well as xcoord and ycoord coordinates that depict the geographical location of each point in the satellite image. Additionally, the data also includes information from each spectrum band, namely band-0 to band-4, which provide insights into the physical and chemical properties of each observed area. To meet the dataset requirements, the overall procedure carried out is as follows: Firstly, the boundaries of Indonesian district maps for the year 2020 are mapped in Shp format to provide spatial context for the collected data. Next, sample image files in (.tif) format are generated from the L1B images converted using QGIS software. This conversion process allows for further interpretation of visual data from satellite images for advanced analysis.

Furthermore, Aerodrome Meteorological Report (METAR) data is also included in the dataset to provide relevant weather information at the time of satellite image acquisition. This enables the evaluation of the impact of weather conditions on the interpretation of satellite image data. Lastly, relevant airport positions are also included to facilitate analyses related to air transportation infrastructure around the observed areas. Thus, the resulting dataset not only encompasses spatial data but also additional information useful for a deeper understanding of the observed geographical environment.

The process begins with digitizing raster data using the point geometry method, utilizing QGIS software. In this stage, adjustments are made to brightness levels and band levels so that the final image results appear visually better and meet the predetermined analysis requirements. These adjustment steps play a crucial role in ensuring the quality and accuracy of the resulting dataset. Thus, this process becomes key in ensuring that the dataset used for analysis maintains high integrity and can provide reliable results.

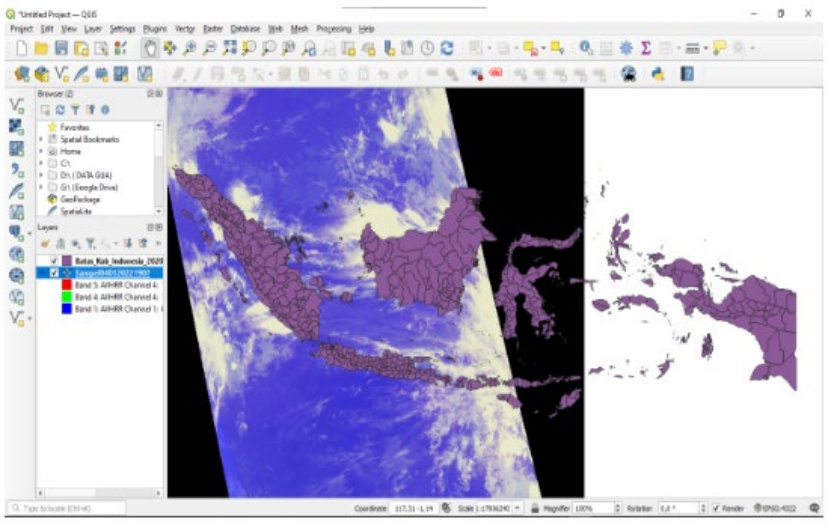

Fig. 4: Digitized satellite image sample file in .tif format.

In Figure 4, we utilized the dataset was built using NOAA L1B-BRIN LAPAN satellite image data manually through QGIS software version 3.14. This process involves setting the Coordinate Reference System (CRS) with parameters CRS: WGS 72 and Authority ID: EPSG 4322 to ensure coordinate conformity. Subsequently, the raster image is juxtaposed with METAR (Meteorological Aerodrome Reports) data to identify which airports are affected by Cumulonimbus clouds (Cb). The next step is to retrieve the METAR reports to ascertain the meteorological conditions around Kualanamu airport as a

sample, while satellite images are matched with the geographical position of the airport. Categories for Cb are determined based on METAR data, enabling accurate grouping based on the characteristics of these clouds.

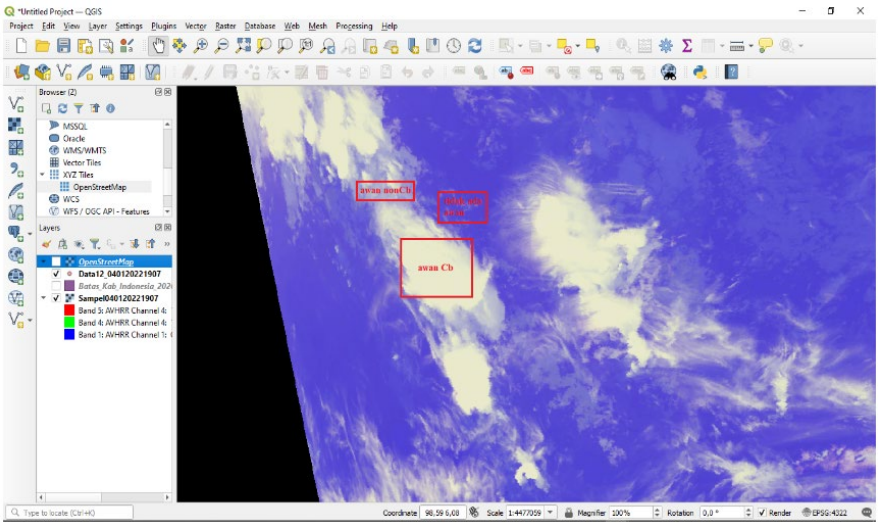

Fig. 5: Display of the locations of three categories of clouds in satellite imagery selected for digitization based on METAR data

After that in Figure 5, check the prepared METAR data. Considering that METAR data uses time in Coordinated Universal Time (UTC+7), it is necessary to adjust the time on the captured image sample from 19:07 WIB to 12:07 UTC. Next, recheck the METAR data to see where Cumulonimbus (Cb) clouds were present at the airport on January 4, 2022, at 19:07 WIB or 12:07 UTC. Cumulonimbus clouds are observed around the Medan area or North Sumatra, particularly around Kualanamu International Airport (WIMM). Points are generated using QGIS software. Initially, points or vectors are placed in the vicinity of Kualanamu International Airport in Medan. Initially, points are assigned for the Cb cloud class, NonCb class, and no cloud class. The quantity of points or vectors in each cloud class is documented. Additional points or vectors enhance the data's granularity. Convert the data into an excel file using MyGeodata Converter to get the coordinates of each point.

