Classification tree analysis (CTA) of smoke detection using Himawari_8 satellite data over Sumatera–Borneo Island, Indonesia

Heri Ismanto^{1,2} · Hartono Hartono² · Muh Aris Marfai²

Received: 22 January 2020 / Accepted: 6 August 2020 / Published online: 16 August 2020
© Springer Nature Switzerland AG 2020

Abstract

This study proposed Classification Tree Analysis (CTA) for automatic smoke detection using Himawari_8 Satellite data over the Maritime Continent of Sumatera and Borneo Islands in Indonesia. Two timestamps of the Region of Interest (ROI) sampling, including Cumulonimbus (Cb) top, low-middle cloud, smoke, bare soil, cirrus cloud, vegetation, and water classes, were used as the input to determine the best CTA models. The CTA model classification was supervised using a collection of 21 single and transformation bands. The study also employed and compared two impurity measures, the Gini Index, and Entropy. The responses of the output of 4 CTA models (Entropy-09, Gini-09, Entropy-10, and Gini-10) were spatially, temporally, and statistically analysed. Furthermore, the CTA models were validated using METAR data (weather airport observation), with results showing that Entropy-10 have the highest Overall Accuracy value of 0.79, and lowest False Alarm Rate Value of 0.11. The computing time shows that Entropy-9 is the fastest with a mean of 19.8 s, followed by entropy-10 with 20.7 s. The accuracy assessment, spatial and temporal analyses, and computing process revealed that the Entropy-10 was the best model. The results of the CTA Entropy-10 are implemented over a small area, such as an airport to justify the work of weather observers and forecasters. This is often based on the objective satellite-based smoke detection product. Furthermore, they serve as information for aviation users in improving their situational awareness of adverse weather conditions related to safety.

Keywords Smoke · Classification tree analysis · Gini Index · Entropy · Himawari-8

1 Introduction

Smoke, a weather hazard specific to Indonesia, has attracted international attention after spreading to neighbouring countries in South-eastern Asia [1–5], negatively affecting many fields, such as health, economy, tourism, and transportation [6]. The worst transportation disaster ever due to smoky conditions was the flight Garuda Indonesia (GA152) on September 26, 1997, that crashed into a mountainous area near Medan International Airport, killing all 234 passengers and crew board [7].

Dry pollutants, smoke, and haze restrict visibility and should be considered during safety planning. Haze is an atmospheric condition where the sky is obscured by dust, smoke, and other dry particles (including fires). Regional haze results in visibility impairment caused by the cumulative air pollutant emissions from numerous sources over a wide geographic area. Since it is composed of microscopic particles, the haze is not individually visible but still effective in visual range restriction. Smokes particles are relatively larger than haze due to their near distance to the sources [8].

✉ Muh Aris Marfai, arismarfai@ugm.ac.id; Heri Ismanto, heri.ismanto@bmkg.go.id; Hartono Hartono, hartono.geografi@ugm.ac.id |

¹Center for Aeronautical Meteorology, Agency of Meteorological, Climatological and Geophysical (BMKG), Kemayoran, Jakarta 10720, Indonesia. ²Faculty of Geography, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.

The methods employed in satellite-based smoke detection are divided into a visual approach using RGB composite [9–11] and automatic smoke detections. The visual method depends on subjective interpretation, while the automatic smoke detection model is based on some training samples. The smoke detection model has been applied using multi-threshold [12, 13], combinations of multi-threshold and K-means cluster [12], multi-spectral transformation [14], and neural network [15, 16]. Another statistical method that has been widely used for over 50 years [17] is the Classification Tree Analysis (CTA), applied in various subjects, such as landslide susceptibility [18] and visibility forecasting due to fog [19]. It is a machine learning algorithm used to choose the best predictor through various impurity measures (Gini, twoing, ordered twoing, and least-squared deviation) [20, 21]. CTA is a potential alternative to the automatic smoke detection method in the Indonesian tropical region.

Himawari_8 satellite data have higher temporal, spatial, and spectral resolutions than the previous generations of Japanese weather satellites. Every single image has data variation of individual objects. Spectrally, this satellite's high instrument, the Advanced Himawari Imager (AHI), is equipped with 4 times more channels than Multi-Functional Transport Satellite 2 (MTSAT2), allowing more accurate object detection. To optimize smoke detection in Indonesia using Himawari_8 data, this research proposed combining single-channel and multi-channel transformation as the input in decision tree analysis. This approach is a modification of the procedures employed in previous studies to obtain a multi-threshold smoke detection using AVHRR, Advanced Very High-Resolution Radiometer and TOMS, Total Ozone Mapping Spectrometer [16, 22]. Based on the type of machine learning, statistically, CTA learns the best threshold based on sample input. Using the combination of the single and multi-channel transformation of Himawari_8 as the sample input in the multi-threshold algorithm for smoke detection is a novelty and an opportunity as an alternative method to obtain better smoke detection results.

This research created an automatic smoke detection over Borneo and Sumatera Islands, the Tropical Maritime Continent in Indonesia, using CTA based on Himawari_8 satellite data. The two different ROI inputs (sample in September and October 2015) in building several CTA models were compared to spatially and statistically acquire the best model and implement it in automatic smoke detection. The research offers many advantages, including support to the other studies related to climate change, weather forecasting, risk and disaster management, remote sensing, health, air pollution, and transportation. Besides, the research provides objective information on

adverse weather conditions for aviation users in promoting safety in their operation.

