COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

YOLOv8-based detection of convective storm clouds for cumulonimbus classification

Yenniwarti Rafsyam*, Shita Fitria Nurjihan, Arief Rinaldi

Department of Electrical Engineering, Politeknik Negeri Jakarta, Depok 16425, Indonesia

Article history:
Received: 3 December 2025 / Received in revised form: 18 December 2025 / Accepted: 22 December 2025

Abstract

Cumulonimbus (CB) clouds are vertically developed convective systems that are capable of producing severe weather phenomena, including turbulence, heavy rainfall, and lightning. These phenomena pose a significant threat to aviation safety. This paper considers an automated CB cloud detection approach using the deep learning algorithm You Only Look Once version 8 on NOAA-19 satellite imagery. The images of 640 × 640 pixels each were labeled into two classes: CB and non-CB. In general, rotation, flip, and random brightening are performed to develop a more robust model. After 100 training epochs, the proposed model produced reliable detection performance, as evidenced by 1,694 TP (true positives), 438 FP (false positives), and 304 FN (false negatives) cases, with a precision of 0.79, recall of 0.84, and an F1-score of 0.81. Validation using METAR reports from the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG) confirmed the consistency of the model with observed weather conditions. The results demonstrated that YOLOv8 could provide a rapid and reliable framework for real-time detection and classification of CB clouds, thereby enhancing situational awareness for aviation operations and facilitating the effectiveness of satellite-based early warning systems in convectively active tropical regions.

Keywords: Aviation safety; cloud detection cumulonimbus clouds; deep learning; YOLOv8

1. Introduction

The Cumulonimbus (CB) clouds are towering convective clouds formed by rapidly rising moist air into the upper atmosphere. They are frequently associated with severe weather such as thunderstorms, heavy rainfall, lightning, and strong turbulence that pose a significant threat to aviation safety, particularly in tropical regions. Conventional CB monitoring using radar and visual observations remains spatially limited and delayed; leading to a need for automated, real-time detection using satellite imagery [1,2]. Previous studies have utilized infrared thresholding to identify overshooting tops as key indicators of CB activity [3]. Furthermore, the integration of geostationary satellite data with radar has been employed to detect both CB and TCU clouds [4,5]. The advancements in image processing and machine learning have rendered deep learning–based models such as YOLO effective instruments for cloud detection and classification [6]. YOLOv8, the latest version, features a lightweight architecture with a backbone, neck, and head for multi-scale feature extraction, thus facilitating rapid and accurate detection of small objects in complex remote sensing imagery [7,8,9,10]. Supported by Ultralytics' open-source framework, YOLOv8 is applicable to a wide range of real-world scenarios [11,12,13]

In this study, YOLOv8 was implemented to automatically detect CB clouds using NOAA-19 satellite imagery. The images were manually annotated into CB and non-CB classes using LabelMe [14,15,16] and trained with data augmentation techniques. The performance of the model was evaluated using precision, recall, and F1-score metrics, while the validation of the model was performed using METAR weather reports from BMKG. The objective of this approach is to facilitate the development of rapid, accurate, and reliable satellite-based early warning systems to enhance aviation safety in regions that are prone to convective weather events [17,18,19,20].

The detection and classification of convective clouds, with a particular focus on Cumulonimbus (CB), have long been central to meteorological research, given their significant impact on aviation safety. Early studies by Baum et al. [21] and Bankert [22] introduced automated cloud classification using multispectral satellite imagery. This work utilized fuzzy logic and probabilistic neural networks applied to AVHRR data. Berendes et al. [23] advanced this work with an adaptive clustering method for classifying convective clouds, while Donovan et al. [24] emphasized the identification of hazardous convective cells over oceans using visible and infrared data to

* Corresponding author.
Email: yenniwarti.rafsyam@elektro.pnj.ac.id
https://doi.org/10.21924/cst.10.2.2025.1854

This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license

support early warning systems. Further progress in this field was achieved by Carbajal Henken et al. [25], who integrated cloud properties from MSG-SEVIRI with radar observations to enhance detection accuracy, and by Mecikalski and Bedka [26], who developed nowcasting techniques to forecast convective initiation using GOES imagery. Rosenfeld et al. [27] further contributed to this field by developing a microphysical approach by retrieving vertical profiles of particle size and thermodynamic phase to identify severe storms. Zinner et al. [28] developed Cb-TRAM, a real-time system for tracking convective storm life cycles from initiation to maturity. This system is conceptually similar to modern computer vision models such as YOLOv8. Schmetz et al. [29] and Levizzani and Setvák [30] highlighted the value of Meteosat and high-resolution multispectral data in identifying overshooting tops and fine storm-top structures. Building on these foundations, this present study employed the deep learning–based YOLOv8 model to develop an automatic, real-time, and precise CB detection system adaptive to extreme weather dynamics, thereby enhancing both satellite-based early warning capabilities and aviation safety. Alvira et al. [31] conducted a review of the various pretreatment technologies for lignocellulosic biomass to enhance enzymatic hydrolysis in bioethanol production. The study evaluated how different pretreatment methods modify cellulose, hemicellulose, and lignin structures, and analyzed their impacts on sugar recovery, inhibitor formation, energy consumption, and process cost. The review emphasized that pretreatment is a critical and cost-intensive step that significantly determines the overall efficiency of lignocellulosic bioethanol conversion. In their study, Kun-Cahyono et al. [32] applied multiple machine learning and deep learning models to predict daily rainfall. These models were trained on open-access remote sensing data processed through Google Earth Engine. The results of the study demonstrated that Support Vector Regression provided the most reliable performance, thus highlighting the feasibility of using satellite-derived atmospheric variables for operational rainfall forecasting.

Despite the considerable adoption of YOLO-based object detection within recent literature, the significance of this paper relates to the examination and validation of YOLOv8 within the context of convective storm cloud detection, as acquired by the use of polar-orbiting satellite imagery supplied by the NOAA-19 satellite within tropical geography. In opposition to the vast majority of recent YOLOv8 literature, which focuses on generic object detection tasks or ground-level images, this paper focuses on the distinct spatial, spectral, and morphological attributes of Cumulonimbus clouds. Additionally, the utilization of METAR data supplied by BMKG in this study relates to examination within a meteorological capacity, rather than solely relying on computer vision evaluation.

2. Materials and Methods

2.1. Dataset and Preprocessing

This study utilized satellite imagery acquired from the NOAA polar-orbiting satellite to detect Cumulonimbus (CB) clouds. The dataset consisted of RGB composite images covering tropical regions with frequent convective activity. The satellite data were obtained through the NOAA Automatic Picture Transmission (APT) system using a self-developed ground receiving station, allowing direct acquisition of raw satellite imagery without reliance on third-party data providers. Rafsyam et al. [33] demonstrated that a double cross dipole antenna operating at 137–138 MHz can receive NOAA satellite signals and converting voice data into cloud images. Fig. 1 illustrates the original satellite image acquired from the receiving system.