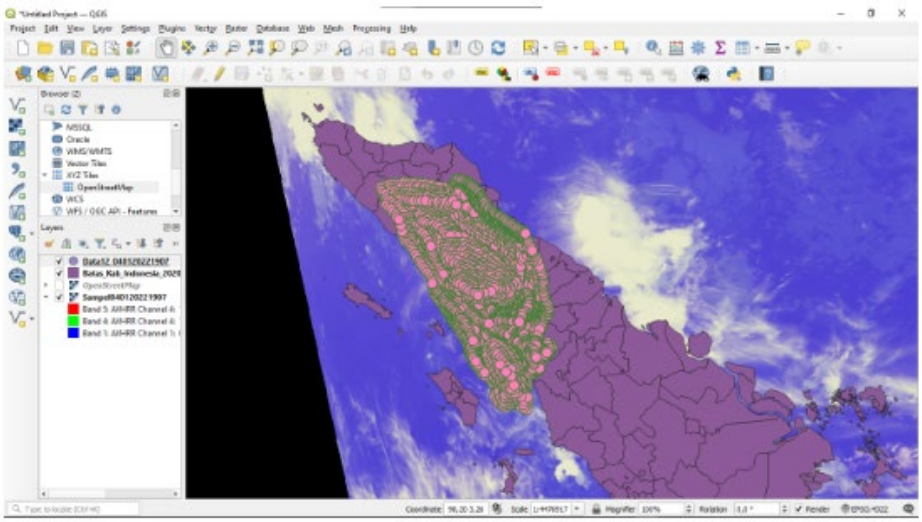

Fig. 6: The results of digitization can be seen in satellite imagery and maps of Indonesia.

In Figure 6, Cumulonimbus clouds are observed around the Medan area or North Sumatra, particularly around Kualanamu International Airport (WIMM). Points are generated using QGIS software. Initially, points or vectors are placed in the vicinity of Kualanamu International Airport in

Medan. The quantity of points or vectors in each cloud class is documented. Additional points or vectors enhance the data's granularity. In conducting the analysis of points or vectors around Kualanamu International Airport in Medan, it is important to consider the variations in clouds that may occur. Firstly, we should begin by marking the areas of Cumulonimbus (Cb) clouds, which typically indicate the potential for storms or adverse weather conditions. After that, attention should also be paid to non-Cb areas, which may indicate clear or slightly cloudy conditions. Furthermore, we need to note the areas where there are no clouds at all, indicating very clear weather conditions.

When collecting data, it is important to record the number of points or vectors in each cloud class. The more points or vectors added, the more detailed the resulting data will be. We can identify weather patterns that may occur around the airport. This will aid in flight operational planning and decision-making related to weather factors that could affect flight safety. Convert the data into an excel file using MyGeodata Converter to get the coordinates of each point.

Table 1: The generated dataset.

id	Class_ID	ycoord	xcoord	class_name	note
1	1	3.599932	97.28009	Cb	
2	1	3.63401	97.31726	Cb	
3	1	3.640206	97.36063	Cb	
4	1	3.658793	97.4164	Cb	
5	1	3.658793	97.47836	Cb	
6	1	3.640206	97.54032	Cb	
7	1	3.652597	97.58988	Cb	
8	1	3.664989	97.62706	Cb	
9	1	3.664989	97.68902	Cb	
10	1	3.63401	97.75717	Cb	
980	2	2.851776	98.73458	NonCb	
981	2	2.833188	98.77175	NonCb	
982	2	2.808404	98.79963	NonCb	
983	2	2.783621	98.82442	NonCb	
984	2	2.752641	98.83991	NonCb	
985	2	2.718564	98.8492	NonCb	
986	2	2.681388	98.8523	NonCb	
987	2	2.650408	98.87089	NonCb	
988	2	2.619429	98.90496	NonCb	
2006	3	2.221341	97.67663	No clouds	
2007	3	2.202754	97.70141	No clouds	
2008	3	2.453688	97.5651	No clouds	
2009	3	2.419611	97.58678	No clouds	
2010	3	2.391729	97.61157	No clouds	
2011	3	2.373141	97.62396	No clouds	
2012	3	2.348358	97.64255	No clouds	
2013	3	2.32977	97.66423	No clouds	
2014	3	2.308084	97.68902	No clouds	
2015	3	2.277105	97.7138	No clouds	

Table. 1, shows the Geospatial Data Abstraction Library (GDAL) algorithm is utilized as the primary library for managing geospatial data in raster and vector formats. GDAL serves as a translator capable of extracting crucial information from datasets, including raster bands and geographic coordinates associated with each pixel value. The dataset constructed in this research is a structured

collection of features essential for further analysis. These features include unique identification (ID), classification (Class_ID), as well as geographic coordinates (longitude and latitude) of each location. Additionally, the dataset includes Digital Number (DN) values, representing the pixel values in each of the 1 to 5 bands (Channels) in NOAA imagery.

It is noteworthy that these DN values represent information that has not undergone radiometric processing, thus, these values do not directly reflect the actual wavelengths. However, DN values remain significant as they depict the characteristics of the detected objects in the image, albeit without radiometric correction. Furthermore, the dataset also stores georeferencing information enabling accurate spatial analysis.

3.2.Feature Retrieval

To obtain the example dataset, which includes a series of raster bands and other important information such as Digital Number values and coordinates points, the digitized data in .csv format and satellite images in .tif format are processed using the algorithm that has been designed. This algorithm relies on the Geospatial Data Abstraction Library (GDAL) in the Python programming language to perform this process. The example output dataset generated during the training process with the GDAL algorithm includes attributes such as id, class_id, xcoord, ycoord, and class_name, as well as values from each spectrum band, namely band-0 to band-4.

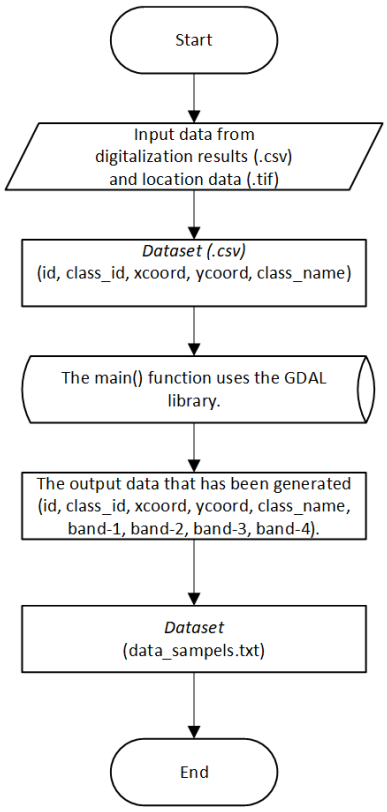

Fig. 7: Flowchart for inputting data (.csv) and training

The flowchart diagram in Fig. 7 illustrates the process flow from inputting data from training to generating output dataset samples in .txt format. The system reads and loads data from a .tif formatted location file using GDAL and a .csv formatted digitization result file, the next step is to process the data. This involves merging location data from satellite images with digitization results, which is done based on spatial coordinates (xcoord, ycoord). This process also includes a preprocessing stage, where data can be filtered or cleaned if necessary, to ensure its quality and integrity.