This paper is structured into research data and methods in Sect. 2, results and discussion, containing CTA model analysis, CTA classification result, and validation, and CTA model performance testing over a small airport area in Sect. 3 and the conclusion in Sect. 4.

2 Research data

The research data included Himawari_8 satellite data and Meteorological Terminal Aviation Routine Weather Report (METAR) from weather observation in airports (white star) in the study area (Fig. 1). The Himawari_8 satellite data was used for the CTA model development, while the METAR was used in validating and assessing the smoke-CTA model's accuracy.

This study used the Himawari_8 (later called H_8) satellite data for the algorithm development. H_8 satellite is positioned at 140.7 E longitude about 35,800 km above the equator. The H_8 data provided by the Meteorology, Climatology, and Geophysical Agency of Indonesia (BMKG) are in SATAID format with a 2 km × 2 km spatial resolution. BMKG receives this data from the Japan Meteorological Agency (JMA) via the Himawari Cloud services. The level 1b data of H_8 has been radiometrically and geometrically corrected [23]. This study used the H_8 data in the 2 km spatial and 10-min temporal resolutions. Furthermore, the H_8 satellite has 3 visible, 3 near-infrared, and 10 infrared channels. The 3 visible channels' central wavelengths are 0.46, 0.51, and 0.64 μm , while that of the 3 near-infrared channels is 0.86, 1.61, and 2.26 μm , as presented in Table 1. The central wavelengths of the 10 infrared channels range from 3.85 to 13.3 μm . This research's focus area is Sumatera and Borneo Islands, the Maritime Continent in Indonesia, and parts of Southeast Asia.

The analysis was based on H_8 satellite data and meteorological observations and reports from 47 airports. The airport locations and observation data were retrieved from <http://aviation.bmkg.go.id/>. METAR contains information from actual weather observation recorded by the aeronautical meteorological stations at an aerodrome. It is used for en-route navigation and flight performance and serves as a basis for weather forecasting. METAR is made by a fully automatic, semi-automatic, or manual observation at an hourly or half-hourly interval (depending on the Regional Agreement/RAN), and transmitted outside the aerodrome for reporting. The report issued includes actual weather conditions such as wind, visibility (including prevailing visibility), weather phenomena, cloud condition, dew point, temperature, and air pressure. However, this study only focused on prevailing visibility. Visibility information is

Wavelengths (μm)	Bands/spectral transformations	Inputs	Physical properties
0.47	1	(B1)	Vegetation, aerosol
0.51	2	(B2)	Vegetation, aerosol
0.64	3	(B3)	Low cloud, fog
0.86	4	(B4)	Vegetation, aerosol
1.61	5	(B5)	Cloud phase
2.26	6	(B6)	Particle size
3.9	7	(B7)	Low cloud, fog, forest fire
6.2	8	(B8)	Mid and upper-level moisture
6.9	9	(B9)	Mid-level moisture
7.3	10	(B10)	Mid- and lower level Moisture
8.6	11	(B11)	Cloud phase, SO2
9.6	12	(B12)	Ozone content
10.4	13	(B13)	Cloud top information
11.2	14	(B14)	Cloud imagery, SST
12.4	15	(B15)	Cloud imagery, SST
13.3	16	(B16)	Cloud top height
	13 (Inverse)	(B17)	Cloud top information
	(Li 2015). 1 $(B1 - B4)/(B1 + B4)$	(B18)	Smoke/aerosol
	(Li 2015). 2 $(B1 - B6)/(B1 + B6)$	(B19)	Smoke/aerosol
	(Li 2015). 3 $B3 + B4$	(B20)	Cloud
	NDVI $(B4 - B3)/(B4 + B3)$	(B21)	Vegetation

To accurately identify the shape of smoke, cloud, or other underlying surfaces, the Region of interest (ROI) of the CTA model for smoke detection was generated from the H₈ data sample. The data on September 4, 2015, 01.00

UTC (sample 1), and October 14, 2015, 02.00 UTC (sample 2) were used as the ROI samples to build the CTA-smoke model based on the Gini Index and Entropy. This procedure yielded 4 combinations of the CTA-smoke model, including Entropy-09 (E09), Gini-09 (G09), Entropy-10 (E10), and Gini-10 (G10).

3 Research methods

3.1 RGB and multispectral analysis

This research used the RGB (red, green, blue) composite for aerosols, as proposed by Daniel Rosenfeld & Lensky [24] (visible 0.6 μm (R-red), 0.8 μm (G-green) and IR 10.4 μm (B-blue, inverse)), natural colour RGB (1.6 μm (R), 0.8 μm (G), 0.6 μm (B)) and Day Microphysics RGB (0.86 μm (R, reflectance), 3.9 μm (G, reflectance), 10.4 μm (B, radiance)). The ROI samples were divided into 7 classes, including Cb-cloud top, low and middle cloud, smoke, bare soil, thin cirrus cloud, vegetation, and water. This approach aimed to minimize the amounts of missing samples of smoke and other classes.

