

Research Article

Classification of Continuous Sky Brightness Data Using Random Forest

Rhorom Priyatikanto ¹, Lidia Mayangsari ², Rudi A. Prihandoko,³
and Agustinus G. Admiranto¹

¹Space Science Center, National Institute of Aeronautics and Space, Bandung 40173, Indonesia

²School of Business and Management, Institut Teknologi Bandung, Bandung 40132, Indonesia

³Department of Computer Sciences and Electronics, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

Correspondence should be addressed to Rhorom Priyatikanto; rhorom.priyatikanto@lapan.go.id

Received 17 September 2019; Revised 26 February 2020; Accepted 27 February 2020; Published 1 April 2020

Academic Editor: Michael Kueppers

Copyright © 2020 Rhorom Priyatikanto et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sky brightness measuring and monitoring are required to mitigate the negative effect of light pollution as a byproduct of modern civilization. Good handling of a pile of sky brightness data includes evaluation and classification of the data according to its quality and characteristics such that further analysis and inference can be conducted properly. This study aims to develop a classification model based on Random Forest algorithm and to evaluate its performance. Using sky brightness data from 1250 nights with minute temporal resolution acquired at eight different stations in Indonesia, datasets consisting of 15 features were created to train and test the model. Those features were extracted from the observation time, the global statistics of nightly sky brightness, or the light curve characteristics. Among those features, 10 are considered to be the most important for the classification task. The model was trained to classify the data into six classes (1: peculiar data, 2: overcast, 3: cloudy, 4: clear, 5: moonlit-cloudy, and 6: moonlit-clear) and then tested to achieve high accuracy (92%) and scores (F -score = 84% and G -mean = 84%). Some misclassifications exist, but the classification results are considerably good as indicated by posterior distributions of the sky brightness as a function of classes. Data classified as class-4 have sharp distribution with typical full width at half maximum of 1.5 mag/arcsec^2 , while distributions of class-2 and -3 are left skewed with the latter having lighter tail. Due to the moonlight, distributions of class-5 and -6 data are more smeared or have larger spread. These results demonstrate that the established classification model is reasonably good and consistent.

1. Introduction

Global progression of human civilization cannot be separated from the multidimensional side effects it causes. As emphasized by Lewis and Maslin [1], human activity affects global environment in many ways, including major changes in biogeochemical cycles, increase of species extinction rates that lead to extreme term of the sixth mass extinction in Earth's history, and significant alteration of land surface, atmosphere, and oceans. In the last decades, extensive uses of artificial lights have also altered the daily rhythm of human activity and even disturbed other species' life cycle. Excess use of artificial lights inundates natural dark sky at

night with light pollution which has various negative impacts ranging from the disruption of the ecology [2], health problems [3, 4], wasted energy [5], and also loss in cultural asset [6]. With those impacts, concerns on light pollution increase significantly these days.

Numerous efforts to quantify the level of light pollution and to mitigate its impacts have been conducted by many researchers through various measures using different types of instruments to obtain one- or two-dimensional sky brightness or even the spectra [7]. Quantification of light pollution on a global scale was conducted by means of accurate modeling of artificial light at night based on satellite imagery [8, 9]. On the other hand, local measurements of sky

brightness at night and the portion of light pollution were performed by means of observation campaigns and continuous monitoring using different instruments [10–12]. Among those measures, utilization of low-cost photodiode-based Sky Quality Meter (SQM) makes sky brightness measuring and monitoring significantly easier.

The increasing use of the SQM for continuous sky brightness measurement opens the stream of data which are valuable for the mitigation of light pollution and other purposes. Measurements conducted in various locations and commenced in long time span increase the volume and variety of the data generated. Typically, SQM produces approximately 50 kilobytes of data in one night when it is operated in one-minute cadence, but multiyear data product will be quite large. A massive dataset can be aggregated into univariate or multivariate distribution of sky brightness for further analysis and inference. Density plots or jellyfish diagrams are often used to visualize the data acquired systematically in long duration [12, 13].

Careful examination of the nightly data is required in order to suppress the noise and to avoid biased inference. In most of the cases, SQM data need to be evaluated whether they are moonlight affected or not [14], whether they are common or outliers, whether they are obtained during clear, cloudy, or overcast night [15, 16]. Classification of the nightly SQM data is also needed for specific research objectives. For example, to evaluate the amplification of ecological light pollution by clouds, Kyba et al. [15] used SQM data acquired from three distinguished observing conditions, namely, clear, cloudy, and overcast night. This kind of evaluation and classification is a laborious task such that it should not be done manually and machines should work automatically for this purpose.

Nightly SQM data are analogous to the diverse light curves of variable stars such that similar automatic classification algorithms [17, 18] are expected to be able to evaluate and classify sky brightness data from SQMs. In variable stars classifications, both periodic and nonperiodic features can be extracted from the light curves and then stuffed to particular machine learning algorithm for obtaining a certain classification model. The algorithm can be classification trees, random forests, Support Vector Machines, likelihood classifier, or artificial neural networks [18, 19].

This study aims to develop a model for classification of nightly SQM data and to evaluate its performance. Such a model is expected to help researchers to handle the data appropriately and to compile good data from an agglomeration of data with different qualities.

2. Data and Methods

2.1. Data and Feature Extraction. Several Unihedron Sky Quality Meters (type LU-DL) were used to measure sky brightness over eight different locations in Indonesia since April 2018 and were still working in early 2020. This type of instrument uses converging lense and color compensating filter such that it has a moderate field of view of about 20° (FWHM) and an effective spectral window between 400 and

600 nm [20, 21]. SQM measures sky brightness in terms of magnitude per square arcsecond (mpsas) with absolute precision of 0.1 mpsas [7]. With a fine performance, SQMs were used extensively in enormous research and monitoring around the world even for the calibration of global map of light pollution [9].

Zenithal sky brightness data used in this study were acquired from eight locations with different conditions in terms of artificial light contamination. As summarized in Table 1, some locations can be considered to be situated in urban areas where artificial light pollution is overwhelming, while the others are in periurban and rural areas. Sky brightness measured at those locations may behave differently as local atmospheric conditions and light pollution influence the observed sky brightness. Clouds may darken the rural sky as it covers a fraction of celestial lights and high altitude atmospheric emissions. On the other hand, clouds partially reflect artificial lights from ground and brighten the sky or amplify the light pollution [15, 16] such that the natural regime of sky brightness is masked [22]. By acquiring data from locations with different conditions, global characteristics and behavior of sky brightness can be understood more thoroughly.