The Input Data from Digitization Results (.csv) and Location Data (.tif) represent two distinct types of data used in geospatial analysis or mapping. Digitization Results Data (.csv): This data is recorded or

entered into a CSV (Comma-Separated Values) format following the digitization process. It may encompass various types of information, including field survey data, laboratory experiment results, or digitally recorded data stored in CSV format, depending on the context. This data stores location information in (.tif) (Tagged Image File Format) files. (.tif) format is commonly employed to store images or maps incorporating spatial information, such as geographic maps or satellite imagery. In this context, "Location Data" may refer to a map or image containing details about specific geographic locations, such as administrative boundaries, land surface alterations, or other natural phenomena.

Both types of data can be used together for geospatial analysis or mapping, where location data from (.tif) files can be paired or linked with digitization results data from CSV files to explore or analyze the spatial relationships between various phenomena or variables. For example, digitization results data can be used to analyze the distribution or patterns of a phenomenon at specific geographic locations identified in the (.tif) file.

After successfully building the dataset, the next step is to train the model to obtain a model that will be used to detect Cumulonimbus (Cb) clouds and their locations using an algorithm that has been designed. The model training process involves the use of several algorithms, including Random Forest, Support Vector Machine, and Neural Network, to determine the overall model accuracy. Thus, through this process, it is expected that a model with high detection capabilities for Cb clouds can be developed and relied upon for applications related to weather and environmental monitoring.

The system designed to process data in the context of machine learning follows a structured and complex flow. Firstly, data obtained from various sources is collected and prepared for analysis. The data collection process may involve sensors, logging software, or other data sources. Next, the collected data is input into the system for preprocessing, including data cleaning, normalization, and feature extraction. Once the data is ready, the next step is to select good data for analysis. This involves the evaluation and selection of the most relevant and significant features for the specified analysis goals. The use of techniques such as correlation analysis and dimensionality reduction can aid in this process. Then, the selected data is used to train machine learning models. The training process may involve various machine learning algorithms, such as decision trees, neural networks, or ensemble methods. These models are fed with the previously selected data and iteratively updated to improve their prediction performance. Once the models are trained, the next step is to classify cloud types, in this case, Cumulonimbus (Cb) and Non-Cb clouds, using the prepared models. The models will receive data about cloud formations and characteristics and then provide predictions on whether they are Cb or Non-Cb.

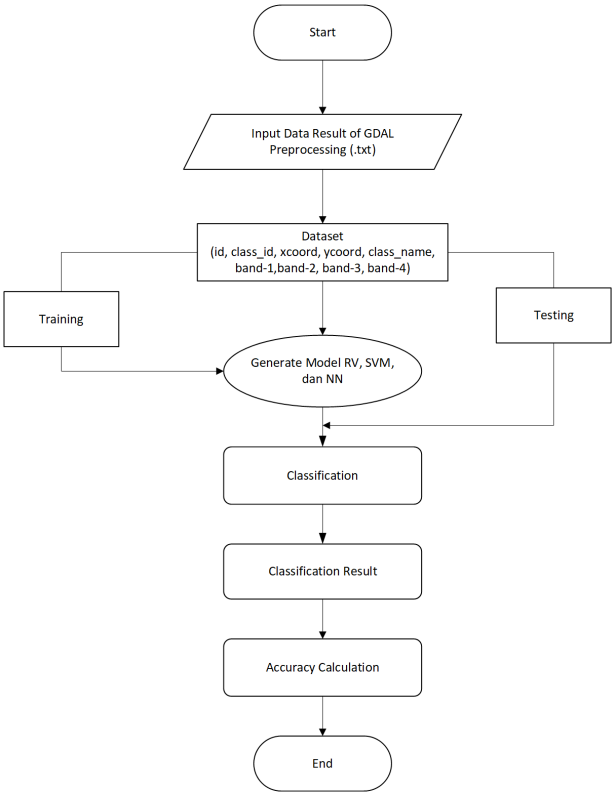

Fig. 8: Flowchart Machine Learning System.

Fig. 8, shows the process begins with the reading of previously processed data using GDAL, typically stored in text format (.txt). This data is then loaded into a dataset that includes various attributes such as id, class_id, x and y coordinates, class names, and values of several bands that may exist in the data.

After that, the dataset is partitioned into training and testing subsets for the training phase and testing processes of the classification model. Classification models such as Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) are then built using the training data. The training process is conducted so that the model can learn the patterns contained within the training data.
Once these models are built, they are evaluated using the testing data to produce the accuracy of each model. This evaluation allows for determining how well the models can classify previously unseen data. The accuracy results of RF, SVM, and NN are then recorded and presented as output. This information provides a clear understanding of the relative performance of each model in classifying Cb and Non-Cb cloud types. After all evaluations are completed, the process flow ends. The output of this process can be used to select the best model according to the needs and gain a deeper comprehension of the model's capabilities in classifying data.

