Indian Journal of Radio & Space Physics
Vol 50, March 2021, pp. 12-18

Segmentation of satellite images using machine learning algorithms for cloud classification

Sruthy Sebastian*, Lakshmi Sutha Kumar, & Pugazhenthi Annadurai

Department of Electronics and Communication Engineering, National Institute of Technology Puducherry, Karaikal 609609, India

Received: 12 February 2021; Accepted: 5 March 2021

Clouds play a significant role in determining the state of a changing weather. Clouds offer useful information for forecasting precipitation and provide measurement for showcasing solar irradiance variability. The influence of specific types of clouds on rainfall prediction and solar radiance has been discussed in this paper. Various segmentation algorithms, clustering algorithms and supervised machine learning algorithms such as K Nearest Neighbors and Random forest have been used to segment/classify the clouds using the dataset obtained from INSAT-3DR satellite. Clouds have been classified into high level clouds (Cirrus clouds), medium level clouds (Alto clouds) and low level clouds (Stratus clouds) in accordance with the altitude and cloud densities. The performance metrics has been found for the segmented images. Parameters that provide optimum results for supervised machine learning algorithms have been explored. On the images, different machine learning algorithms have been compared.

Keywords: Fuzzy-C-Means, INSAT-3DR, Random forest, K-Means clustering

1 Introduction

On Earth, clouds are formed when there is an increase in adequate dampness, generally as water vapors from an adjacent source to raise the dew point to the average surrounding air temperature. Clouds take up a portion of the radiation leaving the earth and keep it from escaping further up to the air and into space. Smoke from fire and vaporizers, and little particles in the air can look like clouds in satellite pictures. The analysis of clouds in the sky is fundamental, as it has been an important effect in giving valuable data to the forecast of precipitation and solar irradiance based variability.

There are three essential sorts of clouds in particular; Stratus, Cirrus, and Cumulus. Clouds can be characterized depending on their elevation. High Clouds are the "Cirrus", the Middle Clouds are the "Alto", and the Low Clouds are the "Stratus"[1]. High level clouds are opaque/thick clouds, and they have a pixel intensity near to 255. The precipitation clouds are the Nimbostratus and Cumulonimbus. From an optical distant detecting perspective, cirrus clouds, comprising a large number of thin non-spherical ice crystals are normally translucent, andare situated at high elevation[2].

Pugazhenthi A *et al.*[3] have classified INSAT-3D clouds into non-cloudy and 3 cloudy regions using various segmentation algorithms. The results have revealed that IFCM outperforms other algorithms considered in the literature. Mathew J Reno and Joshua S Stein *et al.*[4] have suggested that the National Oceanic and Atmospheric Administration (NOAA) characterization of cloud type is helpful for describing the irradiance during the time span by contrasting the hourly cloud pictures with ground estimated irradiance.

Seema Mahajan and Bhavin Fataniya *et al.*[5] have investigated different types of cloud identification like Cloud/No cloud, Snow/Cloud, and Thin Cloud/Thick Cloud utilizing different methodologies of artificial intelligence and traditional calculations. Cloud detection can be performed using cloud optical properties, such as spectral content, near infrared (NIR), visible–infrared (VIR), thermal infrared (TIR), brightness temperature etc[5]. Franke M *et al.*[6] have recommended utilizing a remotely created cloud mask as input, for the calculation of weather situations. Jo Ann Parikh *et al.*[7] have proposed a cloud-type order framework which settle ambiguities in infrared cloud-type marks by an examination of textural measures on known and obscure cloud-type fragments. Zhuli Xie *et al.*[8] have proposed a technique to distinguish appropriate factors and calculations for arranging land cover, forest, and tree species. Six order calculations including Maximum Likelihood Classifier (MLC), K-Nearest Neighbor (KNN), Decision Tree (DT),

[*] Corresponding author (E-mail: sruthyseb503@gmail.com)

Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM) were utilized. RF and SVM gave the best characterization accuracy of about 84%. KNN yielded an accuracy of 76%.

