第 48 卷 第 10 期
2021 年 10 月

湖南大学学报（自然科学版）
Journal of Hunan University（Natural Sciences）

Vol. 48 No. 10
October 2021

Open Access Article

Lithological Boundaries Identification in Dense Vegetation Area Based on Satellite Data Using Rare Training Data

Hary Nugroho[1,2*], Ketut Wikantika[2], Satria Bijaksana[3], Asep Saepuloh[3]

[1] Geodetic Engineering Study Program, Institut Teknologi Nasional, Bandung, Indonesia

[2] Department of Geodesy and Geomatics Engineering, Institut Teknologi Bandung, Indonesia

[3] Department of Geophysical Engineering, Institut Teknologi Bandung, Indonesia

Abstract: One of the most critical geological maps is the boundary of rock types or lithology. Machine learning algorithm such as Random Forest (RF) is a useful classification method for producing lithology predictions. In the lithology mapping that was carried out for the first time in difficult areas to access, the problem faced was the collection of training data. It often happens that the amount of training data that can be collected is very limited, especially if the location is an area with high vegetation density. This study aims to assess the performance of the RF algorithm with hyperparameter tuning in identifying lithological boundaries in dense vegetation areas by using remote sensing data with rare training data. We conducted experiments that simulated remote predictive mapping (RPM) using an RF algorithm using satellite data to obtain a predictive lithological map of Komopa located in Paniai District, Papua Province, Indonesia. This study area has dense vegetation and thick soil layers. We used remote sensing data consisting of Sentinel 2A, ALOS PALSAR and DEM, and 1000 drill log points. The results of nine representative models indicated that the test accuracy of lithological classification was moderate (0.53-0.75), but low values on recall (0.24-0.59), precision (0.24-0.51), and F1 score (0.24-0.39). Meanwhile, the training accuracy achieved by each model was very high (0.92-1.0). Model 9, which only used 50 balanced training points, gives the best classification result. Although its test accuracy and F1-score were relatively low, the resulting lithological boundaries are closest to the existing lithological map. This study shows that the RF classification using balanced training data can provide good classification results in the predictive lithological mapping, even though the number is small.

Keywords: lithological map, remote predictive mapping, machine learning, random forest, remote sensing.

基于卫星数据的密集植被区岩性边界识别利用稀有训练数据

摘要: 地质图中最关键的信息之一是岩石类型或岩性的边界。随机森林等机器学习算法是一种用于进行岩性预测的有用分类方法。在首次在难以进入的地区进行岩性测绘时，面临的问题是训练数据的收集。经常会发生可以收集的训练数据量非常有限的情况，特别是如果该位置是植被密度高的区域。本研究旨在通过使用具有稀有训练数据的遥感数据，评估具有超参数调整的算法在识别密集植被区岩性边界方面的性能。我们通过使用卫星数据进行了使用射频算法模拟远程预测制图的实验，以获得位于印度尼西亚巴布亚省区的预测岩性图。本研究区植被茂密，土层较厚。我们使用了由和以及 1000 个钻孔测井点组成的遥感数据。 9 个代表性模型的结果表明，岩性分类的测试精度中等（0.53-0.75），但召回率（0.24-0.59）、精度（0.24-0.51）和 F1 得分（0.24-0.39）较低。同时，每个模型达到的训练精度都非常高（0.92-1.0）。仅使用 50 个平衡训练点的模型 9 被认为给出了最好的分类结果。虽然其测

Received: June 29, 2021 / Revised: August 27, 2021 / Accepted: September 21, 2021 / Published: October 30, 2021

About the authors: Hary Nugroho, Geodetic Engineering Study Program, Institut Teknologi Nasional, Bandung, Indonesia; Department of Geodesy and Geomatics Engineering, Institut Teknologi Bandung, Indonesia; Ketut Wikantika, Department of Geodesy and Geomatics Engineering, Institut Teknologi Bandung, Indonesia; Satria Bijaksana, Asep Saepuloh, Department of Geophysical Engineering, Institut Teknologi Bandung, Indonesia