At every location, continuous measurement of sky brightness was conducted from April 2018 to December 2018, though measurements at some locations only cover shorter durations. One-minute sampling rate was used such that a dataset obtained in one observing night contains up to 750 rows. The two most important variables recorded and used in the current study are time (in UT and local time) and sky brightness (in magnitude per square arcsecond, mpsas). Using those two variables, light curve for every observing night can be constructed for further analysis. The magnitude recorded by the SQM is in inverse logarithmic scale so that higher value of mpsas means darker sky; e.g., each 1 magnitude increase means that the observed patch of the sky is 2.5 times dimmer. Typical brightness of a clear rural sky is more than 21.6 mpsas while the light polluted sky in the proximity of a city may be hundreds times brighter than the natural sky (~ 16 mpsas).

By examining the light curves visually, experienced users can assess and classify them into several classes (see Figure 1). In this study, six distinguished classes are defined. They are class-1 for peculiar/invalid data, class-2 for data acquired during overcast conditions, class-3 for cloudy nights, class-4 for clear and moonless nights, class-5 for cloudy nights with moonlight, and class-6 for clear and moonlit nights. Weird data are characterized by unreasonable value of sky brightness (e.g., extremely high or low) or light curves with pattern beyond expectation (e.g., perfectly flat light curve during moonlit night). If the light curves exhibit fluctuation or brightening due to clouds for more than 50% of the time, then they are categorized as class-2; otherwise, they belong to class-3. Brightening due to scattered moonlight as a function of its phase and zenith distance is a gradual and smooth process [23] which can be identified quite easily (for class-6 data), unless clouds interfere and the light curves become wiggly (class-5). Class-4 contains smooth light curves with plateau (constant sky

TABLE 1: Locations of the sky brightness measurements conducted from April to December 2018. The fourth column lists the number of observing nights.

	Longitude	Latitude	#Data	Remark
Agam (AGM)	100.320	-0.204	150	Rural
Bandung (BDG)	107.586	-6.894	53	Urban
Biak (BIK)	136.101	-1.174	175	Rural
Garut (GRT)	107.692	-7.649	226	Rural
Pasuruan (PSR)	112.676	-7.567	232	Urban
Pontianak (PTK)	109.366	-0.003	117	Periurban
Subang (SBG)	107.769	-6.562	57	Urban
Sumedang (SMD)	107.837	-6.913	240	Rural

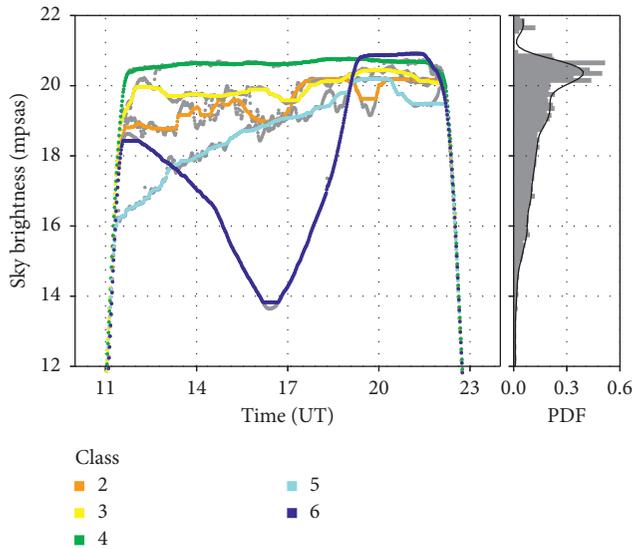


FIGURE 1: Illustration of nightly SQM data acquired in various conditions: overcast (class-2), cloudy (class-3), clear sky (class-4), cloudy moonlit (class-5), and clear moonlit (class-6). Raw data are displayed as gray dots while median filtered data are in color. Probability density function (PDF) constructed using whole month data is displayed on the right panel as reference.

brightness) for the whole night. Those characteristics can be described using a number of features, some of which are often found in the analysis of light curves of variable stars [19, 24, 25]. Table 2 summarizes 15 features extracted from the data.

Most of the features were extracted from a single variable which is the sky brightness. Beside the raw sky brightness data, smoothed light curve which is obtained through median filtering was taken into account. Smoothing window of 61 data points which corresponds to 1-hour continuous measurements was used. This smoothed model was required as the basis to quantify the fluctuation of sky brightness. Lastly, time of measurement was used to define night time and to calculate the ephemeris of the moon such that moon phase and zenith distance were extracted as features.

After the feature extraction process, the dataset which contains 1250 rows and 15 columns is ready for the machine learning stage. Each row in this dataset has its own label (1, 2, ..., 6) according to the manual evaluation by one of the

authors (RP). For every data row, the evaluation was conducted by displaying the nightly light curves superposed with the plot of moon's altitude as a function of time and also the density function constructed from the sky brightness measured in a month. The density function can be regarded as the reference to determine whether the data obtained in a particular night is peculiar or not. According to the characteristics observed from the plot, a certain class was chosen to describe the quality of the data obtained in the whole night. The process was repeated several times in order to suppress the doubt and reduce subjective bias.

The dataset is randomly split into 2 datasets: 875-row training dataset and 375-row testing dataset, while cross-validation data is a subset of the training dataset. Figure 2 summarizes the distribution of classes and training/testing dataset in terms of observation time and locations.

2.2. Random Forest Classifier. Advancement of computing capabilities in the current era enables astronomers to conduct statistical data analysis and inferences more easily. Machine learning has become a popular approach for dealing with massive observational data. Implementations of machine learning technique range from classification of variable stars [18, 19] and regression of photometric redshift [25] to intelligent recognition of predigital images [26].

In this study, Random Forest [27] is utilized for classification tasks since it has several advantageous properties. It works well with data with high dimensionality and it does not require scaling of features or variables. By default, Random Forest Classifier only uses a randomly selected subset of the features such that training or classification process can be quick and parallelization can be performed to make the process faster. This algorithm is considered to be robust to outliers, good in handling nonlinear data, and also able to handle imbalanced data as it tries to minimize the overall error rate. Low bias and low variance are the other advantageous properties of the Random Forest since it is basically an ensemble of decision trees. It is constructed from training dataset through bootstrap aggregation (bagging), where bootstrapped sample is used to grow numerous decision trees, in which sample is recursively divided into more homogeneous subgroups based on the best splitting features and thresholds. Bagging is implemented so that Random Forest becomes robust against overfitting. Gini impurity or information gain is often used as the criteria for selecting the best split during the training. Each tree provides predicted classification, while the average of all trees is the final classification model with higher accuracy compared to single decision tree. In most classification cases, the performance of Random Forest is between Support Vector Machine and Deep Learning [19, 28].