Minakshi Gogoi *et al.*[9] have reviewed about the clouds for prediction and forecasting of rainfall. C Thirumalai *et al.*[10] have proposed a formal prediction of rainfall using machine learning techniques including linear regression method in metrics for effective observation of agriculture in India. Nafish Gasemian *et al.*[11] have presented two RF based calculations, Feature Level Fusion Random Forest (FLFRF) and Decision Level Fusion Random Forest (DLFRF) to incorporate, infrared (IR) and spectral and textural features (FLFRF) for exceptionally exact cloud recognition on distant detecting images. Color histogram based on cloud pixel density is computed which shows the proportion of clouds of each pixel in the image[12]. Pixel 0 represents clear sky/no cloud, whereas pixel 255 represents opaque/dense clouds.

In this paper, segmentation techniques, clustering techniques, and supervised machine learning techniques have been used for cloud segmentation. Thermal Infrared image obtained from Indian National Satellite-3D Repeat (INSAT-3DR) is taken for segmentation. The comparisons of segmentation/ clustering/ supervised machine learning algorithms on the image are provided.

2 Materials and Methods

2.1 Satellite image

Cloud pictures gathered can be either digital or satellite pictures. Digital pictures can be gathered from ground level through computerized camera or through web. While, satellite pictures are gathered either from meteorology division or gathered on regular routine from some meteorology sites. Forecast of precipitation through satellite picture can be accurate, since the satellite picture has clear cloud structure as compared to digital pictures. The Thermal Infrared (TIR) Indian National Satellite-3D Repeat (INSAT-3DR) full disk image is applied with various segmentation algorithms.

The INSAT-3DR satellite was launched on 8 September 2016. It is a major meteorological climate satellite of India arranged with an imaging system and an atmospheric sounder. The brightness temperatures as well as the solar reflectance comprise the 3 channel Red-Blue-Green composites. The satellite monitors the entire disk of the planet at a time resolution better than 30 minutes, that is 48 images per day. The images have the likelihood of clouds in a zone, where the white zones show a high likelihood and the dark zones show a low likelihood of clouds.

2.2 Segmentation techniques

Cloud image segmentation has been done using the region and edge-based detection. Edge detection operators are based on the idea that edge information in an image can be found by looking at the relationship a pixel has with its neighbors. Entropy-based approaches are very stable and efficient in noisy environments. Gaussian edge detection covers wider area around the pixel and also finds correct places of edges but its computational complexity is high. Gradient based detection involves Sobel, Prewitt, and Robert. Sobel operator finds approximate absolute gradient magnitude at each point in an input grayscale image. Thresholding technique is used for region based segmentation.

2.3 Clustering techniques

Fuzzy C Means (FCM) is a clustering technique in which one piece of data belongs to 2 or more clusters. In this paper, 4 clusters have been selected. Every point in FCM has a degree of belonging to other clusters, rather than just completely belonging to just a single cluster. Points at the edge of a cluster maybe in a lesser degree than those points in the center of the cluster. In K-Means (KM) clustering, n observations are partitioned to k clusters, in which each observation belongs to the cluster with the nearest mean. Minimization of the average squared euclidean distance of values (objective function J) from their cluster center is the primary objective of KM. Centre is the centroid or mean of the values in a cluster. The objective function J[13] is given by,

$$J = \sum_{j=1}^{n} \sum_{i=1}^{n} \| x_i^{\,j} - \mu_j \|^2 \qquad \ldots (1)$$

where, *k, n* are the number of clusters and cases respectively, x_i is each data point and μ_j is the centroid for cluster *j*.

2.4 Machine learning algorithms

Machine learning algorithms provide more accurate and efficient segmentation of natural images. Machine learning algorithms can be classified into two, namely, unsupervised and supervised algorithms.

Unsupervised machine learning algorithms rely on thresholds and it requirewell-defined boundaries, whereas supervised machine learning algorithms can capture the shape, inhomogeneity, cloud intensities as well as neighborhood correlations[14]. In this work, supervised machine learning algorithms such as K Nearest Neighbors and Random Forest have been used for prediction. The block diagram for segmentation of the data for a single image using machine learning is shown in Fig. 1.