All 7 classes in the ROI sample have different characteristics in the form (liquid, gas, solid), microphysics (water vapor, ice, dry aerosol), temperature, and particle size of aerosol. Each class emitted different radiative wavelengths and light coming radiative respond. Himawari-8 captured these differences using multi-spectral windows. RGB composite is the method used to establish each class's characteristics based on 3 combinations of satellite channels. The cb-cloud top appears reddish in RGB day microphysics because they have high reflectance of the near infrared channel and are composed of ice. Low and mid clouds are both composed of water vapor and medium-level temperatures compared to the surface and high clouds, therefore they were taken as one class. Smoke is a tiny aerosol with a diameter of less than 1 μm , sensitive only to the lower wavelength. Using RGB aerosol, the smoke appeared yellowish to soft brown due to low reflectance of sunlight and higher temperatures than all other cloud classes. The vegetation appeared greenish due to the high reflectance of near infrared channels (no absorption at all in the green beam) at RGB aerosol. Bare soil appears brown by using RGB day natural color due to increasing reflectance and wavelength. Therefore, the reflectance is higher in the short-wavelength infrared and lower in near infrared and blue channel. Cirrus cloud is categorized as a high cloud. It has low temperatures and low reflectance by using RGB day natural color, therefore, it appears whitish. The water body is revealed as blackish due to the absorption of all band RGB composite day natural color (short-wavelength Infrared, near-infrared, and blue channel).

The feature space of the ROI is presented in Fig. 2. The samples of top Cumulonimbus (Cb) cloud (red dots), water (blue dots), and cirrus cloud over water (magenta dots) separate well among all classes in the 2 ROI samples. The low-middle cloud (cyan dots) of the ROI is placed between top Cb cloud and smoke, vegetation, and bare soil. However, a small amount of mixing with smoke is present in minor pixel. The spectral value of Band 3 separates smoke from vegetation in the ROI. For this reason, automatic smoke detection based on H₈ satellite data was developed using the CTA model with 2 impurity methods (Entropy and Gini Index). Statistically, the ROI sample dataset does not separate in a normal distribution (based on Saphiro-Wilk method analysis). Therefore, to prove the separation of spectral values of each class, the Kruskal–Wallis method was used to solve this statistic distribution of the data among the classes. The Kruskal–Wallis non-parametric test with a significant value of less than 0.05 indicates a statistically significant difference among the dataset [25]. The Kruskal–Wallis method was used to calculate the separation of spectral bands value for all the classes. One result was obtained that represents the separation among these 7 classes. The dataset statistical test results showed that Kruskal–Walli's significant value was < 0.05 , meaning that the ROI sampling had good spectral values separation among the classes (Table 2). Both ROI samples of the month were separated statistically. The spatial distribution of class samples is shown in the attachment Fig. 9.

The CTA analysis input contained all 16 bands of the H₈ satellite data and another transformation band proposed by [15], as summarized in Table 1. The transformation bands B18 and B19 were used to detect smoke. The B20 and B21 (NDVI) were employed to identify cloud and vegetation, respectively. This procedure aimed to establish the response and the best threshold in the Classification Tree (CTA models) from various bands' input.

3.2 Classification tree analysis (CTA)

Classification Tree Analysis, a statistical method used for more than 50 years, was employed to differentiate the data samples into several classes based on splitting. It involves recursive partitioning of a dimensional space defined by the predictors into several groups that share homogeneity in response. This results in a hierarchy of groups within groups, known as a tree, while the group is called a node. The final nodes are called leaves. The two splitting algorithms (impurity functions) were the entropy $\theta(t) = - \sum_j P(j|t) \text{Log} P(j|t)$ and the Gini index $\theta(t) = 1 - \sum_j P^2(j|t)$, where $P(j|t)$ is the proportion of the number of elements in class j to that of node t [17, 21, 26].

In the development of the classification tree, the entropy for the dataset (sample 1, -09) was first

Table 2 Test statistics result

	B1	B2	B3	B4	B5	B6	B7
Chi square	5.771279	5.782264	5.889754	5.929563	10.146518	8.968875	10.771556
df	6	6	6	6	6	6	6
Asymp. sig.	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}
	B8	B9	B10	B11	B12	B13	B14
Chi Square	10.197467	10.162263	10.302658	10.915090	10.847466	10.901275	10.885237
df	6	6	6	6	6	6	6
Asymp. sig.	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}
	B15	B16	B17	B18	B19	B20	B21
Chi Square	10.847697	10.787432	10.901275	9.324185	9.113767	7.093633	6.464232
df	6	6	6	6	6	6	6
Asymp. sig.	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}	1.0×10^{-7}

Kruskal–Wallis test

Grouping Variable: class

The Asymp. Sig in Kruskal–Wallis Test telling that each class is separated well according to the Himawari-8 spectral band and spectral transformation band value (B1 to B21). The Asymp.Sig value is ≤ 0.05