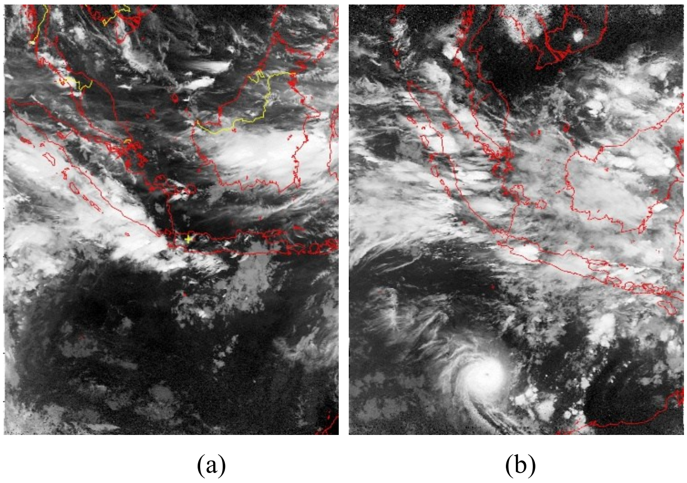
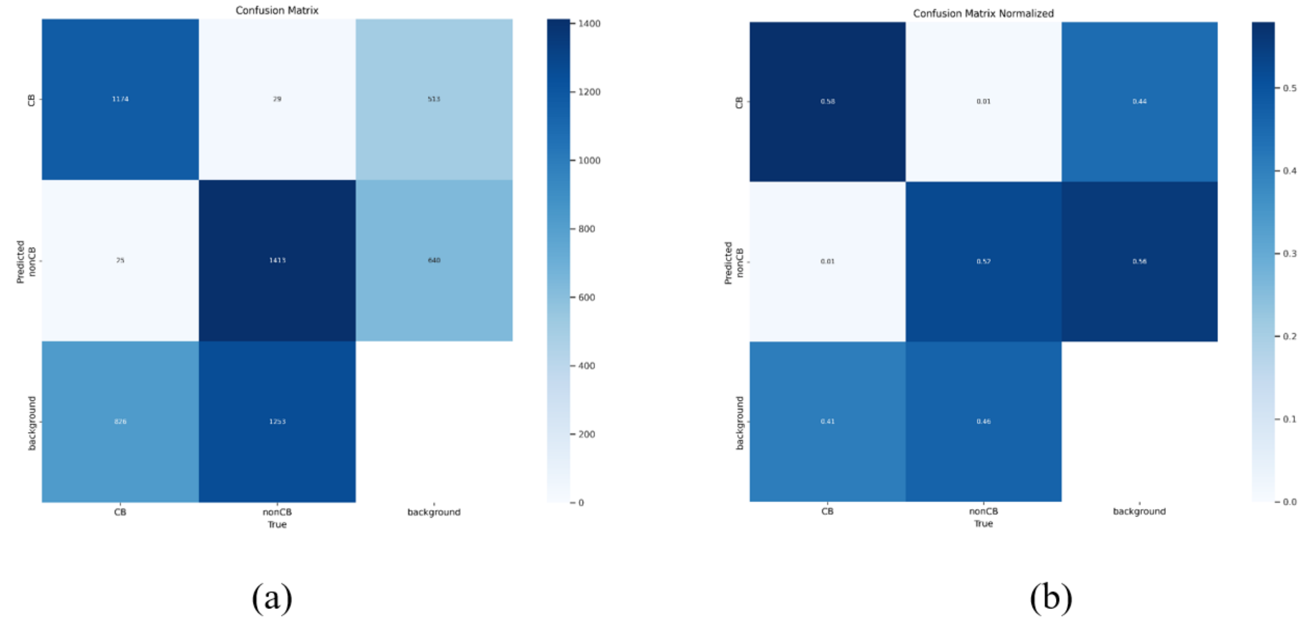


Fig. 1. Example of the original satellite image (a) noaa-19-202202030016-contrastb.jpg (b) noaa-19-202303311318-contrast.jpg

As demonstrated in Fig. 1, the original NOAA-19 satellite images used in this study were obtained under different observation conditions. All images were resized to 640 × 640 pixels to ensure compatibility with the YOLOv8 input requirements. Standard preprocessing steps, including pixel normalization and image scaling, were applied to enhance training stability. The dataset was manually annotated using the Labeling tool and categorized into three classes: Cumulonimbus (CB), non-convective clouds, and background, following established meteorological cloud classification principles.

Fig. 2(a) presents the confusion matrix of the YOLOv8 detection results for three categories: Cumulonimbus (CB), non-CB, and background. The diagonal elements indicate the correct samples, while the off-diagonal entries indicate misclassified instances among the various classes. The system demonstrated an accuracy of 1,694 CB samples, which thus indicated its capacity to recognize mature convective structures of CBs. However, misclassifications of 299 CB samples as background, as well as 2 as non-CB, demonstrated the possibility of inadequate distinctive visual cues in weak or immature convective systems to always be distinguished. Additionally, within the non-CB class, 2,550 samples were correctly classified, but 436 background samples were misclassified as CB. This may result in the conveyance of visual textures or brightness information similar to CBs in specific background areas.



(a)

(b)

Fig. 2. (a) Confusion matrix of YOLOv8 classification results for CB, nonCB, and background classes, (b) Normalized confusion matrix for YOLOv8 classification of CB, nonCB, and background

Fig. 2(b) illustrates the normalized confusion matrix. The columns and rows correspond to the reference classification and the classification, respectively. The value of 0.58 for CB-CB indicates that approximately 58% of the CB pixels have been accurately classified, while the value of 0.44 for CB-background indicates that a significant amount of background pixels has been confused with CB. The occurrence of this confusion is predominantly related to the cloud boundaries, cloud edges, and the stratified regions of clouds that resemble the convective cloud top features observed in the optical satellite images. The normalized confusion matrix illustrates that misclassifications occur more frequently between cloud-related classes than between cloud pixels or clear pixels. The confusion matrix analysis indicates that a trade-off is present between the sensitivity to the convective features and the false alarms in the ambiguous regions as identified by other studies related to cloud classification from satellite images.

2.2. YOLOv8 model architecture

In this work, the proposed detection framework is constructed on the YOLOv8 architecture, specifically designed to perform real-time object detection with robust multi-scale feature representation. Fig. 3 depicts the overall architecture employed in this work and is divided into three parts: a backbone, a neck, and an anchor-free detection head. The backbone of the model uses CSPDarknet53 in extracting hierarchical spatial and semantic features from NOAA satellite imagery. while it uses Conv2D layers, Batch Normalization, SiLU activation functions, and C2f residual blocks to improve feature reuse without compromising computational efficiency. This structure has proven to be of particular important for satellite-based cloud detection, given that cloud boundaries are frequently diffuse, and radiative contrasts tend to be quite subtle.