K-Nearest Neighbor algorithm is a supervised machine learning algorithm that can be used in classification problems. The final classification is obtained by measuring the distances between the test data points and each of the training data points. KNN calculates the distance between two data instances and it locates k most similar data instances and a response is generated from the instances. Euclidean distance[16] between two points or tuples, say, $X_1 = (x_{11}, x_{12}\ldots x_{1n})$ and $X_2 = (x_{11}, x_{12}\ldots x_{2n})$, is:

$$Dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad \ldots(2)$$

Random forest is an ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at training time. Bagging or bootstrap aggregation is performed in order to overcome the challenge of variance-bias trade off. A high variance leads to over fitting whereas a high bias leads to inaccurate results. In order to reduce the correlation of trees in a bagging sample, during the process of tree split, the random forest algorithms chooses a random subset of features. For k branches in a dataset T, with the probability of each class P_i, Gini index[17] is found at each split. It is given by:

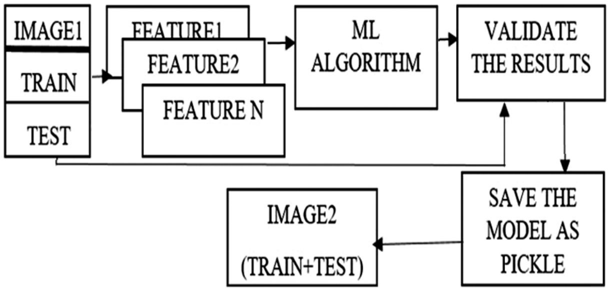

Fig. 1 — Block Diagram for cloud segmentation/classification by machine learning.

$$GiniIndex\ (T) = 1 - \sum_{i=0}^{k} p_i^{\,2} \quad \ldots(3)$$

Various parameters such as mean square error (MSE) and the peak signal to noise ratio (PSNR) are found for the segmented image. The MSE and the PSNR[18] of the predicted/segmented image are found using the equation:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Actual - predicted)^2 \quad \ldots(4)$$

Ideally, the mean square error can range from 0 to infinity. Lower MSE indicates better score.

$$PSNR(dB) = 10\log_{10}\left(\frac{255^2}{MSE}\right) \quad \ldots(5)$$

In ideal case, where the actual as well as the predicted images are the same, the MSE equates to zero and hence, PSNR become infinite.

3 Results and Discussion

Figure 2(a) shows the image chosen for analysis which is the INSAT-3DR full disk TIR image observed on 6/10/2020 at 14:45 pm centered at 20.59° latitude 78.96° longitude. The white zones indicate a high likelihood and the dark zone indicates a low likelihood of clouds. The TIR image selected has a resolution of 4Km with spectral range of 10.3-11.3 µm. The image pixel histogram of the grey scale image is shown in Fig. 2(b). It depicts the various grey level probabilities present in the image. The horizontal axis represents the pixel intensities from 0 to 255 whereas the vertical axis represents the total no. of pixels that belongs to the particular intensity. Left most end (lower intensity region) indicates dark region, i.e. non-cloudy region. Moving towards right, the histogram represents light regions indicating the presence of clouds. The histogram peaks in the dark region at around 30-40 pixels indicating that most of the part of the image belongs to the non-cloudy region.

The thresholding techniques, edge detection techniques, clustering algorithms, and supervised machine learning algorithms are applied to the image under test for region based segmentation. The results and related discussions are provided next.

a Thresholding technique

Thresholding technique is applied on the image as shown in Fig. 2(c). A local threshold is applied in order to segment the image using this fast detection thresholding algorithm on the grey scale image. The image is segmented into two regions based on pixel value observed from histogram of the grey scale image. The threshold is set at 125. Pixels belonging to the range 0 to125 and 125 to 255 are classified as non-cloudy and cloudy region respectively.

b Edge detection technique

An edge is detected when there is a boundary that divides an area of an image into two regions or when a local discontinuity in intensity is detected. The results of a few edge detection techniques namely, Sobel and Entropy, are shown in Fig. 2(d - f).