试精度和相对较低，但得到的岩性边界与现有的岩性图最接近。本研究表明，使用平衡训练数据的分类，即使数量很少，也可以提供相当好的分类结果，可用于预测岩性成图。

关键词： 岩性图、远程预测映射、机器学习、随机森林、遥感。

1. Introduction

RPM is an integrated methodology in which the geological features on the geological map can be recompiled based on interpreting various types of geoscience data, both from remote sensing and field data, resulting from accompanying fieldwork [1]. Generally, RPM is applied to areas that have not been mapped before. The features obtained can be used as preliminary data before geologists conduct geological surveys directly in the field.

The techniques utilized in data processing are Machine Learning Algorithms (MLAs) [2, 3]. The inductive MLAs, such as RF, are carried out through pattern recognition and empirical data generalization. This machine-learning algorithm provides a practical way to predict lithology or other spatially varying physical features.

MLAs implementation requires training data. In areas with a very small amount of training data that can be collected, the prediction of lithological boundaries is a challenging issue. Therefore, the application of RF in such locations is essential to be studied.

1.1. Machine Learning Random Forest Algorithm

RF is a supervised machine learning method that uses training points to "learn" and constructs a model for classification [4]. There must be sufficient training points for the proper classification and well distributed to represent the lithology classes to be classified sufficiently. However, high accuracy is difficult to achieve due to the amount of training point data taken [5, 6] and the sampled surface condition, for example, dense vegetation cover of tropical forest. Density and type of plants will affect spectrally up to 50% [7]. Therefore, it is necessary to suppress the signal from vegetation to reduce its influence on data inaccuracy.

The distribution of training points often causes imbalanced data. This condition is because we do not know the actual lithology conditions in the field. As a result, lithology types with a large or dominant area will dominate the training points collected. The imbalanced training points in lithology classes can create difficulties for machine learning algorithms in the classification process because these algorithms always assume the distribution of training points in each category is relatively balanced [8, 9].

RF algorithm [10] is a supervised ensemble classification algorithm and an extension of the decision tree method [11]. The parameters that determine the number and size of the decision tree are called hyperparameters. Hyperparameters are like the settings of an algorithm that can be adjusted to improve classification performance [12]. However, hyperparameter tuning relies more on experimental results than theory. Thus, the best method to determine the optimal settings is to try many different combinations to evaluate each model's performance [13].

In hyperparameter tuning, the tuning is generally performed on several hyperparameters, such as the number of decision trees. The criteria for achieving a split at the node (for classification, it used gini or entropy) maximum depth of the individual trees, minimum sample to split on at an internal node, the maximum number of leaf nodes, number of random features, and the size of the bootstrap dataset [14]. This study performed hyperparameter tuning on the number of decision trees and the minimum sample split. Many decision trees will make a robust aggregate model with fewer variants and a longer training/computation time [15]. The minimum sample split is a value that affects the decision tree size.

1.2. Objectives

The objectives of this study were: (i) to assess the performance of RF algorithm in identifying lithological boundaries in dense vegetation areas by using remote sensing data with rare or very limited training data, (ii) to examine the effect of hyperparameters tuning and sample size and distribution of imbalanced training points on the accuracy of classification results, and (iii) to find out the differences between the lithological boundaries generated by the RF classification and the existing lithological map.

We used remote sensing data consisting of Sentinel 2A, ALOS PALSAR and DEM, and 1000 drill log points. The trials were conducted by using training data points with varying amounts and distributions. Variations in the distribution of training points were obtained from two kinds of sampling techniques, i.e., simple random sampling and stratified random sampling. The tuning of hyperparameters included the number of decision trees and minimum sample split. RF performance was evaluated by train accuracy, test accuracy, precision, recall, and F1-score. Subsequently, the lithological maps generated from several models were visually compared with Mine Serve International's existing lithological map produced in 2000.