Random Forest Classifier from the *Python* Scikit-Learn [29] is used in this study. Several hyperparameters can be imputed to the algorithm to obtain optimal classification model, among which are the number of trees (`n_estimators`), criteria to measure the quality of a split (using Gini impurity), and the maximum number of features used (`max_features`). As demonstrated by Criminisi et al. [30],

TABLE 2: Description of the features extracted from nightly sky brightness data. The rightmost column summarizes feature importance from Random Forest Classifier with uncertainties calculated using 10-fold cross-validation.

Features	Description	Typical values	Importance
Phase	Moon phase at midnight, calculated using ephemeris package in python	[0, 1]	0.164 ± 0.003
ZDist	Zenith distance of the moon at midnight, calculated using ephemeris	[0, 180]	0.036 ± 0.002
P05	5th percentile of the raw sky brightness, from sunset to sunrise	[0, 25]	0.049 ± 0.004
P50	Median value or 50th percentile of the raw sky brightness	[0, 25]	0.046 ± 0.002
P95	95th percentile of the raw sky brightness	[0, 25]	0.043 ± 0.002
Skew	Skewness of the raw sky brightness	[-10, 10]	0.066 ± 0.003
Kurt	Kurtosis of the raw sky brightness	[-10, 10]	0.043 ± 0.002
Range	P95-P05. This definition is considered to be more robust than max-min	[0, 25]	0.045 ± 0.002
STD	Standard deviation of the raw sky brightness	[0, 10]	0.050 ± 0.002
IQR	Interquartile range of the raw sky brightness	[0, 10]	0.099 ± 0.003
MAD	Mean absolute deviation of the raw sky brightness	[0, 10]	0.084 ± 0.003
PDMed	Percentage of data that deviates more than 2 mpsas from the median value	[0, 1]	0.063 ± 0.003
PDMod	Percentage of data that deviates by more than 0.2 mpsas from the model. The median filtered data (with smoothing window of 61 data points) are defined as the model	[0, 0.5]	0.080 ± 0.005
MADslope	Mean absolute deviation of the slope between two consecutive data points	[0, 0.5]	0.029 ± 0.002
MADRes	Mean absolute deviation of the residue (raw-model)	[0, 0.5]	0.103 ± 0.003

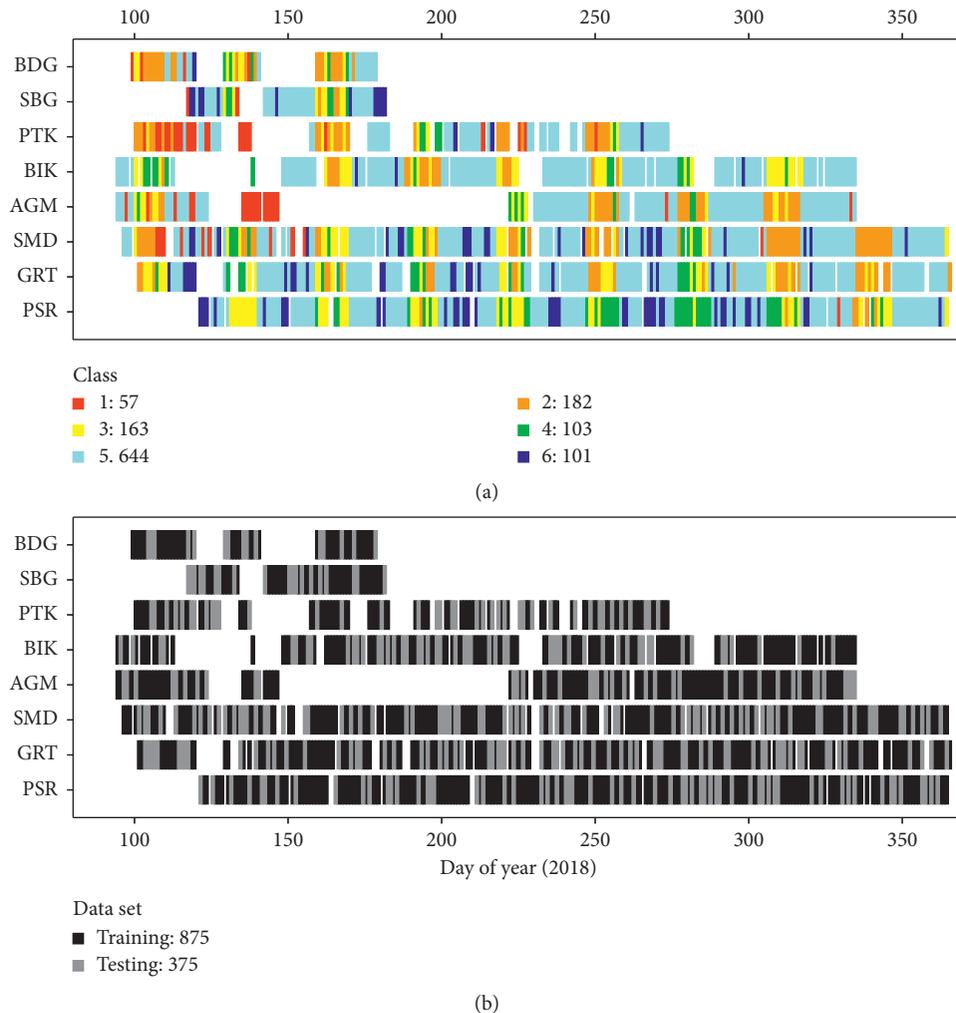


FIGURE 2: Distribution of classes and training/testing dataset.

the number of trees or the size of the forest influences the smoothness of the posterior probabilities resulting from the classifier and also the confidence. The depth of the trees (`max_depth` in Scikit-Learn) also influences the prediction confidence, but Random Forest with too deep trees tends to overfit the training data. In this study, `max_depth` is not set so that the nodes are expanded until all leaves are pure. Balanced Random Forest which assigns class weights is applied, in order to reduce the bias due to imbalanced data.