The sobel filter serves the purpose of edge detection. Edges are detected when there is a large increase from light to dark region, i.e., cloudy and non-cloudy region. This is performed by calculating the gradient of the image intensity. Entropy image is calculated from the

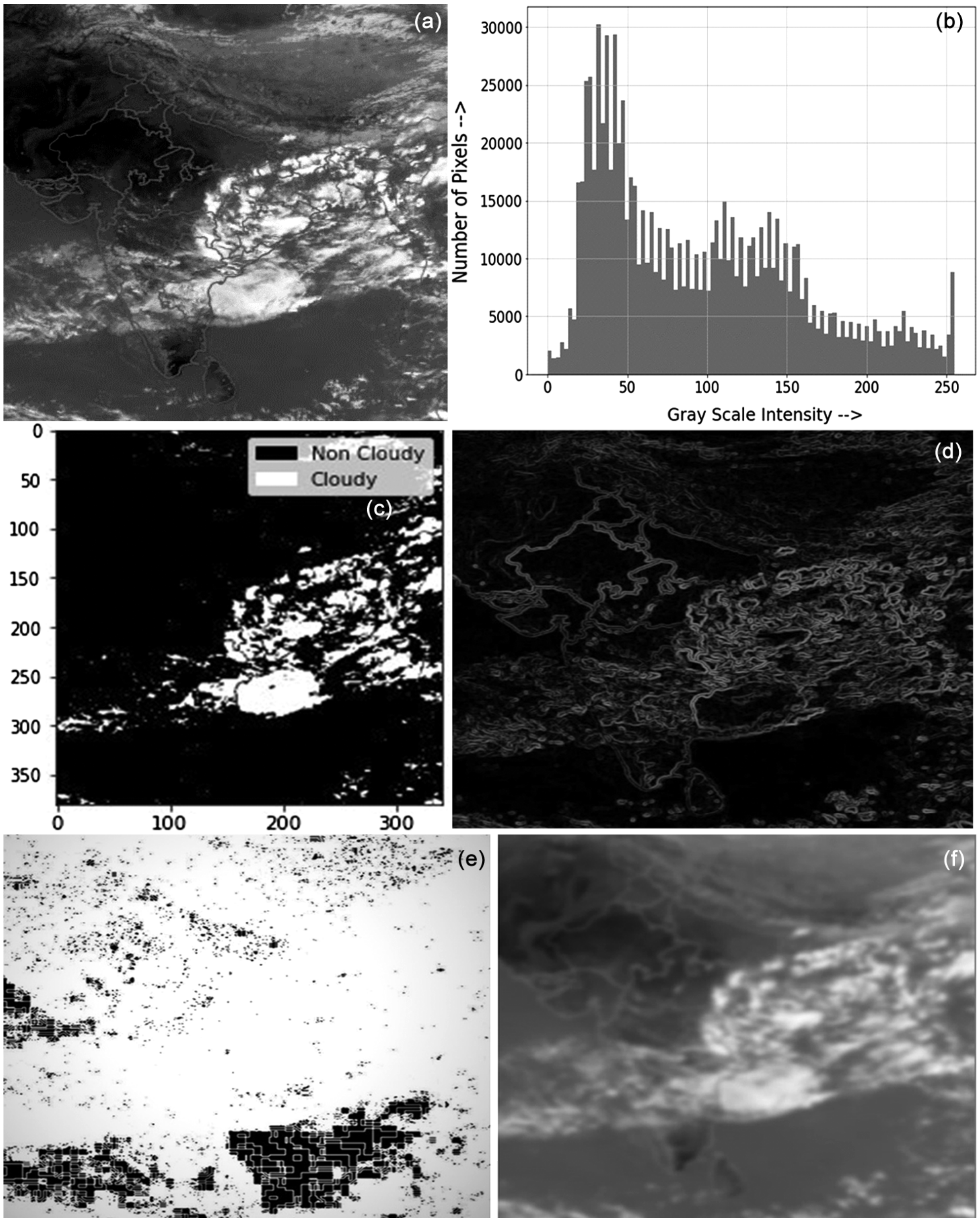

Fig. 2 — (a) Image under test, (b) image histogram, (c) segmented image using thresholding technique, (d) sobel image, (e) entropy image, and (f) gaussian image.