1.3. Geology Setting

The study area is Komopa, located in Paniai District, Papua Province, Indonesia (Figure 1). This area is located along the northern margin of deformed platform rocks, comprising the Early Tertiary New Guinea Limestone Group and Jurassic to Late Cretaceous Kembelangan Group. As elsewhere within the Irian Jaya Mobile Belt, these cover rocks are characterized by south-to-southwesterly verging thrusts and folds that, to a large measure, control the present distribution of the Kembelangan and New Guinea Limestone groups. Moreover, the Komopa-Dawagu area lies along strike from the Grasberg-Ertzberg District concerning these structures.

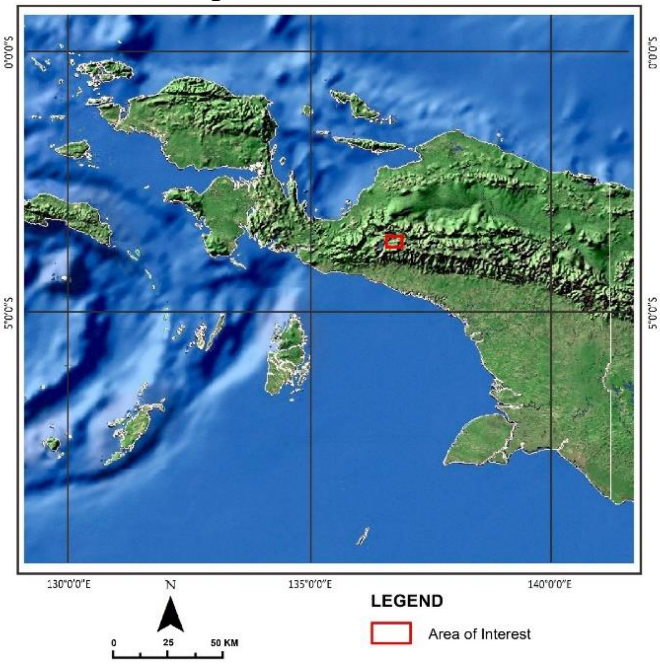

Fig. 1 Area of interest in the middle of Papua Province, Indonesia

These structures parallel the Derewo Metamorphic Belt's southern boundary, marked by the Derewo Fault Zone. They may reflect deep basement faults parallel to the Australian continental plate's attenuated margin formed during the Triassic break-up of Gondwanaland. In detail, they are reflected by the east-west strike of the Komopa and Dawagu porphyries. Alkalic, monzodioritic to dioritic igneous complexes with which the known porphyry copper-gold zones at Komopa and Dawagu are probably younger than the north-south crustal shortening [16].

The Komopa region is covered by dense vegetation, and thick soil layers (humus) (Figure 2a) cover the ground. The existing lithological map [17] is shown in Figure 2b, produced by Mine Serve International in 2000. The types of lithology found in the study area are quaternary alluvium, pseudo gossan, inferred porphyry, sedimentary rocks, and undifferentiated porphyry. Pseudo gossan is a more detailed lithology class compared to others. Nevertheless, Mine Serve International used this class as guidance for their mineral exploration.