A feature map is obtained at three different spatial resolutions, 80×80, 40×40, and 20×20, to capture the small-, medium-, and large-scale cloud structure. These multi-resolution representations facilitate the model to capture the extensive range of spatial variability exhibited by that Cumulonimbus clouds, ranging from compact convective cells at early development stages to large, mature systems with extensive anvils. This neck module amalgamates these features via upsampling and concatenation operations inspired by Feature Pyramid Networks and Path Aggregation Networks, respectively. These operations jointly leverage both fine-grained local information and high-level contextual features. The adoption of enhanced C2f and CBS modules serves to reinforce cross-scale feature fusion, which is of great importance to discriminate CB clouds from surrounding non-convective cloud fields and backgrounds.

This study, as illustrated in Fig. 3 employs the YOLOv8 architecture, that encompass the backbone, neck, and head components. These components are applied for the detection of CB clouds at multi-scale. Multi-resolution feature extraction has been identified as a key component in the analysis of spatial variability related to the structure of convective clouds. This approach has been identified as a primary requirements for cloud detection in previous works using deep learning models [7,10]. The anchor-free prediction mechanism is considered in the detecting head, which contains three parallel branches corresponding to different object scales. Each branch predicts the coordinates of bounding boxes, confidence scores, and class labels for CB and non-CB clouds, being flexible to handle irregular shapes with variable aspect ratios that characterize the formations of convective clouds. Non-Maximum Suppression (NMS) is finally applied to remove redundant detections arising from overlapping predictions, ensuring that each CB cloud system is represented by a single, most confident bounding box. The architecture, as illustrated in Fig.3, facilitates the effective detection of Cumulonimbus clouds at multiple scales from polar-orbiting satellite imagery, thereby forming the basis for the subsequent experimental observation and results.

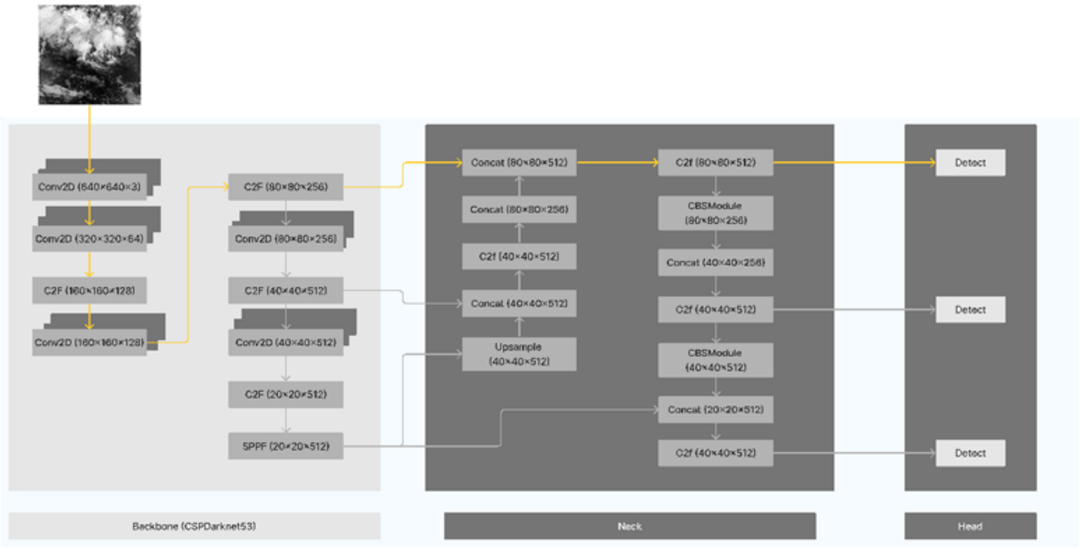


Fig. 3. YOLOv8 architecture used in this study to detect CB clouds from NOAA satellite imagery

To ensure that the performance of the proposed YOLOv8-based detection framework is not biased by a specific data partition, a cross-validation-based dataset validation strategy was applied prior to model training. The annotated NOAA satellite image dataset was divided into K mutually exclusive subsets (folds) of approximately equal size, while maintaining representative spatial and temporal variability of cloud patterns. In each iteration, one-fold was designated as the validation set, with the remaining K−1 folds being allocated for training. The overall performance of the model was computed by averaging the evaluation metrics across all folds, which is expressed as:

$$CV_{score} = \left(\frac{1}{K}\right) \sum_{i=1}^{K} M_i \qquad (1)$$

where K denotes the number of folds and M_i represents the evaluation metric obtained from the iii-th fold, including precision, recall, mAP@0.5, and mAP@0.5:0.5:0.95.

This validation strategy ensures that the multi-scale detection capability of the YOLOv8 architecture, as illustrated in Fig. 4, is evaluated under diverse atmospheric conditions and cloud morphologies. Fig. 4 presents a schematic overview of the cross-validation procedure employed in this study.

The detection visualization, as illustrated in Fig. 4, is an indication of the result of the proposed YOLOv8 model's approach to the NOAA satellite images of the region employed in the K-Fold cross-validation process. Each sub-figure of the image denotes the capacity of the proposed YOLOv8 model to perform multi-scale detections of cloud features, which fits well with the aim of testing the proposed approach to different atmospheric conditions. The [Pink/Red] bounding-box rectangles around the images illustrate the positioning of detected cloud features (e.g., spiral bands and convective centers), along with the detected class and the confidence level of each proposed YOLOv8 approach. The dense distribution of bounding-box rectangles around images indicates that the proposed approach to YOLOv8 architecture performs well in handling different detections of images at various scales. Nevertheless, it also signifies that the approach might have double-detections and the complexity of demarcating boundaries between the detected cloud features in images.

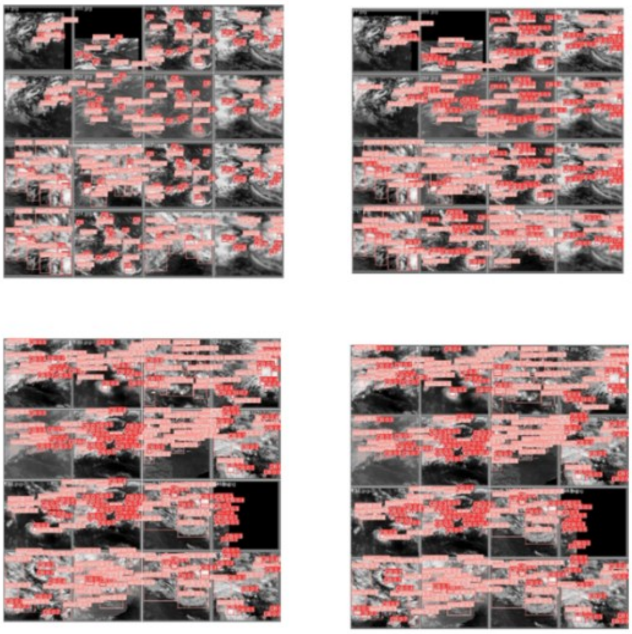


Fig. 4. Visualization of YOLOv8 detection results on NOAA satellite imagery using K-Fold Cross-Validation

2.3. Training strategy and hyperparameters

Model training is conducted using a supervised learning approach with a five-fold cross-validation strategy with the aim of enhancing generalization and reducing overfitting. Each fold maintains a balanced ratio of CB and non-CB samples. To enhance robustness against cloud variability, data augmentation techniques including horizontal flipping, color jittering, and random cropping, are applied. The model has been trained using standard YOLOv8 hyperparameters provided by the Ultralytics framework.