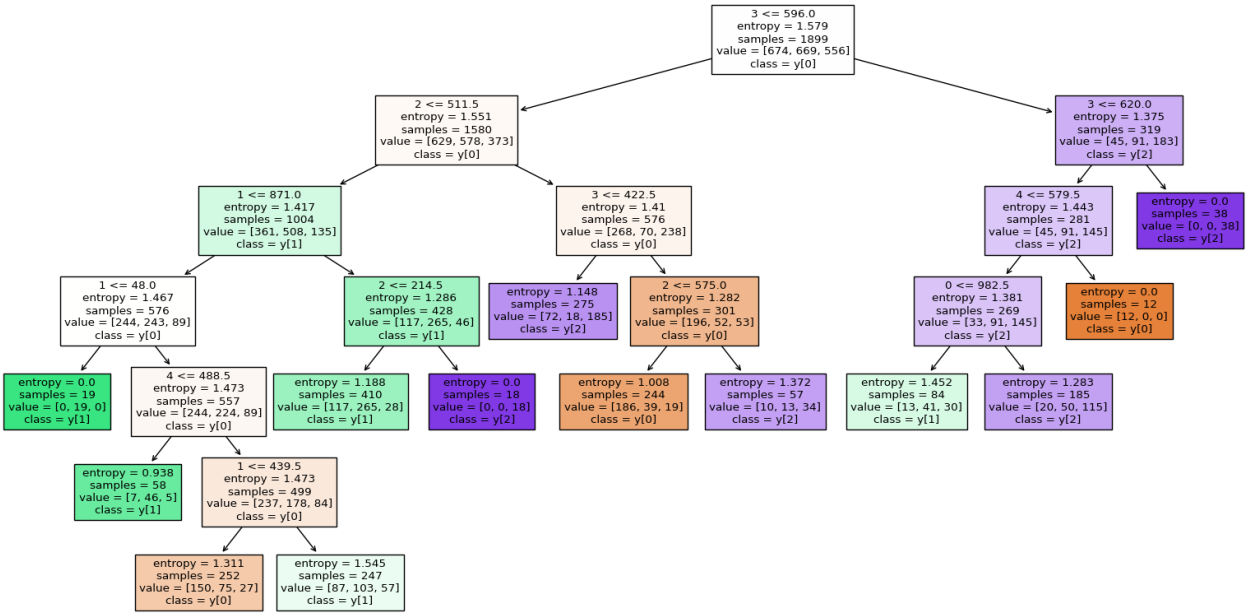

Fig. 9: Random Forest Algorithm Architecture.

In fig. 9, shows the code above is a program written in the Python language utilizing libraries such as NumPy, sci-kit-learn, and Matplotlib to perform various tasks related to data processing and machine learning. Initially, the program loads data from the “data1_samples.txt” file and separates features and labels. Then, the data is split into training and testing sets with a proportion of 70% and 30% respectively using the `train_test_split` function from scikit-learn.

Next, the program trains a RandomForestClassifier classifier with specific parameters defined beforehand. After training the model using the training set, the program saves the model to the "model1.sav" file using the pickle library. Then, the program prints the overall accuracy of the model on the testing set using the `model. score` method.

Finally, the program visualizes one of the trees in the trained random forest. This is done by selecting one of the estimators (trees) from the model and plotting it using the `plot_tree` function from the Matplotlib library. The size of the figure can be adjusted using the `fig size` parameter, and the feature names and class names can also be customized as needed.

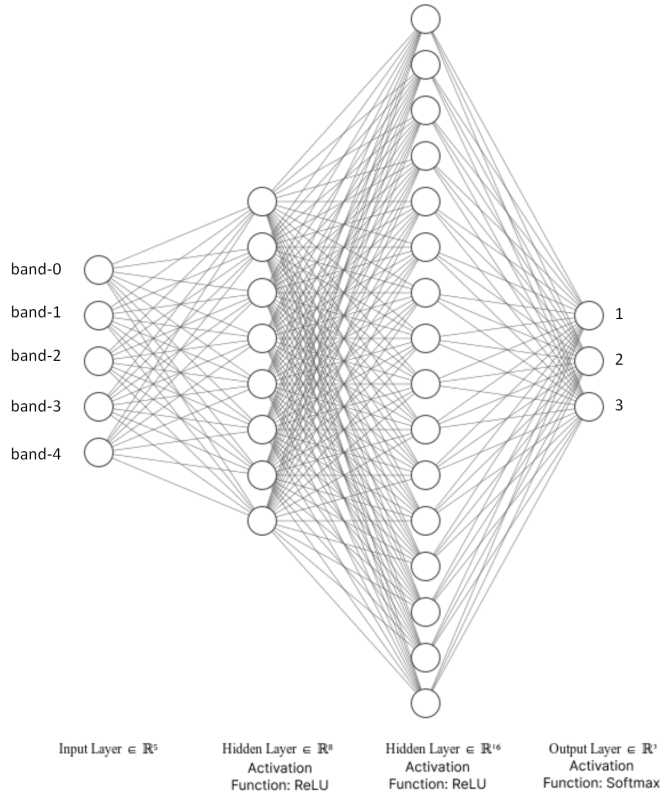

Fig. 10: Neural Network Algorithm Architecture.

In the realm of neural network architecture in Fig. 10, a sophisticated framework unfolds, comprised of intricate layers designed to process and extract meaningful insights from data inputs. This architecture boasts a structure consisting of two hidden layers, one input layer, and one output layer, each playing a pivotal role in the network's computational prowess. At the forefront of this architecture lies the input layer, a gateway through which data streams, comprising five distinct inputs. These inputs serve as the initial fodder for the neural network's analytical machinery, priming the subsequent layers for deeper processing and analysis.

In the neural network's journey, the first hidden layer, powered by ReLU activation, initiates computation with eight nodes, uncovering initial patterns. Progressing, the second hidden layer intensifies complexity with sixteen ReLU nodes, delving deeper into abstraction. At the finale, the output layer emerges, presenting three nodes for potential classifications. Utilizing softmax activation, it delivers the highest classification, symbolizing the network's analytical mastery in revealing three distinct classes, with the top class epitomizing its prowess and potential.

The configuration of each model has been predetermined to ensure fair comparison, to compare and analyze the performance of three different types of machine learning models: Neural Network (NN), Support Vector Machine (SVM), and Random Forest (RF), in solving a classification problem. This study will compare the performance of the three models using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. The configuration of each model has been predetermined to ensure fair comparison, to compare and analyze the performance of three different types of machine learning models: Neural Network (NN), Support Vector Machine (SVM), and Random Forest (RF), in solving a classification problem. This study will compare the performance of the three models using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score.

Firstly, for the Neural Network (NN) model, the network structure has been specified with two hidden layers having 8 and 16 units respectively, followed by ReLU activation for the hidden layers and Softmax activation for the output layer. Optimization is done using the Adam optimizer with a learning rate of 0.001, and the training process is carried out with a batch size of 5 for 200 epochs. Categorical

crossentropy is used as the loss function to measure the model's performance.