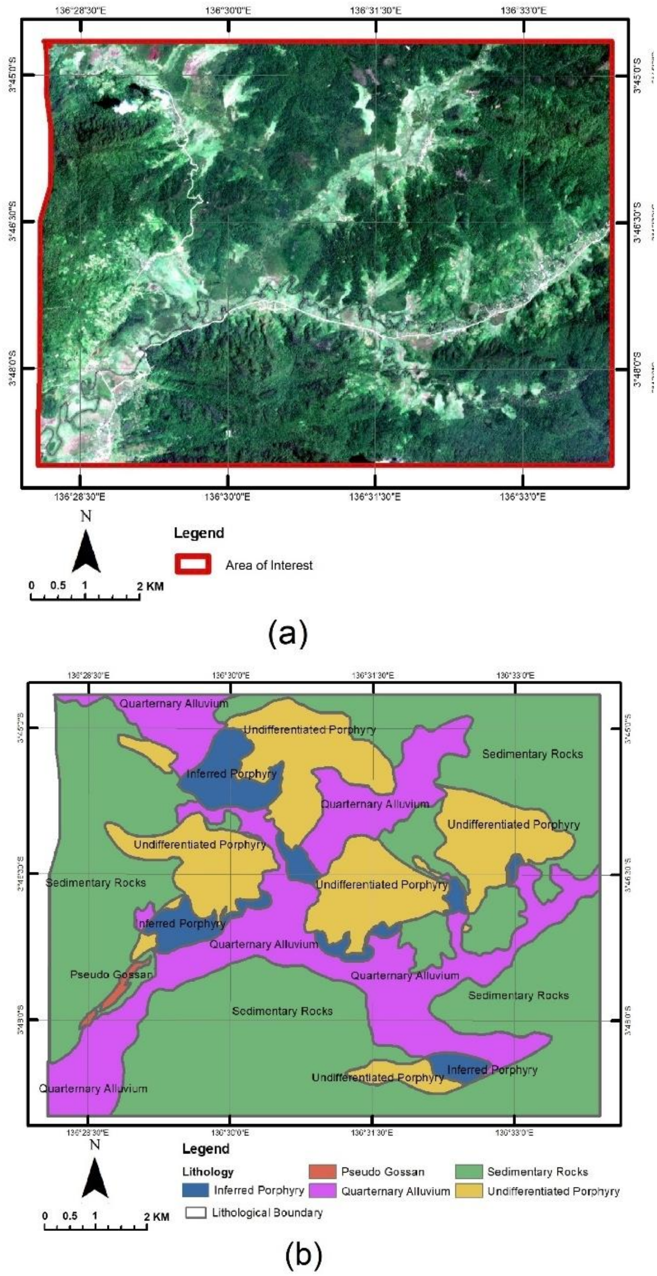

Fig. 2 (a) Study area, Komopa, Paniai District, Papua Province presented by the color composite of Sentinel 2A R, G, B=Band 4,3,2; (b) Local lithological map of Komopa scale 1:25.000, which Mine Serve International released in the year 2000, shows the variation of rock types in different colors [17]

2. Data and Methods

2.1. Data Compilation

Sentinel-2A image was acquired on January 9, 2019. This study only used the Sentinel-2A bands with a spatial resolution of 10 m and 20 m because the 60m resolution containing all components of the 20m product was resampled to 60 m [18]. Image code is the granule at Level-2A acquired from the USGS Earth Explorer [27]: S2A_MSIL2A_20190109T011721_N0211_R088_T53MPR_20190109T032340. Before being used, vegetation's influence on this image data has been removed using the Vegetation Suppression tool in the ENVI software. This process aimed to eliminate spectral vegetation signatures from multispectral images using the red and NIR bands [19].

The radar data used in this study area was ALOS-PALSAR Fine Beam Double polarization data (FBD), L-band, 3.17 m x 14.9 m (azimuth x range) resolution dual-polarization (HH+HV). The data was acquired on July 2, 2007, with Single Look Complex (SLC) data type by ALOS and preprocessed (orthorectification, slope correction, and mosaicked) by JAXA-EORC (Japan Aerospace Exploration Agency – Earth Observation Research Center). SAR calibration provides imagery in which the pixel values can be directly related to the radar backscatter of the scene [20]. The granule was acquired from [28] (Image code ALPSRP076507100-L1.0). The data was obtained in the form of digital number intensity. This data was then processed into HH-HV polarization backscatter [21].