Fig. 5(a) illustrates the F1-Confidence Curve, in terms of the correlation of model confidence and F1-score. From the data given, the optimal F1-score of 0.88 is attained at the confidence threshold of 0.355. NonCB maintains a constantly high F1-score, while CB exhibits more fluctuations, thereby facilitating the determination of the threshold balancing the detection accuracy and misclassification. Fig. 5(b) presents the number distribution of the detected instances and spatial pattern of bounding boxes. It is observed that NonCB instances exhibit a slight numerical advantage over CB. From the x-y scatter plot, it is relatively even in space. Most of the bounding boxes are small, which is indicative of the typical cloud size and informs further anchor box scales and preprocessing. Fig. 5(c) presents pairwise relationships of bounding box parameters (x-center, y-center, width, height). The majority of boxes are characterized by compactness, with weak correlations between width and height for most boxes, indicating varied cloud shapes. This approach thus provides insights into the bounding box regression to further improve irregular cloud formations detection.

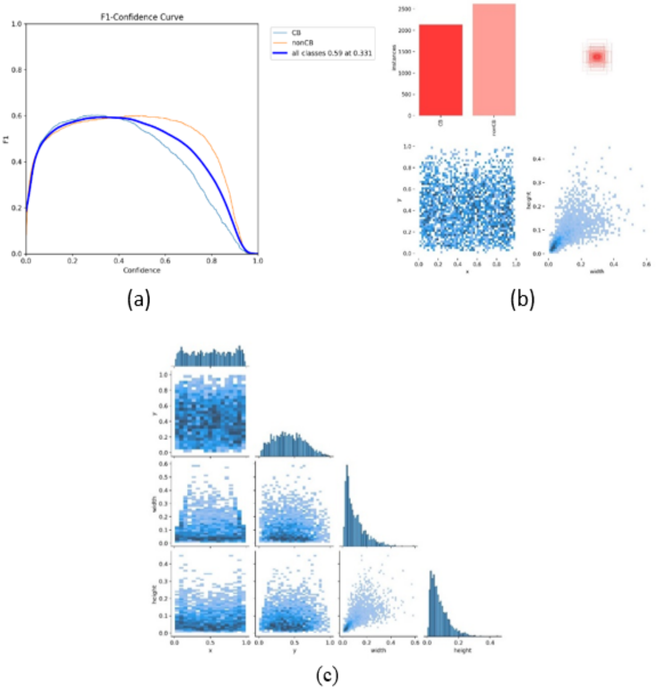


Fig. 5. (a) Confusion matrix of YOLOv8 classification results for CB, nonCB, and background classes, (b) Normalized confusion matrix for YOLOv8 classification of CB, nonCB, and background, (c) Airwise plot of bounding box parameters: x-center, y-center, width, and height

Fig. 5 provides more detailed model's behavior during detection than do overall accuracy metrics. The F1–Confidence curve illustrates the sensitivity of detection performance to confidence threshold selection, whereas bounding box distribution and pairwise plots demonstrate the geometric properties of detected cloud objects. Similar analyses have been previously employed in several YOLO-based remote sensing studies to evaluate the reliability of detection and the spatial consistency of detection [6,13].

2.4. Evaluation metrics

Standard object detection metrics include Precision, Recall, F1-score, and mean Average Precision, which are utilized to evaluate the model's performance. Precision and recall are designed to measure correctness and completeness, respectively, whereas the F1-score is a balanced metric that provides both of these qualities. The mAP metric evaluates the performance of detection at various confidence thresholds and IoU criteria. Another supplementary indicator for operational reliability is the confidence-threshold-based accuracy metric.

Fig. 6(a) illustrates Precision-Confidence Curve that describes how model precision changes regarding its confidence threshold. The maximum precision equals 1.00 for the threshold of 0.823. The NonCB class maintains its precision at a high level even for lower thresholds, although CB precision increases more sharply with confidence. This can be effective to identify operational thresholds when it is deemed necessary to minimize false positives in Automatic Weather Alert Systems. Precision-Recall Curve, as illustrated in Fig. 6(b) depicts the balance of accuracy and coverage. Class CB reaches an mAP@0.5 value of 0.589, while the class nonCB reaches 0.546. The overall combined mAP@0.5 is 0.568. Curves that are closer to the top right demonstrate better performance and provide additional insights into model consistency across different classes and thresholds. Fig. 6(c) presents Recall-Confidence Curve. How does recall change when the confidence threshold increases? Overall recall reaches the value of 0.74 for the threshold of 0.000 and decreases as the model becomes more conservative. Classes CB and nonCB are represented separately. It can facilitate the selection of threshold settings in operational systems where it is important to ensure that no real events are overlooked, thus minimizing false negatives.

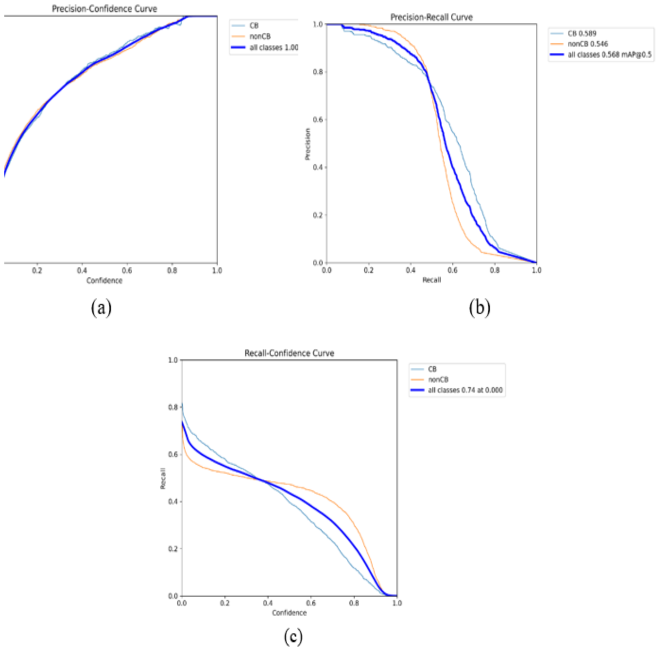


Fig.6 (a) Precision-Confidence Curve of the YOLOv8 model demonstrating how precision changes with different confidence values, (b) YOLOv8 Precision-Recall Curve showing the balance between accuracy and instance coverage, (c) Recall-Confidence Curve showing how recall varies with increasing confidence thresholds

The robust detection performance of the YOLOv8 model can be attributed to its anchor-free design and multi-scale feature aggregation, which facilitate the effective localization of irregular and spatially diverse cloud structures. The employment of multi-resolution feature maps allows the model to capture both compact convective cores and broader cloud anvils, which are the characteristic of mature CB systems. In addition, data augmentation techniques such as rotation and brightness adjustment have been shown to enhance the robustness of satellite observations against variations in viewing geometry and illumination conditions.