3. Results and Discussion

In order to get an optimum model for classification of sky brightness data, the training dataset, which contains 875-night sky brightness data with 15 features, was stuffed to Random Forest Classifier. Tenfold stratified cross-validation was used to produce unbiased measure of classification performance. The following analyses are based on the Random Forest Classifier with 300 trees which accounts for a maximum of 10 features. These optimum hyperparameters are deduced from randomized optimization based on the out-of-bag errors resulting from multiple training processes with varied number of trees and maximum features accounted for. In this optimization process, the first hyperparameter is randomly selected between 10 and 1000, while the second one ranges from 1 to 15. Execution time increases linearly as the numbers of trees and maximum features increase.

3.1. Feature Importance. Not only does Random Forest Classifier produce the most accurate model for classifying the data, but it also identifies which features are the most efficient for the classification. In Random Forest, feature importance basically is the average value of importance obtained from all decision trees. In each decision tree, feature importance is a function of purity such that a more important feature is able to split the samples into two subsamples with lower impurity. The obtained feature importance is normalized such that the sum of all values is one.

Feature importance which represents predictive strength of any feature can be utilized to rank feature and to avoid dimensionality curse. In this study, 15 features cannot be considered as a big problem of dimensionality that needs to be pruned, especially when Random Forest is the machine. As mentioned before, this algorithm only selects a portion of the features (with typical size of the square root of the total number of features) such that the execution time grows moderately by $O(n^{1/2})$. However, there are other benefits of analyzing feature importance such as getting a better understanding of the underlying processes that generate the data and reducing the storage requirements [31]. The analysis of feature importance may serve better understanding of the behavior of sky brightness variation in many conditions. Proper feature selection might be of great importance when other machine learning algorithms are used, e.g., Support Vector Machine which is more affected by unimportant or noisy features.

Figure 3 displays the importance of 15 features extracted from the sky brightness data for multiclass case and binary classification case. In the former case, feature importance is byproduct of the Random Forest during the classification of the training dataset into six different classes. Table 2 also summarizes the obtained feature importance in which the uncertainties are the standard deviations from 10-fold cross-validation. In binary classification case, feature importance is calculated separately during yes-no classification, e.g., determining whether a sample belongs to class-1 or not. The overall importance or rank of the features may change slightly by the stochastic nature of the algorithm and also change slightly if the training dataset is altered, but the composition of the top half features does not change.

In general, the Phase becomes the most important feature since this variable effectively classifies the majority of the data. This result is in line with the fundamental classification of sky brightness data into moonlit or moonless class [14]. As displayed in Figure 3, the Phase is ranked first for identifying data with class-2, -3, or -5 while identification of clear sky data (class-4 or -6) requires other features such as MADRes or IQR that works better. Smoothness of the sky brightness data is an important key for finding clear sky data from a pile of data, while for moonless nights, flat sky brightness means small dispersion which can be described by IQR. Apart from those cases, pairwise feature importance shows that identifying data as class-1 requires statistical quantities that describe both dispersion and the shape of distribution. Skew, Kurt, and P05 represent the shape of distribution, while STD and Range describe dispersion. From the perspective of experienced users, class-1 data usually deviates from expectations such that detecting outliers is the key.

3.2. Feature Selection. Some features have significant correlations with other features since they describe similar characteristics. For instance, Range, STD, IQR, MAD, and PDMeD represent the dispersion of the data while PDMod, MADSlOpe, and MADRes represent the data fluctuation as a function of time. Moon Phase and the measures of dispersion are moderately correlated because sky brightness increases as the moon with higher illumination (or Phase) approaches instrument's field of view. Consequently, sky brightness deviates from moonless level and the dispersion increases. Range and STD also correlate with P05 and P95 since those measures of dispersion are more sensitive to the tail of data distribution compared to IQR and MAD. However, some features may be more effective in classifying the data compared to the others.

Figure 4 maps the coefficient of determination (R^2) between two features. In machine learning, some features that highly correlate with other features can be excluded in order to reduce dimensionality and optimize the computation. The inclusion of such features may not increase the classification performance significantly. Based on the correlation analysis and the feature importance scoring, five features can be excluded for further training/prediction. They are Range, MAD, and PDMeD which describe the

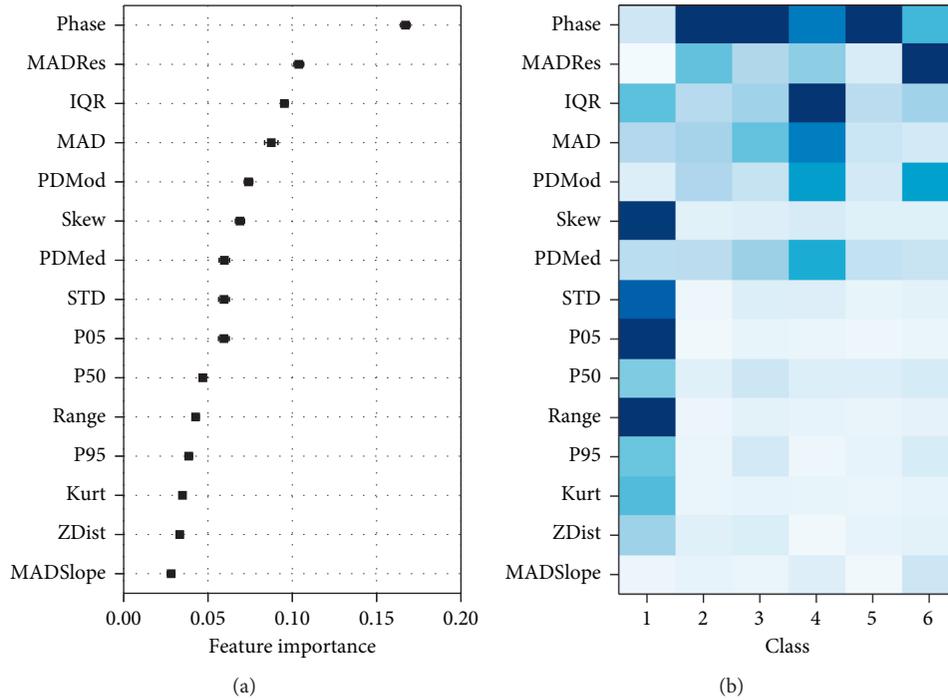


FIGURE 3: Feature importance produced by Random Forest Classifier for multiclass case (a) and binary classification case (b). In the right panel, features with higher importance are represented by darker colors.