Fig. 2 The feature space of the ROI in each class. The 3 bands were the components of RGB for aerosols, as proposed by [24]. The data were obtained from the Himawari_8 satellite on September 4, 2015, 01.00 UTC (Sample 1, left) and October 14, 2015, 02.00 UTC (Sample 2, right). This figure is also used in [11]

computed using impurity function entropy $\mathcal{H}(\mathbf{t})$. Every feature, band input for the CTA ($\mathbf{H}(\mathbf{t})$), was calculated. The entropy average feature $\sum \mathbf{P}(\mathbf{t})\mathbf{H}(\mathbf{t})$ was used to obtain the split's total entropy. Information gain (IG) was used in calculating the decreasing entropy: $\text{IG} = \text{Entropy all dataset} - \text{Entropy average feature}$. The sequence process was repeated for the next feature, from band 1 to all the band input. The process continued until the highest Information gain was revealed. The feature with the highest IG was chosen as a node. The process continued until the last leaves. The procedure was then applied to the other impurity method (Gini Index) and the other sample (sample 2, -10).

The trees produced from this CTA are shown in Fig. 3. Before being applied as an automatic smoke detection model, each tree must be subjected to spatial, statistical, and time consumption analyses.

Overall accuracy, probability of detection (POD), and false alarm rate (FAR) were employed to calculate the accuracy of the Classification Tree Analysis model (Fig. 3). The analysis was limited to daytime data. This study performed an overall accuracy assessment, OA. This is the probability that the model was correctly classified, expressed as the sum of the true positive and negative, divided by the total number of datasets [27]. The probability of detection POD was also tested. It is the ratio of the number of correct

Fig. 3 The CTA results using H₈ data on September 4, 2015, 01.00 UTC (named: -09) and October 14, 2015, 02.00 UTC (named: -10). **a1** Entropy-09: the tree of Sample 1 using the Entropy method; **a2** Gini-09: the tree of Sample 1 using the Gini Index; **b1** Entropy-10: the tree of Sample 2 using the Entropy method; **b2** Gini-10: the tree of Sample 2 using the Gini Index; Class 1 represents the Cb-cloud, class 2 represents the Low and Mid cloud, class 3 represents the smoke, class 4 represents the bare soil, class 5 represents the thin cirrus cloud, class 6 represents the vegetation, while class 7 represents the water body

model result hits (A) to the number of the observed event (A + C), as seen in Table 3 [28]. Lastly, the false alarm rate, FAR, was assessed, which indicated the existence of the smoke phenomenon, while reality did not [29]. A contingency table was used to calculate the hit and miss of the CTA classification model based on a previous study [11] (Table 3). The CTA-based smoke detection model is categorized as a hit when it has the same result as the METAR from airport observation for both no-smoke and smoke occurrence, while its opposite is called a miss. In this model, a detected smoke was labelled with 'smoke,' while the class of vegetation, bare soil, or water was classified as 'no-smoke.' This procedure was conducted after cloud screening, meaning the assessment was only for free-of-cloud pixels. Each pixel of the Himawari-8 CTA model nearest each aerodrome's geographic position was evaluated to identify the best accurate CTA model compared to the actual condition.

The CTA-based smoke detection models' accuracy assessment was evaluated using the data from 5 randomly selected timestamps. The data was taken from 1 July 30, 2015, 03.00 UTC, August 16, 2015, 01.00 UTC, September 29, 2015, 08.00 UTC, October 19, 2015, 02.00 UTC, October 21, 2015, 02.00 UTC and November 2, 2015, 01.00 UTC. The CTA-based smoke models from the satellite data were then compared with 47 METARs from airport observations. Aside from the statistical accuracy assessment, spatial performance, and the CTA model's execution time were also considered.

Table 3 Contingency table and overall accuracy (OA), probability of detections (POD), and false alarm rate (FAR) formula

	Smoke Himawari-8 classification	
	Yes (smoke pixel)	No (vegetation/bare soil pixel)
Smoke airport observation		
Yes	A	B
No	C	D

A, B, C, and D denote the number of occurrences for each case. Overall accuracy is define as $OA = (A + D) / (A + B + C + D)$; Probability of detection is define as $POD = A / (A + B)$; and False Alarm Rate define as $FAR = C / (A + C)$

3.3 Spatial performance of the classification tree analysis (CTA) results

The spatial performance of the CTA results was analysed from the monthly average of the smoke pixel occurrences. The CTA models were implemented in each timestamp of H₈ in one month, producing 8 images out of the 4 CTA models in September and October 2015. The images were analysed for their spatial consistency and then compared with previous research results in the same area. The modified MODIS aerosol retrieval [31] is used as spatial data comparing because this research explored the aerosol characteristic during the 2015 forest fire event over Indonesia.

3.4 Smoke information scheme based on the CTA result in a small airport

H₈ satellite data have a 10-minute temporal resolution, with higher precision or information than weather observation. However, the smoke pixel in H₈ is sometimes blocked by cloud pixels. Accordingly, combining several timestamps of the H₈ satellite was proposed to minimize this disadvantage. The concept of smoke information in the H₈ satellite is merging the nearest timestamp of satellite observation with the mandatory time of airport weather observation. Using this approach, H₈ provides smoke detection with a 30-minute temporal resolution. As a result, it creates more precise information than the mandatory observation time, which, according to the Regional Navigation Agreement (RNA), is one hour in several airports.