Equation (1) was used to calculate accuracy, representing the percentage of detections with confidence scores equal to or greater than the defined threshold of 0.50. Meanwhile, Eq. (2) was used to determine the error rate, indicating the proportion of detections with confidence scores below the threshold.

$$Accuracy\ (\%) = (TP\ /\ Total)\ x\ 100\ \% \tag{2}$$

$$Error\ Rate\ (\%) = (FP + FN)\ /\ Total)\ x\ 100\ \% \tag{3}$$

In this evaluation, the detections with confidence levels greater than or equal to 0.50 are denoted as true positives (TP) as the predictions would have been made with high confidence. The detections with confidence levels lower than 0.50 would be denoted as false positives (FP) as they would have fallen below the minimum confidence level for proper classification. The threshold value of 0.50 was selected for the purpose of comparison, with the objective of differentiating between high confidence and low confidence.

2.5 METAR-based validation

In this study, the utilization of METAR from operational weather stations is also considered for the purpose of independent meteorological validation. The METAR is utilized in validating the occurrence of convective weather events including thunderstorms, heavy rain, or cumulonimbus cloud occurrences at the time when the satellite overpass is performed. For the NOAA image to be considered in the analysis of the evaluated scene, the METAR reports within a short time frame of the satellite data acquisition time are reviewed.

It should be pointed out that METAR data are point measurements, which are taken at predefined aerodrome stations, while satellite images are used to observe clouds in a region in terms of continuous data. Consequently, the identification of Cumulonimbus clouds at specific points in relation to the predefined stations may not always be possible, particularly in cases of rapidly changing and displaced convective systems. In addition, temporal discrepancies may occur between satellite passes and METAR intervals in view of their differing frequency. This has the potential to result in partly incongruent image-based and METAR-based observations of clouds.

In spite of these limitations, the validation using METAR is of great practical importance for determining whether the identified convective system is in accordance or not with surface weather reports. In this case, rather than employing the METAR as an objective reference for validation, validation by METAR is employed as supplementary evidence in justifying the validity and practical significance of the proposed method for detection using YOLOv8 in respect to aviation safety and early warning.

3. Results and Discussion

The YOLOv8-based Cumulonimbus (CB) detection model was trained for 200 epochs, exhibiting three distinct training phases. During the initial phase, a rapid decrease in loss values was observed (box loss from approximately 3.29 to 2.20), accompanied by sharp increases in precision, recall, and mAP@0.5 from near-zero values to approximately 0.18, 0.21, and 0.12, respectively. This phase is indicative of effective learning of basic object localization and feature representations.

In the middle phase, loss values decreased steadily while precision and recall exceeded 0.50, and mAP@0.5 approached 0.50. This indicates a refinement in feature extraction and bounding box regression. In the final phase, the model reached convergence, characterized by a stabilized box loss near 1.0, precision values between 0.70 and 0.73, recall between 0.50 and 0.52, and mAP@0.5 exceeding 0.55. The learning rate decayed to approximately 3.3×10^{-5}, thereby facilitating stable weight updates and preventing oscillations during convergence. Minor fluctuations in evaluation metrics were observed and are considered normal due to batch variability and stochastic optimization.

The confusion matrix analysis indicates that most misclassifications occur between the Cumulonimbus (CB) and background classes. This behavior can be primarily attributed to the inherent limitations of single-sensor RGB satellite imagery. Early-stage or weakly developed convective clouds frequently exhibit visual characteristics that closely resemble surrounding background clouds, such as similar brightness, texture, and spatial continuity. In the absence of additional spectral or thermal information, these subtle convective signatures prove challenging to distinguish exclusively based on RGB features.

In addition, there are no physical parameters such as cloud height or cloud moisture in RGB images that are frequently employed by IR, multispectral, and radar technologies in cloud detection techniques. Consequently, there might be cases where the model is prone to misinterpreting the thin convective clouds or the cloud borders for the background in conditions where the atmosphere is in the state of transition. These results are in line with those from previous cloud classification studies that utilized satellite images, in that the imaging technology still faces similar issues where there is less use of spectral images.

These misclassification patterns facilitate comprehension of the observed trade-off between precision and recall as discussed in the subsequent performance curves. As demonstrated in Fig. 2 (and Fig. 4) the confusion matrix reveals that the YOLOv8 model achieves high true positive detection for the CB class, while the majority of misclassifications occur between cloud and background regions. This behavior is consistent with earlier studies on satellite-based cloud classification, where diffuse cloud boundaries often lead to background confusion [14,18]. The normalized confusion matrix further highlights the relative difficulty in distinguishing weak convective signatures from non-convective cloud formations.

The Precision–Recall and confidence-based performance curves as depicted in Fig. 6 (Fig. 5) further confirm the trade-off between precision and recall under varying confidence thresholds. This trade-off has also been reported in earlier cloud detection and object detection studies, emphasizing the importance of threshold selection for operational applications such as aviation safety and early warning systems [3,26].

Fig. 7 illustrates the training dynamics of the YOLOv8 model, including loss convergence and performance stabilization over 200 epochs. The decrease in training and validation losses, combined with consistent enhancement in mAP metrics, indicates the stability of the learning behavior without significant overfitting, which is comparable to training trends as reported in similar YOLO-based remote sensing studies [6,15].

Fig. 7 illustrates the YOLOv8 training process over 200 epochs, with both loss dynamics and performance metrics. The training losses—train/box_loss, train/cls_loss, and train/dfl_loss exhibited a consistent decrease, indicating an enhancement in the accuracy of bounding box accuracy and feature learning. The corresponding precision and recall metrics increased from low initial values (~0.1) to above 0.5, reflecting reduced false positives and better true positive detection. The validation losses followed a similar downward trend, slightly higher than training losses, suggesting the absence of overfitting. The Mean Average Precision metrics, mAP50 and mAP50-95, also exhibited consistent improvement, indicating the capacity for robust object localization under varying thresholds. The gradual warm-up and decay of the learning rate supported convergence and training stability. Overall, the figure confirms effective model learning, consistent performance gains, and strong generalization to unseen data.