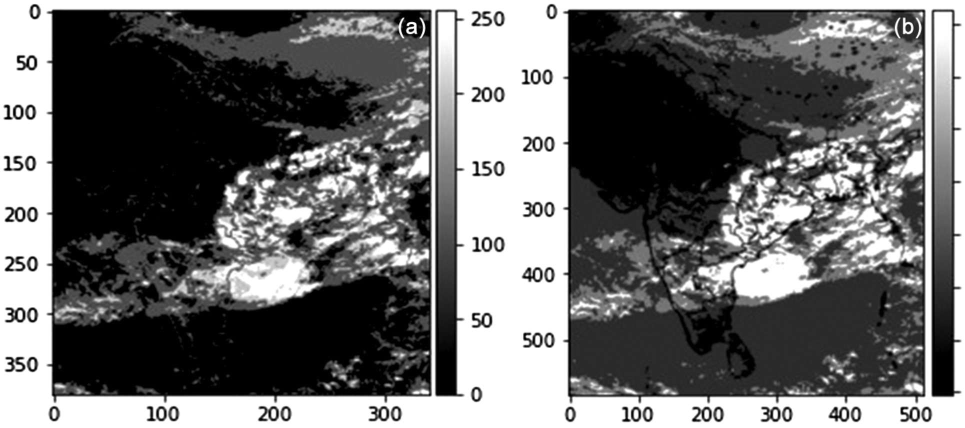

Fig. 3 — (a) Segmented image by FCM, and (b) segmented image by KM

image histogram. The entropy is also known as the average information of an image. The gaussian image has a smoothened edge and it achieved a better PSNR compared to sobel, entropy image and gaussian image.

c Clustering techniques

Fuzzy-C-Means and K-Means clustering algorithms are used to segment the image into 4 regions on the basis of high intra-cluster property and low inter-cluster property.

The segmented images are shown in Fig. 3(a - b) respectively. The 4 types of clusters are non-cloudy (black region), thin clouds (dark grey region), medium clouds (light grey region), and dense clouds (white region). FCM and KM have successfully segmented the image. The PSNR of the segmented images by clustering algorithms is higher than those performed by classical segmentation approaches indicating a better segmentation outcome of KM algorithm than FCM algorithm.

d Supervised machine learning algorithms

The steps involved in applying machine learning algorithms on the image are listed below.

1 3D image is converted to 2D.

2 Onto this 2D image features namely: Gabor filters step size [0.05-05], Canny edge, Roberts, Sobel, Prewitt, Gaussian [sigma=3, 7], Median Filters [size=3] are added.

3 Feature ranking is performed in order to identify the features that contribute to the performance of the algorithm.

4 Label is generated for the 2D image based on pixel values. The data frame generated is of dimension (298424 × 43). 298424 indicates the total number of data points.

5 Prediction is done using KNN and random forest on a single image onto the 298424 data points.

6 70% and 30% data points are converted into training into testing data respectively.

7 In order to choose the best parameters for obtaining optimum results, accuracy vs. the number of neighbors (K) and accuracy vs. the number of estimators (N) are plotted in Fig. 4(a and b) for KNN and Random Forest respectively. This is performed for both test and train datasets.

8 Feature importance is performed and the results are validated.

A smaller value of K results in noise and a large value makes it computationally expensive. Considering these and from the results obtained, K can be chosen as 2 or 3. Since for even number of classes, k as an even value may result in wrong voting of the class therefore, K=3 is selected as the best choice. As the number of estimators increase in random forest, the computational complexity increases. Considering this challenge from the results obtained in Fig. 4(a), thirteen estimators will be the ideal choice for obtaining the best results with low computational complexity for RF. Table 1 shows the accuracy of KNN with K=3 and random forest using 13 neighbors for testing as well as training data.

Random forest has yielded the best accuracy for both; the test and train data. The feature importance is found. Median s3 and Gabor 32 are the features with the topmost importance. They play a crucial role in determining the accuracy of prediction. Median filters preserve edges in the image. Gabor filters give the best response at edges and at locations of texture change.