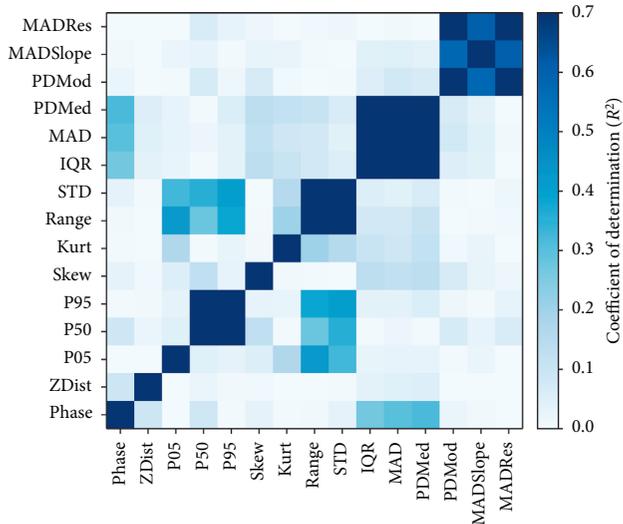


FIGURE 4: Correlation between features. Darker color represents a higher coefficient of determination (R^2) between two features.

dispersion of the data; PDMoD which describes fluctuation; and also P50 which highly correlates with P95. The dispersion of the data is sufficiently represented by IQR which gains 10% of importance while the fluctuation is described better by MADRes which is ranked second. P95 is preferred instead of P50 as P95 is better in detecting outliers or identifying class-1 data.

The evolution of misclassification error rate as the increase of maximum features utilized can be viewed in Figure 5. Classification based on single feature is approximately

50% accurate, and the accuracy increases abruptly by the introduction of some other important features. The average misclassification rates are below 20% when five or more features are used in training. Minimum errors of about 18% are achieved by using 15 features or 10 carefully selected features. However, training with 10 features can be performed 20% faster, even though the difference is not significant when dealing with a considerably small sample of size 10^3 .

3.3. Classification Performance. The overall performance of the classifier can be evaluated by implementing it on the testing dataset. Confusion matrix which summarizes the accuracy of the prediction compared to the expert classification is displayed in Figure 6. Using only 10 selected features, average accuracy of 92% was achieved with the minimum accuracy of 89% for identifying class-5 (cloudy moonlit sky) data and the highest accuracy of 98% for identifying class-1 (invalid) data. However, these high accuracy rates might be less informative.

As discussed in the literature [32, 33], accuracy is not a good metric to evaluate the performance of any classifier handling imbalanced data because it has low informativeness, produces less distinctive and less discriminable values, and also is in favor of dominant class. Alternatively, other metrics such as F -score and G -mean can be computed using the following equations:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

$$G = \sqrt{\text{precision} \times \text{recall}}$$

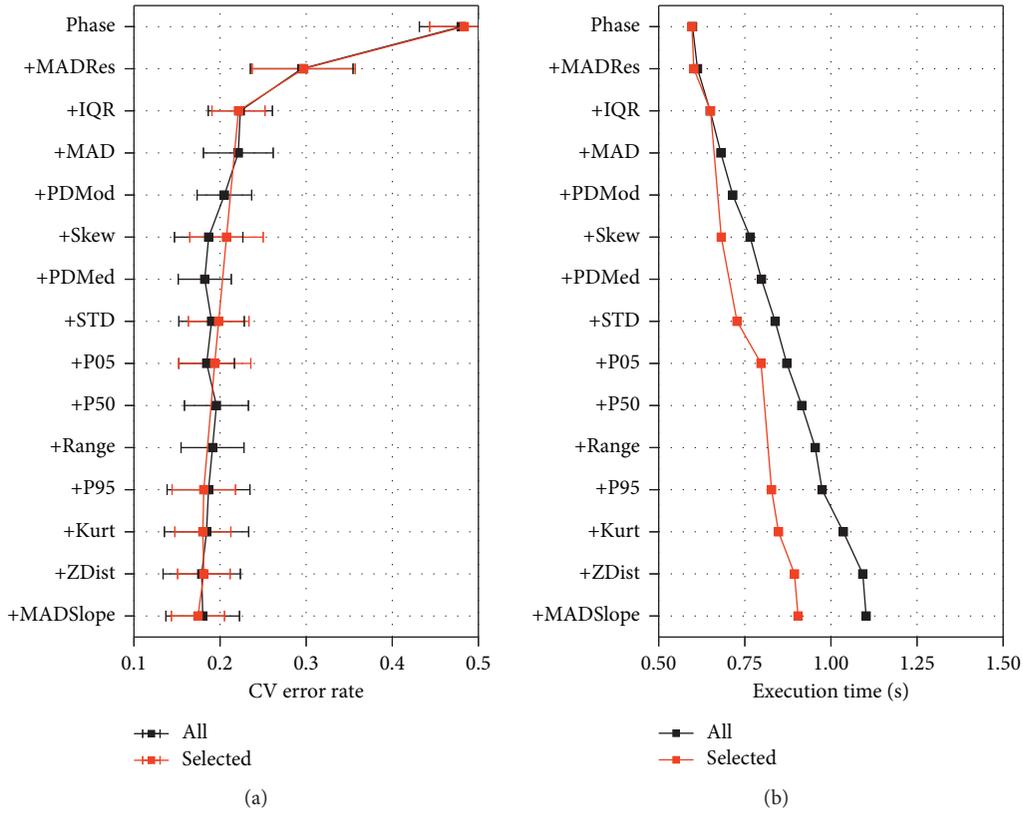


FIGURE 5: The evolution of cross-validation error rate as an additional feature is included in the classification (a). Execution time for each run is also displayed (b).