Even with this promising performance, instances of failure in the results of the cloud detection were observed. False positives were identified in the dense stratiform clouds and the boundaries of the clouds with textured characteristics of developing Cumulonimbus clouds. Such areas tend to have higher values of reflectance and steep gradients that can easily be confused with convective systems by the model. Conversely, false negative instances were mainly identified in the early-developing and partially occult CB clouds that did not yet have evident vertical growth in the satellite image. This finding indicates that the model is better at detecting fully developed convective systems rather than developing ones.

Fig. 8 depicts the result of testing the YOLOv8 algorithm model for its applicability in the classification results validated by METAR.

As illustrated in Fig. 8, representative examples of CB and nonCB detection are presented on NOAA-18 and NOAA-19 satellite imagery. The spatial distribution and confidence levels of detected objects demonstrate the model’s capability to localize convective cloud structures under different atmospheric conditions. The visual results provide qualitative validation of the quantitative metrics and are consistent with previous satellite-based CB detection approaches [10,27].

Fig. 8 presents the results of CB and non-CB cloud detection in the NOAA-18 and 19 images. The details of the detection data results by the YOLOv8 model are outlined in Table 1.

Table 1 summarizes the detection results obtained from three representative satellite scenes. The high proportion of detections with confidence scores above 0.50 supports the robustness of the YOLOv8 model, while the limited number of low-confidence detections highlights remaining challenges in ambiguous cloud regions. Similar confidence-based evaluations have been reported in prior studies on cloud detection to assess model reliability [11,18].

As depicted in Table 1, a total of 53 objects were detected across the three images. Of these, 49 detections had confidence scores above the threshold (TP), while 4 detections fell below it (FP). Substituting these values into the equations above yields an overall detection accuracy of 92.45% and an error rate of 7.55%. These results indicate that the model has a capacity to detect and classify cloud features with a high level of confidence, with only a small portion of predictions falling into the lower-confidence category.

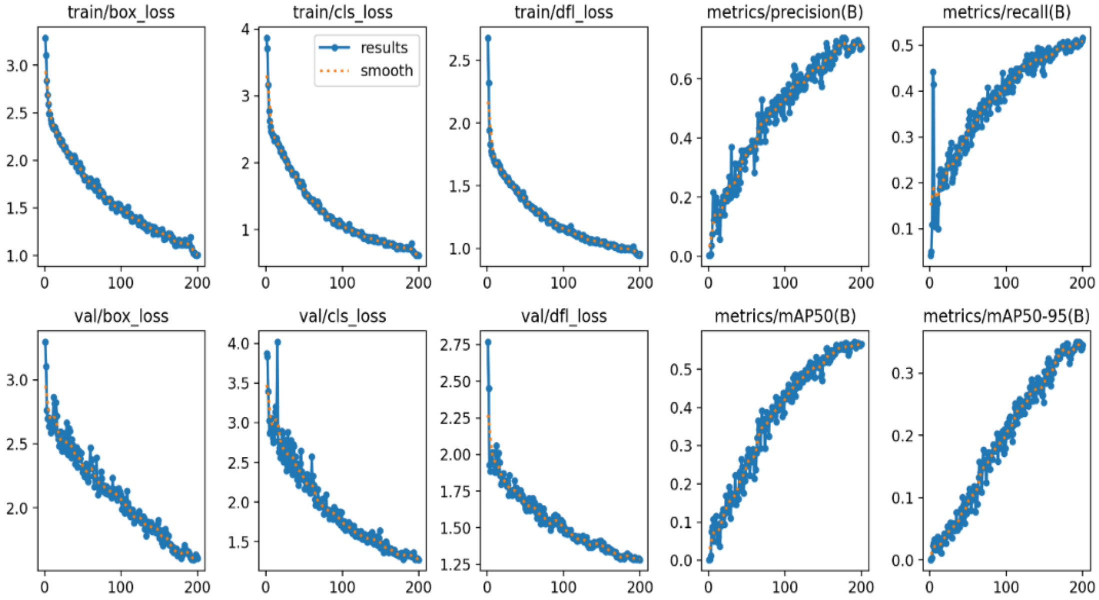


Fig. 7. Training progress of the object detection model over 200 epochs, including loss and performance metrics

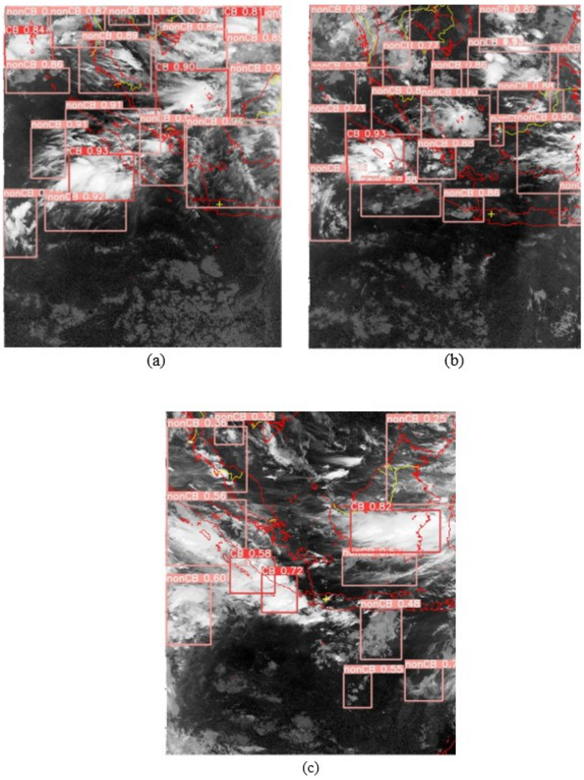


Fig. 8. (a) shows CB and nonCB detection on a NOAA-18 image from 4 January 2022 at 03:15 UTC after contrast enhancement. YOLOv8 detected 20 objects: 4 CB (confidence 0.79–0.93) and 16 nonCB (0.81–0.94), with nonCB dominating the central and right regions. (b) presents detection results from 5 January 2022 at 03:03 UTC. A total of 20 objects were identified: 2 CB (0.90–0.93) and 18 nonCB (0.52–0.93). CBs were located center to lower-left, while nonCBs were more widely distributed. (c) illustrates detection on a NOAA -19 image from 3 February 2022 at 00:16 UTC. YOLOv8 found 13 objects: 3 CB (0.58–0.82) concentrated in the middle to upper-right and 10 nonCB (0.25–0.70) distributed broadly with variable confidence scores

Table 1. Summary of the detection results produced by the YOLOv8 model

Name	Total Detections	Detections ≥ 0.50	Detections < 0.50	Confidence Range
noaa-18-202201040315-contrastb	20	20	0	0.79 - 0.94
noaa-18-202201050303-contrastb	20	20	0	0.52 – 0.93
noaa-19-202202030016-contrastb	12	9	3	0.25 – 0.82

To evaluate the model's performance, accuracy and error rate were calculated using the following equations:

$$Accuracy\ (\%) = (49 / 53) \times 100\ \% = 92.45\ \% \qquad (2)$$

$$Error\ Rate\ (\%) = (4 / 53) \times 100\ \% = 7.55\ \% \qquad (3)$$

It is imperative to note that, in this study, rather than using precision, recall, or mean Average Precision (mAP) as a measure of accuracy as the standard for most object detection methods, this accuracy is actually defined based on the confidence threshold. This approach serves as an interpretation of the detection accuracy at an operational level. It is pivotal to note that, in the confidence level for the detected object being 0.50 or higher, the object is considered correctly detected. In contrast, those with lower confidence levels are considered uncertain outputs of detection. In this case, this definition of accuracy represents the number of detections that meet the specified confidence level.