Next, for the Support Vector Machine (SVM) model, default parameters are used, such as the error penalty parameter (C) with a value of 1.0, and the 'rbf' kernel as the kernel type used. The gamma parameter is also set to the default value 'scale' for the RBF, polynomial, and sigmoid kernels. Finally, for the Random Forest (RF) model, 100 decision trees are used with predetermined specifications, including the maximum number of features considered, the minimum number of samples required to be a leaf node, and the minimum number of samples required to split an internal node. The bootstrap setting is set to False, meaning bootstrap is not used when building the trees.

4. Result and Discussion

After going through meticulous and structured steps in processing data meticulously designed, this research presents its findings through a presentation that clearly and comprehensively reveals the performance of the three considered machine learning algorithms. The generated data not only displays the success rates of each algorithm but also provides deep insights into the analysis process of Cb weather classification. In the presented format, every aspect of Cb classification is carefully explored to ensure the accuracy and sustainability of the results. The model accuracy percentages at the testing stage become the main focus, allowing researchers to understand the extent to which each algorithm can predict different categories of this weather phenomenon. Additionally, the analysis delves deeper to highlight the performance of each algorithm in identifying typical patterns that may be contained within complex and diverse Cb data. Not stopping at accuracy alone, this presentation also explores the capabilities of the three algorithms in handling extreme situations or nonlinear patterns that may occur in Cb classification. This provides a comprehensive view of the accuracy and reliability of each algorithm in diverse and complex contexts.

4.1. Result Model Testing Percentage

In this study, to enhance the model's ability to predict 3 cloud classes, further steps have been taken by incorporating original image data in addition to the raster satellite image data previously used. This addition allows for obtaining more detailed and accurate information regarding the visual characteristics of clouds. A comparison of prediction results from the three models proposed by previous research has been conducted. This includes the Random Forest (RF) algorithm as utilized by (Zhang et al. 2016), Support Vector Machine (SVM) as proposed by (Hemalatha et al. 2021), and also Neural Network (NN) as presented by (Kumar Verma & Mohanty 2013).

Table 2: The dataset tested prediction results.
(a). Accuracy of RF, SVM, and NN algorithm models with splitting dataset 0.7: 0.3

	0.7: 0.3											
	RF				SVM				NN			
Amount of data	Precision	Recall	Accuration	F1 score	Precision	Recall	Accuration	F1 score	Precision	Recall	Accuration	F1 score
7500	0.72	0.70	0.72	0.70	0.74	0.45	0.51	0.38	0.6	0.59	0.59	0.57
8991	0.75	0.72	0.74	0.73	0.70	0.45	0.51	0.39	0.62	0.58	0.58	0.54
10000	0.69	0.63	0.75	0.64	0.55	0.34	0.65	0.28	0.60	0.59	0.64	0.57
14090	0.69	0.65	0.68	0.66	0.64	0.36	0.51	0.30	0.48	0.46	0.46	0.41

(b). Accuracy of RF, SVM, and NN algorithm models with splitting dataset 0.8: 0.2

Amount of data	0.8: 0.2											
	RF				SVM				NN			
	Precision	Recall	Accuration	F1 score	Precision	Recall	Accuration	F1 score	Precision	Recall	Accuration	F1 score
7500	0.74	0.73	0.75	0.73	0.74	0.46`	0.55	0.40	0.17	0.41	0.41	0.24
8991	0.73	0.71	0.72	0.72	0.70	0.45	0.52	0.40	0.55	0.52	0.52	0.50
10000	0.72	0.63	0.74	0.65	0.53	0.34	0.61	0.26	0.17	0.41	0.65	0.24
14090	0.69	0.66	0.69	0.67	0.69	0.39	0.53	0.33	0.55	0.52	0.50	0.50

(c). Accuracy of RF, SVM, and NN algorithm models with splitting dataset 0.9: 0.1

Amount of data	0.9: 0.1											
	RF				SVM				NN			
	Precision	Recall	Accuration	F1 score	Precision	Recall	Accuration	F1 score	Precision	Recall	Accuration	F1 score
7500	0.71	0.69	0.71	0.70	0.72	0.44	0.51	0.37	0.18	0.42	0.42	0.25
8991	0.70	0.68	0.69	0.67	0.69	0.44	0.50	0.38	0.52	0.45	0.45	0.37
10000	0.71	0.63	0.76	0.66	0.55	0.34	0.67	0.28	0.60	0.65	0.65	0.52
14090	0.69	0.65	0.68	0.66	0.69	0.38	0.52	0.32	0.25	0.5	0.5	0.33

In building the model, the dataset tested was varied as can be seen in Table. 2 (a), (b), and (c). The total dataset produced was 14090 data originating from 3 satellite image samples used, namely satellite image dated 040120221907 (local) 1207 UTC, 060120220603 (local), and 0701220552 (local) 1207 UTC. Subsequently, this created dataset will be utilized as training data for constructing machine learning (ML) models. The use of GDAL as the primary tool in dataset construction ensures that geospatial data can be processed and understood accurately before being used in the ML model development process. Thus, the integration between GDAL and the ML model development process is key in this research to ensure the accuracy and reliability of the analyses conducted. To test the prediction results, the dataset is divided into a training set and a test set. In this study, three different data splitting sets are used: 70%:30%, 80%:20%, and 90%:10%.

- Data splitting with a 70%:30% ratio means that 70% of the data is used as the training set, while the remaining 30% is used as the test set.
- Data splitting with an 80%:20% ratio means that 80% of the dataset is allocated for training purposes, while the remaining 20% is designated for testing.
- Data splitting with a 90%:10% ratio means that 90% of the data is used as the training set, while the remaining 10% is used as the test set.

By using these three different data splitting sets, it is expected to comprehensively test the model's performance and gain a better understanding of how well the model can generalize to unseen data. This will help evaluate and compare the model's performance in various data-splitting scenarios. The results of predictive testing using three different machine learning algorithms, namely Neural Network, Support

Vector Machine, and Random Forest, on a dataset divided in a 70:30 ratio, yielded interesting findings. The dataset used varied in size, with 7500, 8991, 10000, and 14090 data points.

The accuracy percentages of Cb classification for Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) algorithms across various split scenarios are presented in the following table. These scenarios encompass different proportions of the dataset allocated to the training and testing phases. In the initial split scenario, where 70% of the data is designated for training and 30% for testing, the Random Forest algorithm exhibits the highest accuracy, achieving an impressive 75%. In comparison, the SVM and NN algorithms yield slightly lower accuracies, hovering around 65%.