The case study involved a small island named Tanjung Pandan, which has one major airport with mandatory hourly weather observation. A 30-minute temporal resolution of smoke information based on H₈ satellite data was combined with the satellite scanning at 3 timestamps. The 3 actual weather conditions reported in METAR (weather observation) were used to evaluate the CTA model performance with thick smoke, partly cloudy and thin smoke. The best CTA model was then implemented in this small airport.

4 Results and discussion

The Entropy-09 model (Fig. 3a1) statistically distinct underlying surface and cirrus over water from aerosol (smoke, Cb-Cloud top, and low-middle cloud) according to the difference in the reflectance values (B1) of H₈ satellite. Both aerosols and low and middle cloud showed different B4 values from smoke and Cb-cloud top. Based on the B7 pixel value, the Cb-cloud top's microphysics characteristics

Fig. 4 Box plot of the CTA model execution time. The tests for each impurity method used inputs from the same timestamp. The test was using 60 datasets (Himawari-8 data on September 1–2, 2015 daytime in 10-min resolution)

are different from that of smoke. The Gini Index-09 model (Fig. 3a2) first separated the water body from all classes based on the differentiation in their B4 values. The Cb-cloud top, smoke, thin cirrus, and vegetation were then separated from the low-middle cloud and bare soil using the B5 value. The Cb cloud's microphysics is relatively different from smoke, bare soil, thin cirrus, and vegetation. The smoke was distinguished from bare soil, thin cirrus, and vegetation according to the B2 value.

The Entropy-10 (Fig. 3b1) model output showed that the Cb-cloud top had different aerosol microphysical properties (B7) from the 6 other classes. The low and middle cloud differed from smoke, bare soil, cirrus cloud over water, vegetation, and water body in cloud top information (B13) and reflectance characteristics (B1). Furthermore, the classes of the water body and cirrus cloud over water were split from smoke and the underlying surface by B4 values. The smoke was then revealed as an independent class because its reflectance value was different from bare soil and vegetation. The Gini Index-10 (Fig. 3b2) yielded a CTA model similar to the CTA E-10 model in distinguishing the Cb-Cloud top from other classes and smoke from the underlying surface (vegetation and bare soil). Although the impurity methods used the same data input (sample no.2), they created different trees with few similarities.

Figure 4 shows the execution time of the CTA model. The CTA model developed with Entropy performed faster than the one generated using the Gini Index. According to the execution time, both Entropy-09 and Entropy-10 are the best and most efficient methods. Therefore, they sufficiently develop CTA-based smoke detection models using H₈ satellite data.

Figures 5 and 6 show the CTA model's spatial performance for September and October 2015. The maps display the frequency of smoke pixel occurrence in the CTA-based smoke detection models. The blue and yellow colours indicate the maximum and minimum frequencies of smoke occurrences in 1 month. The white colour shows that the smoke occurrence in the CTA-based smoke detection models is less than 20%.

The CTA models from ROI sample 1 (Fig. 5 left, Entropy-09, and Gini Index-09) show a wide area of smoke detected over the land and oceans. Nearly the entire south-western part of Borneo and the central part of Sumatera Island were covered by smoke in September 2015. Also, the Indian Oceans on the south-west of Sumatera and the majority of the south-eastern part of the South China Sea was covered by smoke. The highest frequencies of smoke occurrences were found above the water bodies on the west of Borneo and east of Sumatera. However, several pixel areas of smoke occurrences above the land are in contrast to the nearby oceans. Less than 20% of smoke pixels above the ground and more than 20% over the adjacent seas were detected in Sumatera Island's western and southern parts. Also, there was a difference in the western and south-western parts of Borneo, to the nearby ocean. These results indicate that the spectral mixture of smoke over the land may differ from the one over the oceans.

A different result is presented in Fig. 5 right (Entropy-10 and Gini Index-10). Based on the CTA models using the data of sample 2, the weather was dominated by frequencies of smoke pixel occurrence above the land in September 2015. The highest frequencies were detected over the southern part of central Borneo and the western part of central Sumatera Island. The same frequency over the ocean was considerably low.

Figure 6 shows the frequency of smoke pixel occurrence using the CTA models in October 2015. It depicts several similarities with Fig. 5, except for the frequency and the extent of the area covered with the smoke pixel. Several contrasts in smoke pixels' appearance were also revealed in the CTA model developed from sample 1. The smoke frequency was dominant above the land in the CTA model generated from sample 2 in October 2015.

Based on these eight images, the CTA model generated from sample 1 was concluded as spatially and logically unacceptable. It also overestimated the results because of 2 reasons. First, there is a gap area, evidenced by the low smoke pixel occurrences in the nearby oceans where frequent smoke incidence occurs. Second, high frequencies of smoke pixels spread from above the sea on the north of Java Island to the northern shore of Java Island. These conditions require further validation using the smoke data above the ocean.