L-band SAR can penetrate through dense vegetation, such as forest canopies [22]. Penetration depth tends to be longer with longer wavelengths (Figure 3). DEM data used in this study is DEMNAS (National DEM) from the Geospatial Information Agency, Republic of Indonesia. The National DEM is built from several data sources, including IFSAR (5m resolution), TERRASAR-X (5 m resolution), and ALOS PALSAR (11.25m resolution) data, by adding the stereo-plotting mass point data. The spatial resolution of DEMNAS is 0.27-arcsecond (8.1 m), using the EGM2008 vertical datum [23].

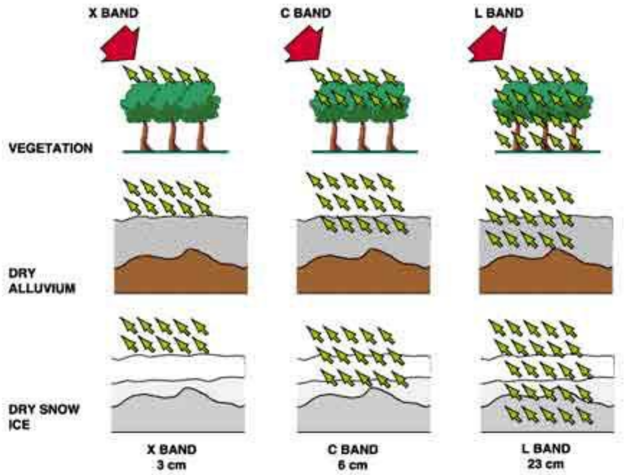

Fig. 3 SAR Data Penetration showed L-band microwave can reach the ground partially penetrating through vegetation to obtain information of vegetation and ground surface [21]

2.2. Training and Testing Samples

Training point is a point that connects data feature, and the object sought. The distribution pattern of training points can determine whether classification results are of good quality [24]. There were 1000 drill log points in this study, where 500 points were used as training points and 500 as testing points. PT provided all drill log data.

This study has two types of training point distribution patterns: (i) simple random sampling with sample sizes varying from 25, 50, 100, 200, 300, 400, and 500; and (ii) stratified random sampling with sample size variation of 25 and 50.

Data samples using simple random sampling are imbalanced data, and the distribution of sample points depends on the progress of drill log points acquisition in the field (Table 1). Therefore, not all lithology types were represented at sample sizes of 25 and 50 with simple random distribution. Meanwhile, the data sample using stratified random sampling is balanced data. This study only made two models with a stratified random sampling distribution, i.e., 25 and 50, because the number of samples in the minority lithology type is sufficient for these two data models. In the data processing, all data layers were resampled to a 20 m grid size.

2.3. Methods

The research methodology used in this study is illustrated in Figure 4. The lithological classification was then performed by applying an RF algorithm using the RF module from the Scikit-learn library [14], with varying sample sizes and tuning some hyperparameters. Scikit-learn is a Python module integrating many machine learning algorithms for medium-scale supervised and unsupervised problems [25, 26]. The RF application's hyperparameter values include the number of decision trees (n_estimator) and the number of minimum sample split (min samples split). The number of decision trees starts from 25, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000, while the minimum number of split samples starts from 5, 10, 15, 20, 25 30. With variations in the sample size and hyperparameter values, this experiment resulted in 420 classification results.

Table 1 Number and distribution of training points

No.	Distribution Type and Sample Size	Training Points Distribution in Each Rock Type					Total Number
		Inferred Porphyry	Pseudo Gossan	Quaternary Alluvium	Sedimentary Rocks	Undifferentiated Porphyry	
1	Simple Random 25 points	1	0	0	6	18	25
2	Simple Random 50 points	8	0	1	13	28	50
3	Simple Random 100 points	10	1	4	24	61	100
4	Simple Random 200 points	7	2	8	42	141	200
5	Simple Random 300 points	8	5	12	70	205	300
6	Simple Random 400 points	10	6	14	98	272	400