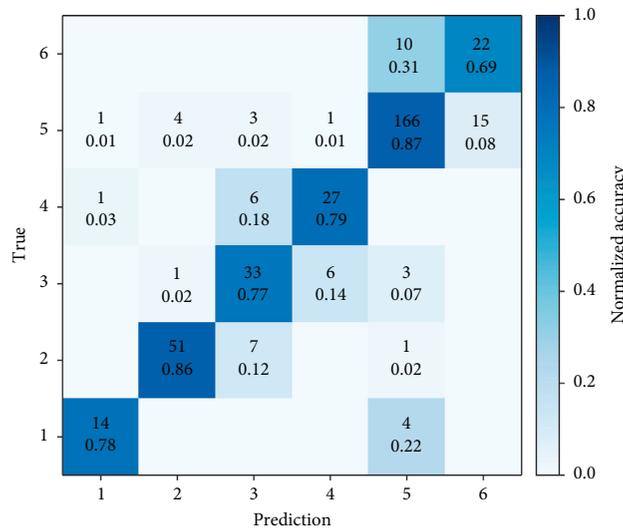


FIGURE 6: Confusion matrix for the Random Forest classification using 10 features from the training dataset. Each cell shows the number and the fraction of data points of a certain class which are assigned to a particular predicted class. Fractions in the diagonal are the recall scores.

where

$$\begin{aligned}
 \text{recall} &= \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \\
 \text{precision} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}}.
 \end{aligned}
 \tag{2}$$

F-score is basically the harmonic mean between recall and precision while the *G*-mean is the geometric mean between the two. Recall or true positive rate measures the ability of the classifier to find data with certain class while precision describes the fraction of correct classification in a retrieved class. In general, recall anticorrelates with precision such that adjustment in the classifier hyperparameters may

yield high recall while precision decreases such that proper averaging between the two scores is needed.

Table 3 summarizes performance scores calculated from the confusion matrix in Figure 6, but for binary classification in every class. The average values of the scores (weighted by the number of data points in each class) are considerably good while F -score and G -mean tend to be convergent. Nevertheless, misclassification occurs such that some scores are below average.

Some peculiar data of class-1 were misclassified as moonlit data since some features (e.g., moon phase, measure of dispersion, and fluctuation) resemble the characteristics of moonlit data. Significant portions of misclassification can also be found for class-3, class-4, class-5, and class-6 where the justification of cloudiness during the night is the key aspect for identifying data of those classes. However, those relatively low scores are not necessarily caused by a poor classification model. Instead, inconsistency of manual classification by experienced users also contributes to the scores achieved since they are calculated by assuming that manual classification is true. This argument may explain the misclassification observed during dichotomy between class-3 and class-4 and also between class-5 and class-6.

Aside from the aforementioned scores, the overall performance of the classifier can also be evaluated according to the posterior distribution of the data. Figure 7 shows posterior probability distribution functions (PDF) of sky brightness over eight selected stations after classification processes. Data of class-1 is not displayed in the chart as it is considerably meaningless.

In most cases, clear moonless sky brightness (class-4, green) shows a sharp distribution with full width at half maximum of about 1.5 mpsas, except for the data from Agam (AGM) where multimodal distribution is obvious. Examination of the AGM dataset reveals that this multimodal distribution is produced by several flat light curves with different levels categorized as class-4. The peaks of the distribution vary from ~ 16 mpsas for data acquired at light polluted urban regions to ~ 21 mpsas for data obtained at considerably dark rural areas. This implies that the classification model works well for sky brightness data with different degrees of light pollution. Additionally, the classifier is expected to perform equally well for the data acquired at high latitude regions where the duration of dark nights can be shorter and clear moonless sky does not necessarily produce flat light curve [12]. For data with such characteristics, the importance of IQR as the distinguishing feature is expected to be overtaken by other features that represent data fluctuation (PMod, MADRes) rather than dispersion.

Next, data of class-2 (orange) and class-3 (yellow) have broader and more left skewed distribution as cloud appearance amplifies the light pollution and brightens the night sky [16, 17]. Actually, darkening of the night sky was observed several times in rural station of Agam (AGM) and urban station of Bandung (BDG) during overcast conditions where thick clouds, fog, covered the sky or even precipitation occurred. At those moments, darkening started before twilight and the observed brightness reached 22.5 mpsas (in

TABLE 3: Performance scores achieved in classification of testing dataset. The average values are weighted by the number of data in each class.

Class	Accuracy	Precision	Recall	F -score	G -mean
1	0.98	0.87	0.78	0.82	0.82
2	0.97	0.91	0.86	0.89	0.88
3	0.93	0.67	0.77	0.72	0.72
4	0.96	0.79	0.79	0.79	0.79
5	0.89	0.92	0.87	0.89	0.89
6	0.93	0.59	0.69	0.64	0.64
Average	0.92	0.84	0.83	0.84	0.84

AGM) with less fluctuation compared to the data obtained during partially cloudy conditions. This pattern was also reported by Jechow et al. [16]. Presence of dew and water droplets in front of the sensor container also influences the measurements. BDG data show darkening during overcast conditions even though it was acquired in the urban area where brightening due to the clouds is expected. Because of this anomaly, data with indications of darkening was manually classified as peculiar (class-1) and omitted in further statistical analysis. For instance, Rezky et al. [34] developed an algorithm based on Hough transform to identify the actual clear sky brightness from SQM data which is served as density plot (or scotogram). This value (or function) is expected to be linear during the ideal night without any interference from clouds, moonlight, or Milky Way patches. The inclusion of peculiar or invalid data in this analysis is likely to obscure the linear pattern observed.

Based on Figure 7, data from Pasuruan (PSR) seem to be outlier because the posterior distributions of class-2 and class-3 have bumps on the right side (darker sky). On several nights, flat light curves suddenly jump to higher magnitude at around 3 am local time without any indication of clouds. The attenuation of sky brightness due to clouds in this region is highly unlikely since the location is situated close to the city and the brightening is expected during cloudy and overcast nights. On the other hand, darkening by smoke is plausible since the measurement was conducted at the location which is somewhat close to the industrial complex. Due to the bumps, the algorithm correctly identifies overcast conditions.