Fig. 5 The frequency of smoke occurrences in September 2015, as estimated with 4 CTA Models (Entropy-09, Gini Index-09, Entropy-10, and Gini Index-10)

The CTA models developed from sample 1 in September and October 2015 (Entropy-09 and Gini Index-09) showed an overestimation in the areas of smoke pixel occurrence for both months (Figs. 5 and 6). There was a broad area of smoke pixels in the Indian Ocean, the Strait of Malacca, the South China Sea, and the Java Sea. A gap area was identified between the frequency of smoke pixel occurrence on the land and above the nearby ocean. These results are different from [30]. From the MODIS Aerosol Optical Depth, this research found that smoke pixel occurrence frequency dominated the weather above the land with a stable dispersion to the nearby oceans without any frequency gap. This was probably caused by the different spectral mixture of the smoke properties above the ground and water body, which were caught during the sampling process of ROI 1 (sample 1).

The CTA model based on sample 2 (Entropy-10 and Gini Index-10) in September and October 2015 showed that the smoke pixel was dominant above the land area but present in minor quantities above the sea. These results are logically acceptable due to the absence of any contrasts between land and the nearby ocean. They are also in line with Shi et al. [30] in which the smoke pixel mainly distributes above land.

A conclusion from both samples includes using the peak of forest fire season for the ROI sampling in October. When forest fire reaches its maximum, the smoke becomes easily distinguished from the other objects' background. With a careful process, better ROI sampling is achieved during this phenomenon's peak, creating a better CTA-based smoke detection model.

In this research, the validation used the miss and hit in the smoke detection model. The CTA models developed from H₈ satellite data produced several classes. A class is labelled with 'miss' when the model categorizes it as vegetation or bare soil, but the airport's actual observation shows obscuration by smoke. A class categorized as 'hit' when the smoke appearance is supported by the weather observation at the airport (smoke occurs). Because the smoke classification had already excluded cloud objects, the validation was not performed on pixels classified as clouds. The cloud obscures the spectral smoke information.

Based on the statistical accuracy assessment, the Entropy-10 model of smoke detection is the most accurate among the 4 CTA models because of 2 reasons. First, it shows the same result as the actual weather condition. This is because the relationship between the actual and the corresponding modelled data is more than 80%.

Fig. 6 The frequency of smoke occurrences in October 2015, as estimated with 4 CTA models (Entropy-09, Gini Index-09, Entropy-10, and Gini Index-10)

Second, Entropy-10 has the lowest FAR value, indicating very low false in the classification result. A low FAR provides the capacity to minimize the false classification of smoke. Based on the spatial analysis, accuracy, and execution time, CTA E-10 is considered the best model (Fig. 7).

Figure 8 shows the CTA-E10 model results for 3 conditions, including smoky, partly cloudy, and very transparent

smoke. The CTA-E10 model outputs combined 3 timestamps of Himawari_8 data having a 10-min resolution, creating a time length of 30 min. For example, in the image produced by this model, the time stamp shows October 25, 2015, 02.00 UTC, meaning that it merges the information recorded from 01.40 to 02.00 UTC. According to the CTA E-10 model's object identification, the yellow

Fig. 7 A bar chart of the accuracy assessment results of the CTA models using samples from 7 different times (July 30, 2015, 03.00 UTC; August 18, 2015, 01.00 UTC; September 29, 2015, 08.00 UTC; October 19, 2015, 02.00 UTC; October 19, 2015, 05.00 UTC; October 21, 2015, 02.00 UTC; November 2, 2015, 01.00 UTC)

Fig. 8 The CTA model output over Tanjung Pandan Area (WIKT). The satellite data on October 24, 2015, 05.00 UTC showed smoky condition (a), September 26, 2015, 07.00 UTC was partly cloudy (b), October 25, 2015, 02.00 UTC had transparent smoke (c). The images were created from the last 3, 10-minute Himawari_8 data. METAR, containing data of weather conditions, prevailing visibility, and cloud information, represents the actual state during the time of observation. The green and white circles represent the weather within 16 km and 8 km, respectively, from the aerodrome centre. *wx* weather condition, *vis* prevailing visibility over the aerodrome, *cld* cloud cover amount and cloud base height

pixel represents the smoke, the blue pixel is cumulonimbus, low or middle cloud, the grey pixel is cirrus cloud, and the white pixel is bare soil, vegetation or water body. The green and white circles are areas within the radius of 16 km and 8 km from the aerodrome centre, respectively.

The CTA model output detected a smoke area (Fig. 8a, on October 24, 2015, 05.00 UTC). Smoke covered more than 50% of areas located 8 and 16 km from the aerodrome centre. According to the ICAO's document number 8896 about the manual of aeronautical weather observation, weather conditions for air navigation information must represent the area within the radius of 8–16 km from the centre of the aerodrome. The weather information within 8 km is used as based data described in the local routine report. It is the crucial information required for take-off and landing. Similarly, the weather information within 16 km is the based data of METAR. This includes the information of actual weather over the aerodrome, which is then disseminated to the outside.

In a partly cloudy condition (Fig. 8b, on September 26, 2015, 07.00 UTC), the CTA model output was very satisfying. There was no smoke in most of the aerodrome, and clouds were present in 30–40% of the area. This model output is in line with the actual weather condition at the aerodrome, recorded in METAR. Although a few smoke pixels were identified in the aerodrome, the condition was too insignificant to be considered in the weather observation.