7	Simple Random 500 points	14	8	18	123	337	500
8	Stratified Random 25 points	5	5	5	5	5	25
9	Stratified Random 50 points	10	10	10	10	10	50
		Testing Points Distribution					
		8	9	17	124	342	500

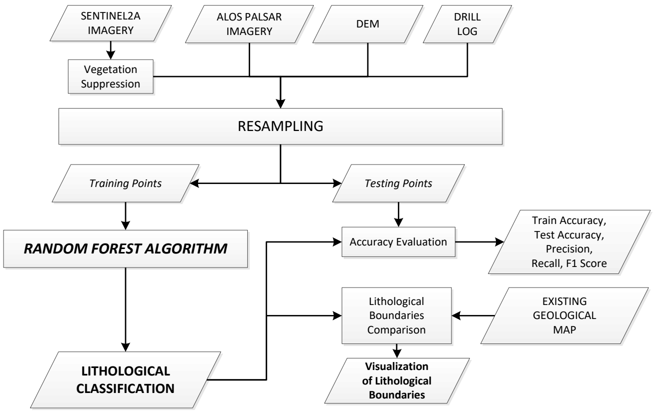

Fig. 4 Research methodology showed the workflows to obtain a lithological map

2.4. Accuracy Evaluation

The performance of RF classification results was evaluated using a confusion matrix based on precision, recall, F1-score, train accuracy, and test accuracy. We used testing points from drill log data to calculate those metrics. Precision quantifies the number of positive class predictions that belong to the positive class. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset. If being used alone, neither precision nor recall tells the whole story [9]. F1-score provides the harmonic mean of precision and recall. Besides, we also calculated the training accuracy using training points as cross-validation in the RF classification. Apart from doing a quantitative evaluation, we also compared the distribution of lithology types resulting from RF classification with the lithology boundaries on the existing geological map from Mine Serve International.

3. Results and Discussion

3.1. Effect of the Number and Distribution of Training Points on Classification Accuracy

This analysis took several examples in simple random and stratified random distributions, with the decision tree (DT) values being 500 and 1,000 and the minimum sample splits (MSS) value being 5. The performance on the other numbers of the decision tree was similar.

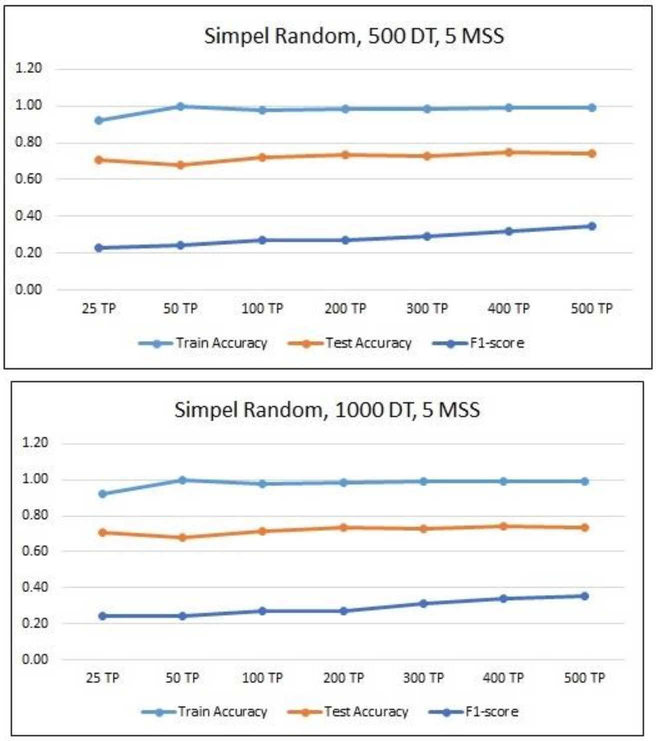

Fig. 5 The effect of the number of training points on F1 score, train accuracy and test accuracy for simple random distribution
Notes: TP - training points, DT - decision tree, MSS - minimum sample split