Data obtained during clear moonlit sky (class-6) have the widest distribution whose left tail reaches 12 mpsas in most cases. Data of class-5 (cloudy moonlit) has shorter tail because clouds attenuate scattered moonlight and make moonlit sky appear a bit darker. Considering the shape of the distribution, it is noteworthy that failure in discriminating data acquired in overcast conditions may lead to overestimation of night sky brightness up to 1 mpsas. In order to properly obtain a single representative value of local sky brightness, it is highly recommended that only data of class-4 is used. This careful selection is comparable to the significant magnitude which is defined as the average of the highest third of night sky brightness values recorded during ideal night ($m_{1/3}$) [35]. The ideal night in this regard is defined as the condition of astronomical darkness, when the sun is below -18° and the moon is below -5° . Recently, it is

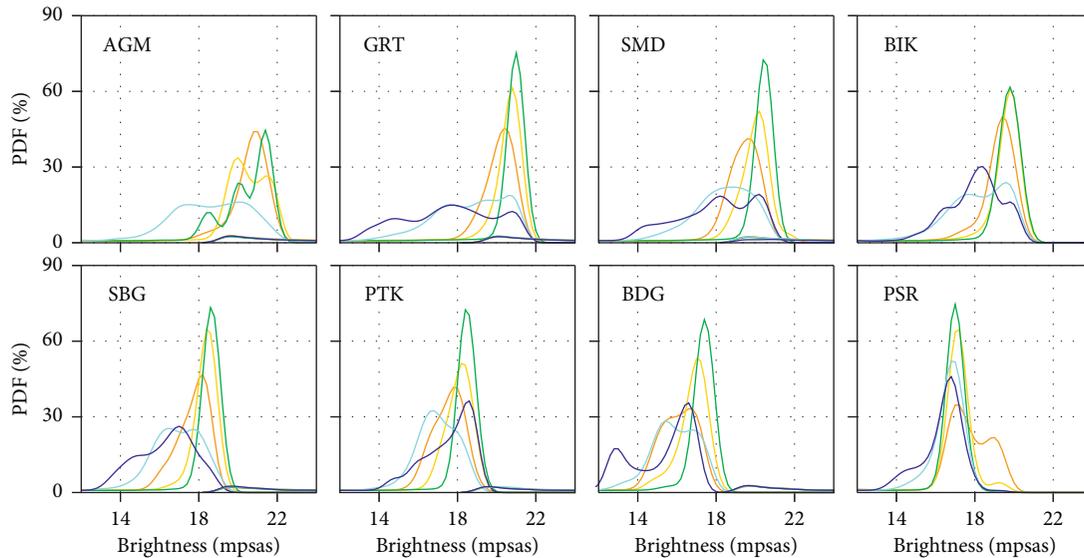


FIGURE 7: Distribution of sky brightness over different stations: constructed from categorized data. The color code is similar to that in Figure 1.

suggested that data within the full width at half maximum (m_{FWHM}), instead of the highest third, are used to calculate the significant magnitude for long term monitoring of sky brightness [36]. In short, these statistical selections are the other ways of selecting clear and moonless data that represent the true value of night sky brightness. The stability of sky brightness during the whole night becomes the main key to the selections both in the studies by Bará et al. [35, 36] or in the study reported here. The difference is that Bará et al. [35, 36] used statistical approaches to the density distribution constructed using long term data while the current study evaluates nightly characteristics to determine the class. For the case of data obtained in tropical region where the clear sky fraction is somewhat lower [37], there is a caution that the sharp distribution of the ideal data is likely to be overwhelmed (or smoothed) by data obtained during cloudy or overcast conditions.

In general, the classification model performed well in following the patterns of data it used in training. Beyond this point, it is normal to question whether the classification is also in agreement with the actual weather data. Additional data from instruments such as all-sky camera and weather station or even *ex-situ* observation will be required to confirm the weather conditions deduced from SQM data. Not all stations invoked in this study are equipped with operational weather stations so that further data compilation from various sources will be required to discuss this issue.

4. Conclusions

In this study, a Random Forest Classifier of 300 trees that invokes 10 features from nightly sky brightness data has been trained and tested. Observation data in the form of light curves need to be categorized into six different classes according to their qualities and observing conditions (peculiar data, overcast, cloudy, clear, moonlit-cloudy, and moonlit-clear). Initially, 15 features were extracted from the data, but some of the features were not selected for further classification as they highly correlate with other better

performing features. Among those features, the moon phase becomes the most important feature for classification followed by some measures of dispersion and fluctuation.

Evaluated using 375 testing data, the classifier achieves considerably high average accuracy (0.92), F -score (0.84), and G -mean (0.84). There is a slight confusion to discriminate data obtained in cloudy, overcast, and clear nights such that recall and precision scores achieved are below average. This confusion cannot be judged as pure misclassification because the foundation of the Random Forest Classifier is training dataset evaluated or labelled manually. This laborious task is not independent from error and inconsistency while machine learning technique serves better consistency. Apart from that, the classifier produces a good and reasonable set of posterior distribution functions where the distribution of class-4 (clear sky) data is the most concentrated than others. The distributions of class-2 (overcast) and -3 (cloudy) data are left skewed following the notion that the presence of clouds tends to amplify the light pollution, while data of class-5 (cloudy moonlit) and -6 (clear moonlit) have the largest dispersion/spread. However, further study using actual weather data (either *in situ* or *ex situ*) to examine the conformity of the classification model is interesting to do in the future.

Currently, open source software, such as PySQM developed by a group in Universidad Complutense Madrid, is available for SQM user community with the main functions of visualization and standardization of data produced using SQM. Evaluation and classification tasks can be regarded as fundamental and crucial tasks to do after those basic visualization and standardization processes.

Data Availability

The data used in this study can be provided by the corresponding author upon request. Codes related to this study are available through <https://github.com/rhorom/skybright>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

RP had a leading role in conceptualizing, conducting, and reporting the study. LM and RAP contributed to result analysis. AGA managed the whole project of sky brightness measurement through the National Institute of Aeronautics and Space.

Acknowledgments

This study was funded by the National Institute of Aeronautics and Space through an in-house program. LM was supported by Riset Pengembangan Kompetensi Institut Teknologi Bandung. The authors acknowledge the data acquisition team for maintaining the SQMs and for providing continuous sky brightness data.