The disadvantage of satellite-based smoke detection is the low scanning sensitivity when the weather is filled with thin smoke. On October 25, 2015, 02.00 UTC (Fig. 8c), there was no smoke in the aerodrome. However, the METAR indicated a presence of smoke with a visibility of 3500 m. The CTA model would indicate a cloud coverage of 10–20% of the aerodrome, which is in line with the information published in METAR (FEW020 means that the lowest layer of cloud is present in almost 10–20% of the sky area). This phenomenon is referred to as thin smoke. It disables the spectral value of the satellite data from distinguishing aerosol from the underlying surface.

This CTA model output is handy for weather observers and forecasters because it provides highly accurate information based on remote sensing products. This offers comprehensive information for aviation users to improve their situational awareness of the smoke hazard (Fig. 9).

Smoke detection as adverse weather provides numerous benefits for weather information providers, aviation operators, users, and stakeholders. Weather station, as an information provider, provides spatial information from the orbital technology. This information justifies the actual observation and offers consideration for forecaster in predicting the weather condition. This adverse weather visualization is convenient for users of aviation meteorological

Color legend of each class		
	Smoke	
Low and Mid Cloud	Thin Cirrus Cloud	
Vegetation	Water Body	
Bare soil	Cumulonimbus Cloud Top	

Fig. 9 Spatial Distribution of ROI sample date October 14, 2005, 02.00 UTC (sample:-10), with the basemap is using RGB Day Natural Color. The red beam is band 1.6 μm , the green beam is band 0.8 μm , and the blue beam is band 0.6 μm

products and services to improve their situational weather awareness [31].

5 Conclusions

CTA-based smoke detection model detects smoke automatically from the orbital satellite data. This statistical approach builds a path of classification based on the sample data. The model's output depends on several factors, such as ROI sampling, band combination input, and adopted impurity methods (Entropy, Gini Index, or other). Entropy-10 (E10) is the best CTA model examined in this research based on its overall accuracy, POD and FAR, spatial performance analysis, and execution time. This model can be implemented in Indonesia, especially Sumatra and Borneo Islands because its ROI sampling was performed during the highest forest fire smoke incident in the last 2 decades. Smoke pixel is undetected, possibly because there is no smoke revealed in said pixel, cloud covers the smoke pixel, or the smoke is transparent (the pixel value is embedded with land/underlying surface).

This research is beneficial to the meteorological service providers, specifically in disseminating spatial information of forest and land fire smoke. This is because the occurrence of thick smoke is dangerous for flight operation safety. Aviation stakeholders need this spatial information to improve their situational awareness related on smoke hazard.