This discrepancy highlights the efficacy of the Random Forest approach in accurately classifying Cb instances under this specific distribution of training and testing data. Upon examination of alternative split scenarios, a consistent trend emerges: the Random Forest algorithm consistently outperforms SVM and NN counterparts, maintaining its status as the top performer. Across varying proportions of training and testing data, Random Forest consistently achieves accuracies exceeding 76%. This robust performance underscores the versatility and reliability of the Random Forest methodology in Cb classification tasks. Furthermore, an important observation arises regarding the relationship between training set size and classification accuracy. As the proportion of the dataset allocated to training increases, there is a discernible uptrend in accuracy across all algorithms.

This trend suggests that a larger training set facilitates better model learning and generalization, leading to improved classification performance across the board. In summary, the results elucidate the superior performance of the Random Forest algorithm in Cb classification tasks across diverse split scenarios. Additionally, they underscore the positive correlation between training set size and classification accuracy, emphasizing the significance of ample training data in achieving optimal model performance.

Based on the result, optimization is still needed to increase the accuracy. Different machine learning or deep learning methods may be used so the accuracy can be improved. The quest for optimizing Cb classification accuracy extends beyond the current algorithms, inviting exploration of ensemble learning, deep learning methodologies, and innovative techniques such as transfer learning, attention mechanisms, and data augmentation. By harnessing the collective power of these approaches, practitioners can propel Cb classification performance to new heights, unlocking unprecedented levels of accuracy and efficacy in meteorological analysis and prediction.

Table 3: Model Testing Percentage (%) of Cb Classification

Model Testing Percentage (%) of Cb Classification								
Split 0.7: 0.3			Split 0.8: 0.2			Split 0.9: 0.1		
RF	SVM	NN	RF	SVM	NN	RF	SVM	NN
75	65	66	76	66	68	77	67	63

After compiling and analyzing a dataset comprising three distinct classes and conducting rigorous training and validation processes utilizing various models, the findings are presented in Table. 3 reveals the performance metrics derived from a different dataset, specifically the data123_samples_random_10000.txt. Among the models evaluated, the Random Forest (RF) algorithm emerges as particularly noteworthy, demonstrating a commendable accuracy level surpassing 70%. Such a robust performance underscores the suitability of the RF model for integration into the ongoing research efforts.

Moreover, the SVM and NN models also exhibit satisfactory accuracy levels, albeit slightly trailing behind the RF model. This collective success in accurately classifying Cumulonimbus (Cb) and Non-Cumulonimbus (NonCb) cloud types underscores the efficacy of these models in capturing the inherent complexities of the dataset. These results instill confidence in the utility of employing the RF, SVM, and NN models within the current research framework. The discernible success achieved in classifying

cloud types imparts a significant level of assurance regarding the potential contributions of these models to the broader research objectives. Overall, the demonstrated performance of these models not only validates their efficacy but also positions them as valuable assets in furthering the objectives of this research endeavor. After conducting the testing, the research visualized the distribution of Cb cloud detection results using a scatterplot as shown in Figure 11.