References

- [1] S. L. Lewis and M. A. Maslin, "Defining the anthropocene," *Nature*, vol. 519, no. 7542, pp. 171–180, 2015.
- [2] T. Longcore and C. Rich, "Ecological light pollution," *Frontiers in Ecology and the Environment*, vol. 2, no. 4, pp. 191–198, 2004.
- [3] R. Chepesiuk, "Missing the dark: health effects of light pollution," *Environmental Health Perspectives*, vol. 117, no. 1, 2009.
- [4] R. M. Lunn, D. E. Blask, A. N. Coogan et al., "Health consequences of electric lighting practices in the modern world: a report on the National Toxicology Program's workshop on shift work at night, artificial light at night, and circadian disruption," *Science of the Total Environment*, vol. 607–608, pp. 1073–1084, 2017.
- [5] T. Gallaway, R. N. Olsen, and D. M. Mitchell, "The economics of global light pollution," *Ecological Economics*, vol. 69, no. 3, pp. 658–665, 2010.
- [6] T. Gallaway, "On light pollution, passive pleasures, and the instrumental value of beauty," *Journal of Economics Issue*, vol. 49, no. 1, pp. 71–88, 2010.
- [7] A. Hänel, T. Posch, S. J. Ribas et al., "Measuring night sky brightness: methods and challenges," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 205, pp. 278–290, 2018.
- [8] P. Cinzano, F. Falchi, and C. D. Elvidge, "The first World Atlas of the artificial night sky brightness," *Monthly Notices of the Royal Astronomical Society*, vol. 328, no. 3, pp. 689–707, 2001.
- [9] F. Falchi, P. Cinzano, D. Duriscoe et al., "The new World Atlas of artificial night sky brightness," *Science Advances*, vol. 2, Article ID e160037, 2016.
- [10] F. Falchi, "Campaign of sky brightness and extinction measurements using a portable CCD camera," *Monthly Notices of the Royal Astronomical Society*, vol. 412, no. 1, pp. 33–48, 2010.
- [11] A. Jechow, S. J. Ribas, R. C. Domingo, F. Hölker, Z. Kolláth, and C. C. M. Kyba, "Tracking the dynamics of skyglow with differential photometry using a digital camera with fisheye lens," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 209, pp. 212–223, 2018.
- [12] T. Posch, F. Binder, and J. Puschnig, "Systematic measurements of the night sky brightness at 26 locations in Eastern Austria," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 211, pp. 144–165, 2018.
- [13] J. Puschnig, A. Schwope, T. Posch, and R. Schwarz, "The night sky brightness at Potsdam-Babelsberg including overcast and moonlit conditions," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 139, pp. 76–81, 2014.
- [14] C. S. J. Pun, C. W. So, W. Y. Leung, and C. F. Wong, "Contributions of artificial lighting sources on light pollution in Hong Kong measured through a night sky brightness monitoring network," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 139, pp. 90–108, 2014.
- [15] C. C. M. Kyba, T. Ruhtz, J. Fischer, and F. Hölker, "Cloud coverage acts as an amplifier for ecological light pollution in urban ecosystems," *PLoS One*, vol. 6, Article ID e17307, 2011.
- [16] A. Jechow, F. Hölker, Z. Kolláth, M. O. Gessner, and C. C. M. Kyba, "Evaluating the summer night sky brightness at a research field site on Lake Stechlin in northeastern Germany," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 181, pp. 24–32, 2016.
- [17] A. Jechow, Z. Kolláth, S. J. Ribas, H. Spoelstra, F. Hölker, and C. C. M. Kyba, "Imaging and mapping the impact of clouds on skyglow with all-sky photometry," *Scientific Reports*, vol. 7, no. 1, Article ID 6741, 2017.
- [18] J. W. Richards, D. L. Starr, N. R. Butler et al., "On machine-learned classification of variable stars with sparse and noisy time-series data," *The Astrophysical Journal*, vol. 733, p. 10, 2011.
- [19] F. Elorrieta, S. Eyheramendy, A. Jordán et al., "A machine learned classifier for RR Lyrae in the VVV survey," *Astronomy & Astrophysics*, vol. 595, p. A82, 2016.
- [20] P. Cinzano, "Night sky photometry with sky quality meter," *ISTIL Internal Report*, vol. 9, 2005.
- [21] C. Fruck, M. Gaug, J.-P. Ernenwein et al., "Instrumentation for comparing night sky quality and atmospheric conditions of CTA site candidates," *Journal of Instrumentation*, vol. 10, no. 4, Article ID P04012, 2015.
- [22] T. W. Davies, J. Bennie, R. Inger, and K. J. Gaston, "Artificial light alters natural regimes of night-time sky brightness," *Scientific Reports*, vol. 3, no. 1, p. 1722, 2013.
- [23] K. Krisciunas and B. E. Schaefer, "A model of the brightness of moonlight," *Publications of the Astronomical Society of the Pacific*, vol. 103, pp. 1033–1039, 1991.
- [24] P. Dubath, L. Rimoldini, M. Süveges et al., "Random forest automated supervised classification of Hipparcos periodic variable stars," *Monthly Notices of the Royal Astronomical Society*, vol. 414, no. 3, pp. 2602–2617, 2011.
- [25] M. C. Kind and R. J. Brunner, "TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests," *Monthly Notices of the Royal Astronomical Society*, vol. 432, pp. 1483–1501, 2013.
- [26] J. Zhang, G. Lin, S. Zeng et al., "Intelligent recognition of time stamp characters in solar scanned images from film," *Advances in Astronomy*, vol. 2019, Article ID 6565379, 9 pages, 2019.
- [27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [28] D. Baron, "Machine learning in astronomy: a practical overview," 2019, <https://arxiv.org/abs/1904.07248>.
- [29] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [30] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning,"

- Foundations and Trends in Computer Graphics and Vision*, vol. 7, pp. 81–227, 2011.
- [31] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [32] Y. Sun, A. K. C. Wong, and M. S. Kamel, “Classification of imbalanced data: a review,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.
- [33] J. Hernandez-Orallo, P. Flach, and C. Ferri, “A unified view of performance metrics: translating threshold choice into expected classification loss,” *Journal of Machine Learning Research*, vol. 13, pp. 2813–2869, 2012.
- [34] M. Rezky, R. Priyatikanto, A. G. Admiranto, and E. Soegiartini, “Pattern recognition as tools for understanding night sky brightness variabilities,” *Journal of Physics Conference Series*, vol. 1231, Article ID 012001, 2019.
- [35] S. Bará, “Anthropogenic disruption of the night sky darkness in urban and rural areas,” *Royal Society of Open Science*, vol. 3, Article ID 160541, 2016.
- [36] S. Bará, R. C. Lima, and J. Zamorano, “Monitoring long-term trends in the anthropogenic night sky brightness,” *Sustainability*, vol. 11, no. 11, p. 3070, 2019.
- [37] T. Hidayat, P. Mahasena, B. Dermawan, T. W. Hadi, P. W. Premadi, and D. Herdiwijaya, “Clear sky fraction above Indonesia: an analysis for astronomical site selection,” *Monthly Notices of the Royal Astronomical Society*, vol. 427, pp. 1903–1917, 2012.