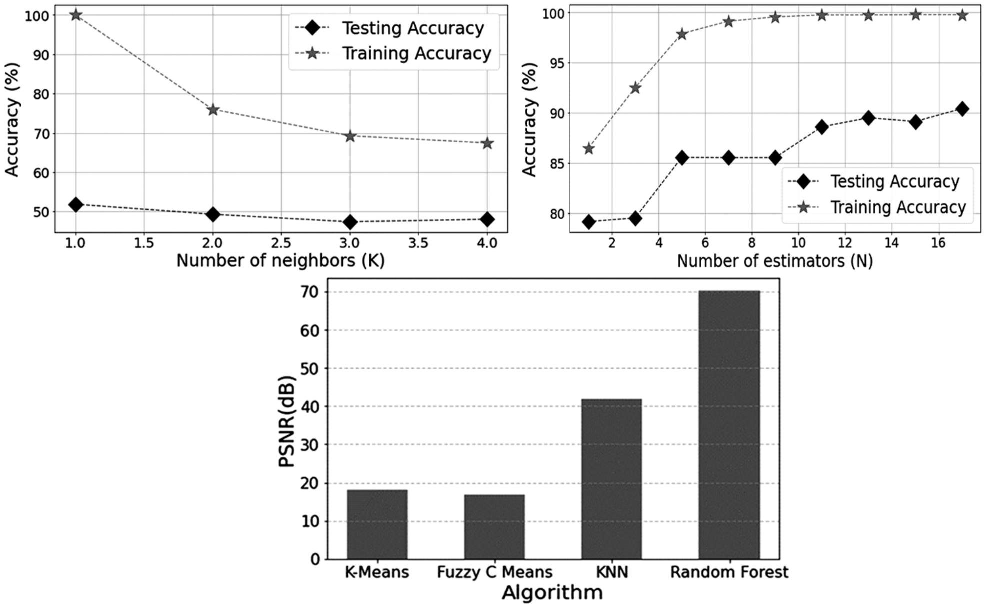

Fig. 4 — (a) KNN[13]- K vs Accuracy, (b) Random forest[14]- N vs Accuracy, and (c) PSNR (dB) vs segmentation algorithm.

Table 1 — Accuracy for machine learning algorithms[16,17]

ML Algorithm	Train Accuracy (%)	Test Accuracy (%)
KNN[16]	70.10	48.56
Random Forest[17]	99.80	90.00

Table 2 — PSNR and MSE for various segmentation algorithms[13,16-18,21-24]

Algorithm	Technique employed	PSNR(dB)	MSE
KM[13]	Clustering	17.89	1056.69
KNN[16]	Machine Learning	41.56	4.53
Random Forest[17]	Machine Learning	70.13	0.00631
Thresholding[21]	Region based segmentation	8.59	8983.43
Sobel[22]	Edge detection	6.58	14279.00
Entropy[22]	Edge detection	6.71	13870.84
Gaussian[23]	Edge detection	24.03	257.00
FCM[24]	Clustering	16.54	1442.00

Based on the features generated, the prediction is performed on the INSAT-3DR image. Accuracy of KNN for 3 estimators has yielded 70.10% whereas random forest yielded an accuracy of 99.2% while using gini index with 13 estimators. The quantitative parameters are calculated using (4), (5) for the algorithms discussed and results are tabulated in Table 2.

It is clear from Table 2 that classical methods such as edge based and region-based segmentation have given poor performance. Clustering techniques have yielded better performance compared to classical segmentation algorithms. However, machine learning has generated the best results. Random forest (shown in bold in Table 2) has performed well followed by KNN (shown in bold and italic in Table 2) as compared to other methods considered based on the calculated PSNR, MSE values. Among clustering techniques, KM has yielded the best results. MSE is bound to $(0, \infty)$. The MSE obtained by performing supervised machine learning has been very low and that indicates better prediction. A comparison of 4 segmentation algorithms from clustering based and machine learning based technique is given in Fig. 4(c).

4 Conclusion

An image taken from INSAT-3DR is applied with thresholding techniques, edge detection techniques, clustering algorithms, and supervised machine learning algorithms. The quantitative parameters are calculated to compare the segmentation results. It is found that supervised machine learning algorithm has yielded the best segmentation results followed by the clustering algorithms. In the machine learning algorithms considered, Random forest has yielded the best PSNR and MSE. The accuracy of prediction of random forest is also higher than KNN for both test and train data. Better PSNR and accuracy helps to achieve better segmentation outcomes and hence better prediction of weather situations. The future focus of the paper is to analyze the cloud segmentation results to predict the rainfall and solar variability with the aid of rain and solar datasets.