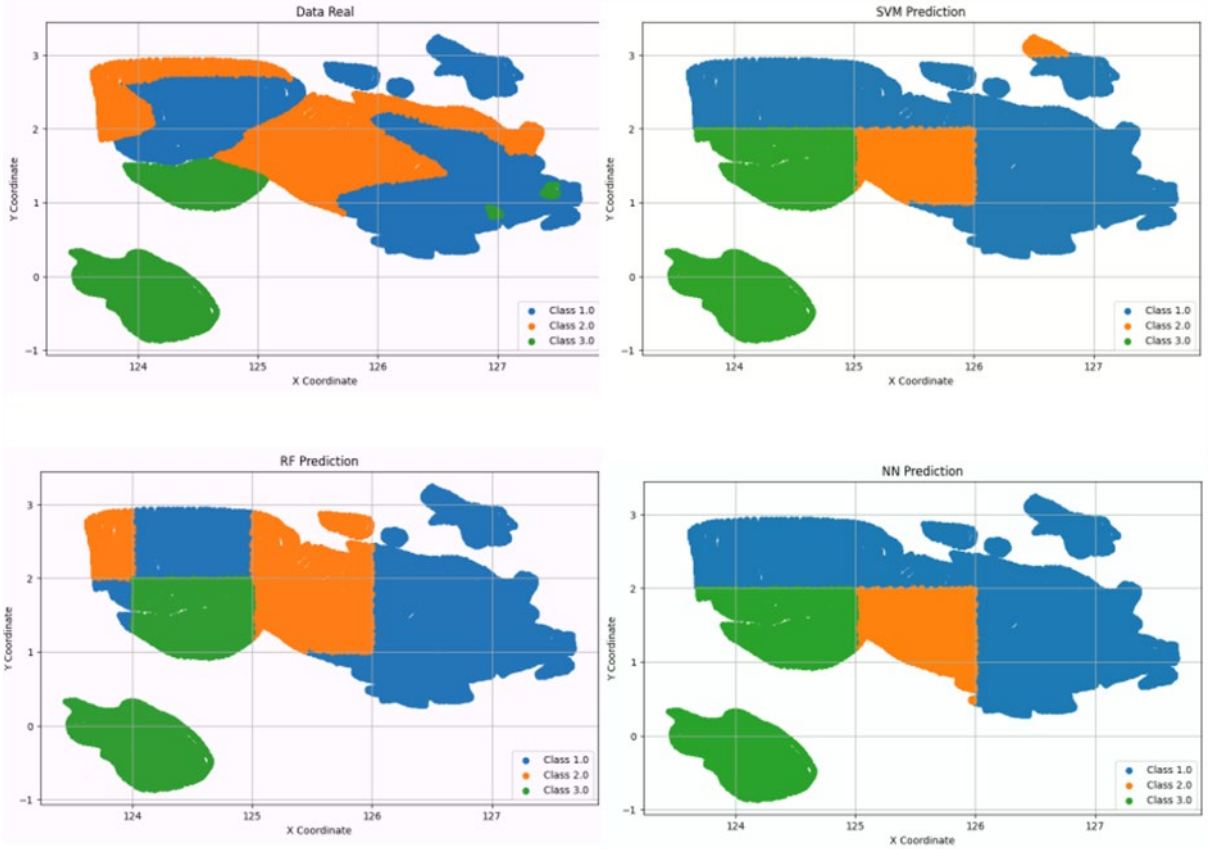

Fig. 11: Scatterplot visualized the distribution of Cb cloud detection

Figure 11 shows the prediction results compared to the real raster, in this case using sample data from January 5, 2022, at 23:03 UTC. The best prediction was produced by the RF algorithm, as it is known that the RF algorithm performs better in classification tasks. The RF algorithm used by Zhang et al. (2016) was capable of producing more accurate storm forecasts than numerical weather prediction (NWP) algorithms. In their research, Lao et al. (2021) estimated rainfall using cloud top temperature data from the geostationary satellite FY-4A. The testing results indicate that the proposed method is able to classify cloud types and cloud locations by examining the X-coordinate (longitude) and Y-coordinate (latitude) for each class_ID. The color blue represents the detection result of Class 1, which indicates detected Cb clouds, while the orange color represents Class 2, which is NonCb clouds, and Class 3 indicates no clouds. The runtime for RF was 0.312 seconds, SVM was 2.83 seconds, and NN was 8 minutes and 30 seconds, all executed on Google Colab.

4.2. Discussion

Three models used in the classification analysis of Cb are Random Forest , Support Vector Machine, and Neural Network. Each model has different strengths and characteristics that can influence the performance and classification outcomes.

Random Forest (RF)

- RF is an ensemble learning method that constructs numerous decision trees randomly during the training process.
- The main advantage of RF lies in its ability to tackle overfitting and enhance model generalization.
- RF can handle data with unstructured features well and has a tolerance towards imbalanced data.

- RF models tend to provide stable and reliable classification results, especially in situations with many features and high complexity. Therefore, RF is more suitable for higher percentage results in Cb analysis cases.

Support Vector Machine (SVM)

- SVM is a learning method used for classification or regression, which seeks the best linear (or non-linear, with kernel) separator between data classes.
- SVM is effective in handling high-dimensional data and can deal with non-linear classification problems through the use of kernels.
- Although SVM can provide good classification results, its performance can be influenced by the choice of kernel and appropriate parameters. SVM also tends to be more sensitive to imbalanced data.

Neural Network (NN)

- NN is a machine learning model inspired by the structure and function of biological neural networks.
- With its ability to handle complex data and capture intricate patterns, NN has become a popular choice in various machine-learning tasks, including image classification.
- However, training NN requires significant computation and can be more challenging to interpret compared to simpler models like RF.
- The performance of NN is highly influenced by architecture, the number of layers, the number of neurons, and the training process used.

In the context of Cb analysis, the higher performance of the RF model may be due to its ability to handle data complexity and its tendency to produce stable models with good generalization. However, SVM and NN still play important roles in cases where data exhibits more complex or non-linear patterns, and they can provide complementary results when used alongside RF.

5. Conclusions

In conclusion, this study presents a viable methodology for predicting Cb clouds by applying machine learning to NOAA satellite images. Specifically, Random Forest models are shown to be most effective, producing over 75% prediction accuracy. While there remain opportunities to expand the dataset size and geographical areas, our initial results demonstrate the promise of this approach. With further optimization and additional datasets, the proposed technique combining NOAA satellite data and Random Forest modeling could generalize effectively to improve operational Cb cloud forecasting.

In the realm of Cb clouds analysis, the elevated efficacy of the Random Forest model can be attributed to its adeptness at managing intricate data intricacies and its inclination toward generating robust models capable of broad generalization. The Random Forest model's strength lies in its capacity to navigate through multifaceted datasets, accommodating a myriad of variables and interactions within the data structure. This adaptability renders it particularly effective in scenarios where the cloud data presents intricate relationships or subtle nuances that necessitate a comprehensive approach for accurate analysis. Furthermore, Random Forest's propensity to yield stable models enhances its reliability in capturing the underlying patterns inherent in Cb clouds formations.

Nonetheless, amidst the sophistication of Random Forest, Support Vector Machine and Neural Network retain significant relevance, particularly in instances where the data exhibits heightened complexity or nonlinear behaviors. Support Vector Machine, renowned for its proficiency in handling high-dimensional data and delineating intricate decision boundaries, excels when confronted with convoluted Cb clouds datasets that necessitate precise delineation between distinct classes or states. Similarly, Neural Network 's ability to discern intricate patterns and nonlinear relationships empowers it to unravel the intricate dynamics of Cb clouds formations, uncovering latent structures that may elude simpler models.

In practice, employing Support Vector Machine and Neural Network in conjunction with Random Forest can furnish a synergistic advantage, leveraging the unique strengths of each model to glean

comprehensive insights from Cb clouds data. While Random Forest provides a robust foundation for analysis, Support Vector Machine and Neural Network offer complementary perspectives, enriching the analytical framework and enhancing the overall accuracy and depth of understanding. By integrating multiple modeling approaches, analysts can harness the collective capabilities of these methodologies to navigate the complexities of Cb clouds analysis with heightened precision and insight.

References

Abdusyukur, F. (2023). Penerapan algoritma support vector machine (svm) untuk klasifikasi pencemaran nama baik di media sosial twitter. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *12*(1).

Adawiyah, R., & Mulyana, D. I. (n.d.). *INFORMASI (Jurnal Informatika dan Sistem Informasi) Optimasi Deteksi Penyakit Kulit Menggunakan Metode Support Vector Machine (SVM) dan Gray Level Co-occurrence Matrix (GLCM)*.

Alita, D., Fernando, Y., & Sulistiani, H. (2020). Implementasi algoritma multiclass svm pada opini publik berbahasa indonesia di twitter. *Jurnal teknokompak*, *14*(2), 86.

A. F. Rais et al., "Prediction Of Cumulonimbus (Cb) Cloud Based On Integrated Forecast System (Ifs) Of European Medium-Range Weather Forecast (Ecmwf) In The Flight Information Region (Fir) Of Jakarta And Ujung Pandang," Jurnal Sains & Teknologi Modifikasi Cuaca, vol. 21, no. 2, Dec. 2020, doi: 10.29122/jstmc.v21i2.4100.