In contrast, mAP values quantify the performance of detection across multiple confidence thresholds and Intersection over Union (IoU) criteria, thus offering a more comprehensive evaluation of localization and classification accuracy. Therefore, the reported accuracy of 92.45% and the mAP@0.5 value of approximately 0.57 represent different aspects of model performance and should not be directly compared. While mAP evaluates overall detection quality, the confidence-based accuracy provides additional insight into the reliability of high-confidence detections for operational applications such as early warning systems.

It is important to note that confidence thresholds are subjected to variation according to performance assessment tasks. Specifically, the confidence threshold value 0.355 is selected according to F1 score maximization, with the objective of identifying an optimal balance point concerning Precision-Recall score values calculated over the evaluation set of images. This confidence threshold is exclusively employed within the context of an evaluation concerning the operating point at which the balance of classifications is most consistently achieved by the detector.

In contrast, a higher confidence threshold of 0.50 is adopted for accuracy analysis to reflect a more conservative operational scenario. In practical applications such as early warning or aviation-related monitoring, high-confidence detections are favored to reduce false alarms. Hence, the accuracy reported at a confidence threshold of 0.50 represents the reliability of detections under stricter confidence requirements, rather than the optimal classification balance. These two thresholds serve complementary functions and should not be interpreted interchangeably.

Recent studies have employed the use of various YOLO models such as YOLOv5 and YOLOv7 for cloud and atmospheric objects. However, these studies in general have shown an incremental improvement over each other regarding the detection capability and efficiency of the method. The YOLOv8 model was selected for the task due to its anchor-free mechanism, its feature aggregation technique, and the steadiness of its training procedure, rendering it more appropriate for cloud objects that tend to have irregular features. Instead of empirically asserting its superiority over other models available, the feasibility of YOLOv8 will be assessed for its functionality from a real-time satellite observation perspective.

It is importantly noted that no direct quantitative comparison is made with other state-of-the-art detectors under identical experimental settings, which require multiple model retraining with the same dataset and hyperparameter and is considered an important direction of prospective work. However, the performance achieved in this paper is consistent and comparable to those reported in previous cloud detection studies using YOLO architecture, thereby validating the proposed approach.

However, a number of potential limitations should be acknowledged. Firstly, the model is based solely on RGB satellite imagery and does not explicitly model spectral or temporal information, which may limit its ability to discriminate visually similar cloud types. Secondly, the efficacy of detection will be determined by the quality of manual annotations and the spatial resolution of the satellite data. Furthermore, the confidence-based thresholding technique introduces a trade-off between false positives and false negatives, and requires threshold values to be carefully selected according to the operational requirements. These results emphasize the necessity to understand model behavior beyond aggregate performance metrics, with a particular focus on safety-critical applications.

3.1. Comparison with previous YOLO-based studies and implications

Different from previous detection studies grounded in the YOLO, which are primarily concerned with generic object or land scene detection, this study aims to exploit polar-orbiting satellite imagery in the DO of Cumulonimbus clouds over tropical areas. Although the results achieved are similar to performance statistics as reported by previous YOLO-based cloud detection studies, this work adds value to existing approaches by verifying meteorologically and using METAR observations. Pragmatically, the findings demonstrate the feasibility of employing YOLOv8 for near real-time CB identification, thereby supporting aviation safety and early warning systems. Scientifically, the observed detection behavior and failure cases provide insights into the strengths and limitations of anchor-free object detectors when applied to diffuse and non-rigid atmospheric targets.

Some recent studies have presented quantitative results for CB detection with infrared thresholds, multispectral satellite data and radar-supported techniques. Such algorithms generally can attain a high level of accuracy in the detection process as they exploit information regarding the physical cloud-top temperature, vertical structure or reflectivity. However the reported performance is based on different datasets, sensors, spatial resolutions of images and evaluation protocol, which prevents us making direct quantitative comparison with YOLO methods for vision modality. In this study, the detection performance was achieved within the reported range of previous CB detector studies, using solely single-source RGB satellite imagery. This highlights the feasibility of using a lightweight, data-driven detector for near-real-time CB identification, particularly in regions where radar coverage or multispectral data availability is constrained.

Although the utilization of METAR observations provides valuable independent validation from an operational meteorological perspective, it is important to acknowledge the inherent spatial and temporal mismatches between satellite-based detections and point-based surface reports. Satellite imagery represents cloud structures over a spatially continuous area, whereas METAR observations are recorded at fixed ground stations and describe atmospheric conditions at specific locations. As a result, a detected Cumulonimbus cloud within a satellite image may not be perfectly aligned with the exact position of a METAR station, particularly for rapidly evolving convective systems.

Furthermore, temporal discrepancies may arise due to differences between satellite overpass times and the reporting intervals of METAR observations. Despite efforts to match observations within a close temporal window, short-term convective development or dissipation may still lead to partial inconsistencies. These limitations are inherent to satellite–surface data integration and should be considered when interpreting the results of validation process. Nevertheless, METAR-based validation remains valuable in providing operational relevance and complementary confirmation of convective activity.

4. Conclusion

The outcome of the evaluation suggests that the YOLOv8 model provides encouraging results for the detection of Cumulonimbus (CB) cloud in individual NOAA satellite imagery, as evident by balanced quantitative metrics and observations. The model has been shown to be capable and efficient in differentiating CB clouds from non-CB clouds, reaching an highest F1 score of 0.88 for an optimal confidence threshold value of 0.355, while an even higher conservative value threshold at 0.50 has been demonstrated to work well for operational accuracy. The results for object detection performance, regardless of CB cloud size and shape, confirm the resilience of the system and demonstrate YOLOv8's superior ability in identifying small-scale and convective CB clouds. This signifies YOLOv8 strength in encoding small-scale convective clouds. Despite attaining mAP scores of 0.568, which are comparatively weaker than standards for general terrestrial object detection tasks, the model can still compete with satellite image detection work, considering the natural complexity involved in atmospheric phenomena for accurate cloud detection. Validation on additional NOAA-18 images provided an overall accuracy level of 92.45%, signifying consistent performance for accurate detection. However, remaining confusion between CB clouds and background regions confirms limitations in utilizing RGB images alone. This suggests the necessity for additional satellite imaging tools or techniques involving multispectral, temporal, or adaptive threshold strategies for further enhancement in resilience for operational aviation weather surveillance and early warning systems.