References

1 Zhao M, Chang C H, Xie W, Xie Z, & Hu J, *IEEE Access, 8 (2020) 44112.*

2 Lima C B, Prijith S S, Rao P V N, Sai M V R S, & Ramana M V, *IEEE Trans Geosci Remote Sens*, 10 (2020) 1.

3 Pugazhenthi A, & Kumar L S, *IET Image Processing*, 14 (2019) 5.

4 Matthew Reno J, & Joshua Stein S, Sandia National Laboratories, *US Department of Energy's National Nuclear Security Administration,* 1 (2013) 2.

5 Mahajan Seema, & Bhavin Fataniya, *Complex Intell Syst*, 6 (2019) 251.

6 Göttsche, Frank M, & Olesen, & Folke, *Multi-scale segmentation of satellite data into image objects and knowledge-based detection and classification of clouds, EUMETSAT Meteorological Satellite Conference on trends in Electronics and Informatics*, 5 (2017) 1114.

7 Ann Jo Parikh, & Azriel Rosenfeld, *IEEE Trans Syst Man Cybern Syst*, 8(10) (1978) 736.

8 Zhuli Xie, Yaoliang Chen, Dengsheng Lu, Guiying Li, & Erxue Chen, *Remote Sensing*, 11 (2019) 1.

9 Gogoi Minakshi, & Devi Gitanjali, *J Adv Res Elect Elect Eng*, 2 (2015) 13.

10 Thirumalai, Chandra Segar, Harsha K, M L Deepak, & K C Krishna, *Heuristic prediction of rainfall using machine learning techniques, International Conference on Trends in Electronics and Informatics (ICEI)*, 1 (2017) 1114.

11 Nafise Ghasemian, & Mehdi Akhoondzadeh, *Adv Space Res*, 62(2) (2018) 288.

12 Ghosh, Pal R N, & Das J, *Int Geosci Remote Sens Symp*, 6 (2003) 3438.

13 Li, Youguo, Wu, & Haiyan, *A Clustering Method Based on K-Means Algorithm, International Conference on Solid State Devices and Materials Science*, 25 (2012) 1698.

14 Seo, Hyunseok, Badiei Khuzani, Masoud,Vasudevan Varun, Huang Charles, Ren Hongyi, Xiao Ruoxiu, Jia Xiao, & Xing Lei, *The International Journal of Medical Physics Research and Practice,* 47 (2020) 148.

15 Ali Ameri, Ali Akhaee Mohammad, Scheme Erik, & Englehart Kevin, *PLOS ONE*, 13 (2018) 2.

16 Li Yu Hu, Min Wei Huang, Shih Wen Ke, & Chih Fong Tsai, *SpringerPlus*, 5 (2016) 2.

17 Sivagama S, & Sundhari, *A knowledge discovery using decision tree by Gini coefficient*, *International Conference on Business, Engineering and Industrial Applications*, 2 (2011) 233.

18 Joshi K, Yadav R, & Allwadhi S, *PSNR and MSE based investigation of LSB*, *International Conference on Computational Techniques in Information and Communication Technologies*, 3 (2016) 218.

19 www.imd.rapid.gov.in

20 Tu L, & Dong C, *Histogram equalization and image feature matching, International Congress on Image and Signal Processing*, 6 (2013) 446.

21 Siva kumar V, & Murugesh V, *A brief study of image segmentation using Thresholding Technique on a Noisy Image, International Conference on Information Communication and Embedded Systems, 1* (2014) 2.

22 Zhang H, Zhu Q, & Guan X, *Probe into image segmentation based on sobel operator and maximum entropy algorithm, International Conference on Computer Science and Service System*, 5 (2012) 238.

23 Wang M, & Zheng S, *A new image denoising method based on Gaussian filter, International Conference on Information Science, Electronics and Electrical Engineering*, 1 (2014) 163.

24 Duan L, Yu F, & Zhan L, *An improved fuzzy C-means clustering algorithm International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1 (2016) 1199.