Acknowledgements This research is part of the dissertation of the first author. Profound gratitude is extended to the Centre for Education and Training of BMKG for funding the Ph.D. programme of the first author in the Faculty of Geography, Universitas Gadjah Mada. Furthermore, gratitude is extended to Mr. Andersen Panjaitan and Mr. Alpon Sepriando from the Centre of Public Meteorology, BMKG, Indonesia for providing the WIS Himawari_8 satellite data and invaluable discussion, as well as Mr. Retnadi Heru Jatmiko and Restu7 from the Faculty of Geography, Universitas Gadjah Mada for their significant input in digital image processing.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Aouizerats B, Van der Werf GR, Balasubramanian R, Betha R (2015) Importance of transboundary transport of biomass burning emissions to regional air quality in Southeast Asia during a high fire event. *Atmos Chem Phys*. <https://doi.org/10.5194/Acp-15-363-2015>
- Behera S, Betha R, Huang X, Balasubramanian R (2015) Characterization and estimation of human airway deposition of size-resolved particulate-bound trace elements during a recent haze episode in Southeast Asia. *Environ Sci Pollut Res*. <https://doi.org/10.1007/s11356-014-3645-6>
- Gaveau D, Salim M, Hergoualch K, Locatelli B, Sloan S, Wooster M et al (2014) Major emission from peat fires in Southeast Asia during non-drought years: evidence from the 2013 Sumatran fires. *Sci Rep*. <https://doi.org/10.1038/srep06112>
- Radojevic M (2003) Chemistry of forest fires and regional haze with emphasis on Southeast Asia. *Pure appl Geophys* 160(1–2):157–187
- Reid J, Hyer E, Johnson R, Holben B, Yokelson R, Zhang J, Chin S (2013) Observing and understanding the Southeast Asian aerosol system by remote sensing: an initial review and analysis for the seven Southeast Asian Studies (7SEAS) program. *Atmos Res*. <https://doi.org/10.1016/j.atmosres.2012.06.005>
- Murdiyarto D, Lebel L, Gintings A, Tampubolon S, Heil A, Wasson M (2004) Policy responses to complex environmental problems: insights from a science–policy activity on transboundary haze from vegetation fires in Southeast Asia. *Agric Ecosyst Environ*. <https://doi.org/10.1016/j.agee.2004.01.005>
- Aiken SR (2004) Runaway fires, smoke-haze pollution, and unnatural disasters in Indonesia. *Geogr Rev* 94(1):55–79
- Isyanto H, Hartono H, Marfai MA (2018) Classification tree analysis (Gini-Index) smoke detection using Himawari_8 satellite data over Sumatera-Borneo maritime continent Sout East Asia. In: The 2nd international conference on environmental resources management in global region (ICERM 2018) 22–23 October 2018, Yogyakarta, Indonesia, Yogyakarta, 2020
- Chung Y, Kim H (2008) Satellite monitoring of forest fires and associated smoke plumes occurring in Korea. *Air Qual Health*. <https://doi.org/10.1007/s11869-008-0012-0>
- Lensky IM, Rosenfeld D (2008) Clouds-aerosols-precipitation satellite analysis tool (CAPSAT). *Atmos Chem Phys* 8(2):4765–4809
- Isyanto H, Hartono H, Marfai MA (2018) Smoke detection and visibility estimation using Himawari_8 satellite data over Sumatera and Borneo Island Indonesia. *Spat Inf Res*: 1–12
- Wang J, Song W, Wang W, Zhang Y, Liu S (2011) A new algorithm for forest fire smoke detection based on MODIS data in Heilongjiang Province. In: Proceedings of the 2011 international conference on remote sensing, environment and transportation engineering (RSETE), Nanjing, China, 24–26 June 2011, pp 5–8
- Xie Y, Qu J, Xiong X, Hao X, Che N, Sommers W (2007) Smoke plume detection in the eastern United States using MODIS. *Int J Remote Sens* 1:2367–2374. <https://doi.org/10.1080/01431160701236795>
- Nagatani I, Kudoh J, Kawano K (2012) A new technique for visualization of forest fire smoke plume using MODIS data. In: Geoscience and remote sensing symposium (IGARSS), 2012
- Li X, Song W, Lian L, Wei X (2015) Automatic detection of fire smoke using artificial neural networks and threshold approaches applied to AVHRR imagery. *Remote Sens* 1:4473–4498. <https://doi.org/10.3390/rs70404473>
- Li Z, Khananian A, Fraser R, Cihlar J (2001) Automatic detection of fire smoke using artificial neural networks and threshold approaches applied to AVHRR imagery. *IEEE Trans Geosci* 1:1859–1870. <https://doi.org/10.1109/36.951076>
- Loh WY (2014) Fifty years of classification and regression trees. *Int Stat Rev* 1:329–348. <https://doi.org/10.1111/insr.12016>
- Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM (2016) Landslide susceptibility mapping using random forest, boosted regression, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslide*. <https://doi.org/10.1007/s10346-015-0614-1>
- Dutta D, Chaudhuri S (2015) Nowcasting visibility during winter-time fog over the airport of a metropolis of India: decision tree algorithm and artificial neural network approach. *Nat Hazards* 75(2):1349–1368
- Kurt I, Ture M, Kurum AT (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl* 34(1):366–374
- Shaaban K, Pande A (2016) Classification tree analysis of factors affecting parking choices in Qatar. *Case Study Transp Policy* 1:88–95. <https://doi.org/10.1016/j.cstp.2015.11.002>
- Baum BA, Trepte Q (1999) A grouped threshold approach for scene identification in AVHRR imagery. *J Atmos Ocean Technol* 16(6):793–800
- Bessho K, Date K, Hayashi M, Ikeda A, Imai T, Inoue H, Kumagai Y, Miyakawa T, Murata H, Ohno T, Okuyama A, Oyama R, Sasaki Y, Shimazu Y, Shimoji K, Sumida Y, Suzuki M, Taniguchi H, Tsuchiyama H, Uesawa D, Yokota H, Yoshida R (2016) An introduction to Himawari-8/9—Japan's new-generation geostationary meteorological satellites. *J Meteorol Soc Jpn Ser II* 94(2):151–183
- Rosenfeld D, Lensky IM (1998) Satellite-based insight into precipitation formation processes in continental and maritime convective clouds. *Bull Am Meteorol Soc* 79(11):2457–2476
- Hipólito RA, Martins CR (2010) Prevalência de alterações da mucosa bucal em adolescentes brasileiros institucionalizados em dois centros de reeducação. *Ciência & Saúde Coletiva* 15(suppl 2):3233–3242
- Li X, Claramunt C (2006) A spatial entropy-based decision tree for classification of geographical information. *Trans GIS*. <https://doi.org/10.1111/j.1467-9671.2006.01006.x>
- Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M (2004) The use of “overall accuracy” to evaluate the validity of screening or diagnostic tests. *J Gen Int Med* 19:460–465
- Sharma K, Ashrit R, Bhatla R, Mitra A, Iyengar G, Rajagopal E (2017) Skill of predicting heavy rainfall over India: improvement in recent years using UKMO global model. *Pure appl Geophys* 174:4241–4250. <https://doi.org/10.1007/s00024-017-1640-9>
- Barnes L, Grunfest E, Hayden M, Schultz D, Benight C (2006) False alarms and close calls: a conceptual model of warning accuracy. *Weather Forecast* 22:1140–1154
- Shi YR, Levy RC, Eck TE, Fisher B, Mattoo S, Remer L, Slutsker I, Zhang J (2019) Characterizing the 2015 Indonesia fire event using modified MODIS aerosol retrievals. *Atmos Chem Phys* 1:259–274. <https://doi.org/10.5194/acp-19-259-2019>
- Chamberlain JP, Latorella KA (2001) Convective weather detection by general aviation pilots with conventional and data-linked graphical weather information sources. In: 20th digital avionics system conference (DASC), 2001

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.