Acknowledgements

This research is supported by the Research and Community Service Center of Jakarta State Polytechnic as stated in the Decree of the Director of Jakarta State Polytechnic Number: 3287/PL3/KP.08.01/2024, dated December 30, 2024, under the Research Scheme for Associate Professors (PLK) of Jakarta State Polytechnic for the 2025 fiscal year, as stated in the Decree of the Director of Jakarta State Polytechnic Number: 725/PL3/PT.00.06/2025, dated April 15, 2025.

References

1. C. Henken, M. Schmeits, H. Deneke, and R. Roebeling, *Using MSG-SEVIRI cloud physical properties and weather radar observations for the detection of CB/TCU clouds*, J. Appl. Meteorol. Climatol. 50 (2011) 349–365.
2. S. Mahajan and B. Fataniya, *Cloud detection methodologies: Variants and development A review*, Complex & Intelligent Systems. 6 (2020) 1–11.
3. K. M. Bedka et al., *Objective satellite-based detection of overshooting tops using infrared window channel brightness temperature gradients*, J.

Appl. Meteorol. Climatol. 49 (2010) 181–202, .

4. M. Kim, J. Lee, and J. Im, *Deep learning-based monitoring of overshooting cloud tops from geostationary satellite data*, GIScience & Remote Sensing. 55 (2018) 763–792.
5. R. Dorfman et al, *Spatio-temporal detection of cumulonimbus clouds in infrared satellite images*, in Proc. IEEE Int. Conf. Science of Electrical Engineering in Israel (2018) 1–5.
6. P. Jiang et al, *A review of YOLO algorithm developments*, Procedia Computer Science. 199 (2022) 1066–1073.
7. M.-T, Pham et al, *YOLO-fine: One-stage detector of small objects under various backgrounds in remote sensing images*, Remote Sensing. vol. 12. no. 15 (2020) 2501.
8. G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLOv8*, GitHub (2023).
9. K. Wada, *LabelMe: Image polygonal annotation with Python*, GitHub (2018).
10. Y. Zhang, S. Wistar, J. Li, M. A. Steinberg, and J. Z. Wang, *Severe thunderstorm detection by visual learning using satellite images,* IEEE Trans. Geosci. Remote Sens. 55 (2017) 1052.
11. M. Dewan et al, *Detection of thunderstorm and its associated weather from satellite image using texture features*, International Conference on Informatics, Electronics & Vision (2012) 594–598.
12. N. Aswin et al, *Application of machine learning in cloud classification using satellite images*, International Journal of Emerging Trends in Engineering Research, 8 (2020) 2751–2755.
13. T. Zhenwei and Z. Yadong, *Remote sensing image object detection based on improved YOLOv3*, EURASIP Journal on Image and Video Processing. 2020 (2020).
14. R. L. Bankert et al, *Automatic cloud type classification using passive satellite data: a comparison of neural network and nearest neighbour approaches*, Journal of Applied Meteorology and Climatology. 48 (2009) 1965–1981.
15. K. Khoshelham et al, *A review of deep learning applications in satellite remote sensing*, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLII-4/W18 (2020) 305–310.
16. B. Baum, V. Tovinkere, J. Titlow, dan R. Welch, *Automated cloud classification of global AVHRR data using a fuzzy logic approach*, J. Appl. Meteor. 36 (1997) 1519–1540.
17. R. Bankert, *Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network*, J. Appl. Meteor. 33 (1994) 909–918.
18. T. Berendes, J. Mecikalski, W. MacKenzie Jr., K. Bedka, dan U. Nair, *Convective cloud identification and classification in daytime satellite imagery using standard deviation limited adaptive clustering*, J. Geophys. Res. 113 (2008).
19. M. Donovan et al, *The identification and verification of hazardous convective cells over oceans using visible and infrared satellite observations*, 12th Conf. on Aviation Range and Aerospace Meteorology. AMS (2006).
20. C. K. Carbajal Henken, M. J. Schmeits, E. L. A. Wolters, dan R. A. Roebeling, *Detection of Cb and TCu clouds using MSG-SEVIRI cloud physical properties and weather radar observations*, KNMI WR 2009-04 (2009).
21. Li, C. et al, YOLOv6: *A Single-Stage Object Detection Framework*, arXiv:2209.02976 (2022).
22. He, K. et al, *Spatial Pyramid Pooling in Deep Convolutional Networks*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 37 (2015) 1904–1917.
23. Lin, T.Y. et al, *Feature pyramid networks for object detection*, CVPR (2017).
24. Misra, D, *Mish: A self regularized nonmonotonic neural activation function*, arXiv:1908.08681 (2019).
25. Zheng, Z. et al, *Distance-IoU Loss: Faster and better learning for bounding box regression*, Association for the Advancement of Artificial Intelligence (2020).
26. J. Mecikalski dan K. Bedka, *Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery*, Mon. Wea. Rev. 134 (2006) 49–78.
27. D. Rosenfeld, W. Woodley, A. Lerner, G. Kelman, dan D. Lindsey, *Satellite detection of severe convective storms by their retrieved vertical profiles of cloud particle effective radius and thermodynamic phase*, J. Geophys. Res. 113 (2008) D04208.
28. T. Zinner, H. Mannstein, dan A. Tafferner, *Cb-TRAM: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel Meteosat-8 SEVIRI data*, Meteor. Atmos. Phys. 101 (2008) 191–210.
29. J. Schmetz, S. Tjemkes, M. Gube, dan L. Van de Berg, *Monitoring deep convection and convective overshooting with Meteosat*, Adv. Space Res. 19 (1997) 433–441.
30. V. Levizzani dan M. Setvák, *Multispectral, high-resolution satellite observations of plumes on top of convective storms*, J. Atmos. Sci. 53 (1996) 361–369.
31. P. Alvira, E. Tomas-Pejo, M. Ballesteros and M. J. Negro, *Pretreatment technologies for an efficient bioethanol production process based on enzymatic hydrolisis: A review, Bioresour.* Technol. 101 (2010) 4851-4861.
32. B. Kun-Cahyonoa, M. Hidayatul-Ummaha, R. Andarua, N. Andikab, A. Pamungkasc, H. Hapsari-Handayanid, P. Atmodiwirjoe, R. Nathan, *Leveraging machine learning and open accessed remote sensing data for precise rainfall forecasting,* Communications in Science and Technology. 10 (2025) 135–147.
33. Y. Rafsyam, L. G. Oktariza, I. Z. Jonifan, and E. E. Khairas, *Identification of Cumulonimbus Cloud using Sensor Data of NOAA Satellite Captured by Low Cost Flower Cross Dipole,* Annual Southeast Asian International Seminar (2022).