A. F. Rais, R. M. Putra, M. A. Fitrianto, and T. Hermansyah, "A Preliminary Comparative Study on the Feasibility of a Multipurpose Numerical Weather Model for Prediction of Cumulonimbus Clouds in Indonesia," in *2022 International Conference on Science and Technology (ICOSTECH)*, Feb. 2022, pp. 1–6, doi: 10.1109/ICOSTECH54296.2022.9829049.

Fahruddin Rais, A., Setiawan, F., Yunita, R., Meinovelia, E., Fadli, M., Wijayanto, B., & Meteorologi Klimatologi dan Geofisika Jl, B. (2020). Prediction Of Cumulonimbus (Cb) Cloud Based On Integrated Forecast System (Ifs) Of European Medium-Range Weather Forecast (Ecmwf) In The Flight Information Region (Fir) Of Jakarta And Ujung Pandang Prediksi Awan Cumulonimbus (Cb) Berbasis Model Integrated Forecast System (IFS) European Medium-Range Weather Forecast (ECMWF) Pada Flight Information Region (FIR) Jakarta dan Ujung Pandang. In *Jurnal Sains & Teknologi Modifikasi Cuaca* (Vol. 21, Issue 2). https://confluence.ecmwf.int//display/FUG/HRES

G. Davis, "History of the NOAA satellite program," *Journal of Applied Remote Sensing*, vol. 1, no. 1, p. 012504, Jan. 2007, doi: 10.1117/1.2642347.

Hemalatha, G., Rao, K. S., & Kumar, D. A. (2021). Weather Prediction using Advanced Machine Learning Techniques. *Journal of Physics: Conference Series*, *2089*(1). https://doi.org/10.1088/1742-6596/2089/1/012059

Hyvärinen, O., Saltikoff, E., & Hohti, H. (2015). Validation of automatic Cb observations for metar messages without ground truth. *Journal of Applied Meteorology and Climatology*, *54*(10), 2063–2075. https://doi.org/10.1175/JAMC-D-14-0222.1

Kim, C. (2011). *The Effects Of Weather Recognition Training On General Aviation Pilot Situation Assessment And Tactical Decision Making When Confronted With Adverse Weather Conditions Recommended Citation*. https://tigerprints.clemson.edu/all_dissertations

Kleber, J., Thomas, R., Domingo, C., & Blickensderfer, B. (2022). Measuring Mental Models: General Aviation Pilots' Understanding of Preflight Weather. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1997–2000. https://doi.org/10.1177/1071181322661216

Kopacz, M., Breeze, V., Kondragunta, S., Frost, G., Anenberg, S., Bruhwiler, L., Davis, S., da Silva, A., de Gouw, J., Duren, R., Flynn, L., Gaudel, A., Geigert, M., Goldman, G., Joiner, J., McDonald, B., Ott, L., Peuch, V. H., Pusede, S. E., … Kalluri, S. (2023). Global Atmospheric Composition Needs from Future Ultraviolet–Visible–Near-Infrared (UV–Vis–NIR) NOAA Satellite Instruments. *Bulletin of the American Meteorological Society*, *104*(3), E623–E630. https://doi.org/10.1175/BAMS-D-22-0266.1

Kumar Verma, A., & Mohanty, W. K. (2013). *Quantifying Sand Fraction from Seismic Attributes using Modular Artificial Neural Network*. https://www.researchgate.net/publication/262188841

Liu, Y., Xi, D. G., Li, Z. L., & Shi, C. X. (2014). Automatic Tracking and Characterization of Cumulonimbus Clouds from FY-2C Geostationary Meteorological Satellite Images. *Advances in Meteorology*, *2014*. https://doi.org/10.1155/2014/478419

L. Peck, "THE IMPACTS OF WEATHER ON AVIATION DELAYS AT O.R. TAMBO INTERNATIONAL AIRPORT, SOUTH AFRICA," University of South Africa, 2015.

Mishra, C., & Gupta, D. L. (2017). Deep Machine Learning and Neural Networks: An Overview. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *6*(2), 66. https://doi.org/10.11591/ijai.v6.i2.pp66-73

Olaganathan, R., & Ham, R. G. (2020). Significance of Incorporating Weather Technology Training for GA Pilots to Curb Fatalities. *International Journal of Aviation, Aeronautics, and Aerospace*, *7*(2). https://doi.org/10.58940/2374-6793.1478

Rossow, W. B., Knapp, K. R., & Young, A. H. (2022). International Satellite Cloud Climatology Project: Extending the Record. *Journal of Climate*, *35*(1), 141–158. https://doi.org/10.1175/JCLI-D-21-0157.1

R. M. Putra et al., "Cumulonimbus cloud prediction based on machine learning approach using radiosonde data in Surabaya, Indonesia," IOP Conference Series: Earth and Environmental Science, vol. 724, no. 1, p. 012047, Apr. 2021, doi: 10.1088/1755-1315/724/1/012047.

Saidah, H., Saptaningtyas, R. S., Hanifah, L., Jaya Negara, I. D. G., & Widyanty, D. (2020). NOAA Satellite performance in estimating rainfall over the Island of Lombok. *IOP Conference Series: Earth and Environmental Science*, *437*(1). https://doi.org/10.1088/1755-1315/437/1/012035

Seth, N. (2024). A Predictive Study of Machine Learning and Deep Learning Procedures Over Chronic Disease Datasets. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, *42*, 34–47. https://doi.org/10.55529/jaimlnn.42.34.47

Temme, M. M., & Tienes, C. (2018). Factors for pilot's decision making process to avoid severe weather during enroute and approach. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2018-September*. https://doi.org/10.1109/DASC.2018.8569357

Tuomola, L., Ylhäisi Heikki Järvinen Examiners, J., & Järvinen Dos Marja Bister, H. (2021). *Cumulonimbus cloud detection with weather radar at Helsinki-Vantaa airport*.

T. Zhang et al., "GraphAT Net: A Deep Learning Approach Combining TrajGRU and Graph Attention for Accurate Cumulonimbus Distribution Prediction," Atmosphere, vol. 14, no. 10, p. 1506, Sep. 2023, doi: 10.3390/atmos14101506.

Wan, L., Gong, K., Zhang, G., Yuan, X., Li, C., & Deng, X. (2021). An efficient rolling bearing fault diagnosis method based on spark and improved random forest algorithm. *IEEE Access*, *9*, 37866–37882. https://doi.org/10.1109/ACCESS.2021.3063929