Table 2 The effect of the number of training points on F1 score, train accuracy and test accuracy for stratified random distribution

	Model			
	500 DT, 5 MSS		**1000 DT, 5 MSS**	
Training Points	25	50	25	50
Train Accuracy	1	0.98	1	0.98
Test Accuracy	0.58	0.52	0.57	0.51
F1 Score	0.3	0.31	0.29	0.31

Notes: DT - decision tree, MSS - minimum sample split

Classification results with a minimum sample split of five were the best results of the overall minimum sample split variation. As shown in Figure 5 and Table 2, as a whole, it can be seen that increasing the number of training points can increase the value of train accuracy, test accuracy, and F1 Score, except for the number of training points of 50, both on simple random and stratified random distributions. Meanwhile, training points 100 provided interesting classification results because the accuracy was generally better than training points 25 and 50 but not significantly different from training points 500. Based on 500 decision trees and five minimum sample splits, it can be seen that the optimal number of training points with a simple random distribution is 100.

3.2. Effect of the Number of Decision Tree on Classification Accuracy

Figure 6 shows an example of the effect of the number of decision trees on train accuracy, test accuracy, and F1 Score for simple random and stratified random distributions, taking 50 training points as an example. The other number of training points has a similar performance. The simple random distribution, increasing the decision tree number from 25 to 200, will increase train accuracy (2-6%). The maximum value of the training accuracy of 1.00 was reached at the 200-decision tree and stayed consistent. The test accuracy is relatively unaffected by the number of decision trees ranging from 0.67-0.69. A different condition occurs in the F1 Score, where increasing the number of decision trees by more than 100 decreases the F1 Score. In the stratified random distribution, in general, the number of decision trees has almost no effect on test accuracy, test accuracy, and F1 Score.

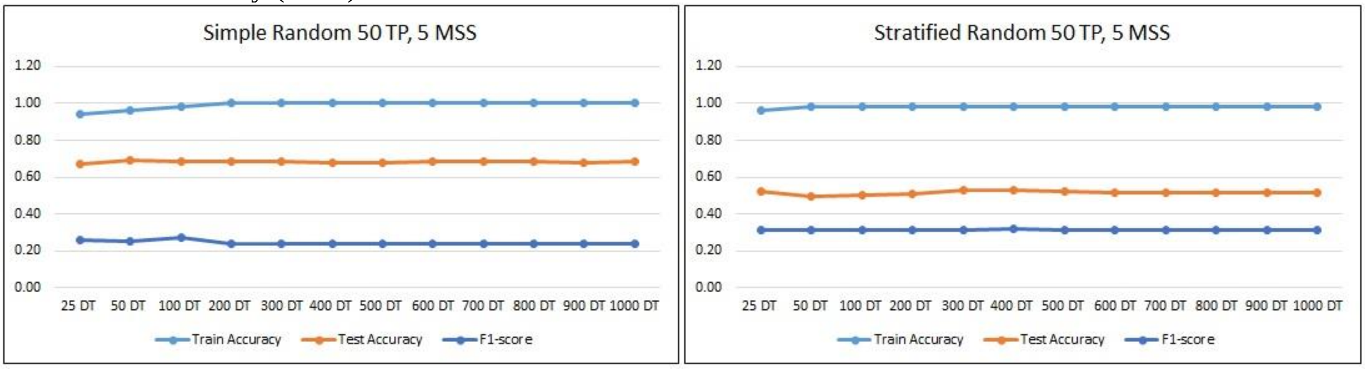

Fig. 6 The effect of the number of decision trees on the train accuracy, test accuracy, and F1 score for simple random and stratified random sampling distributions

Notes: TP - training points, DT - decision tree, MSS - minimum sample split

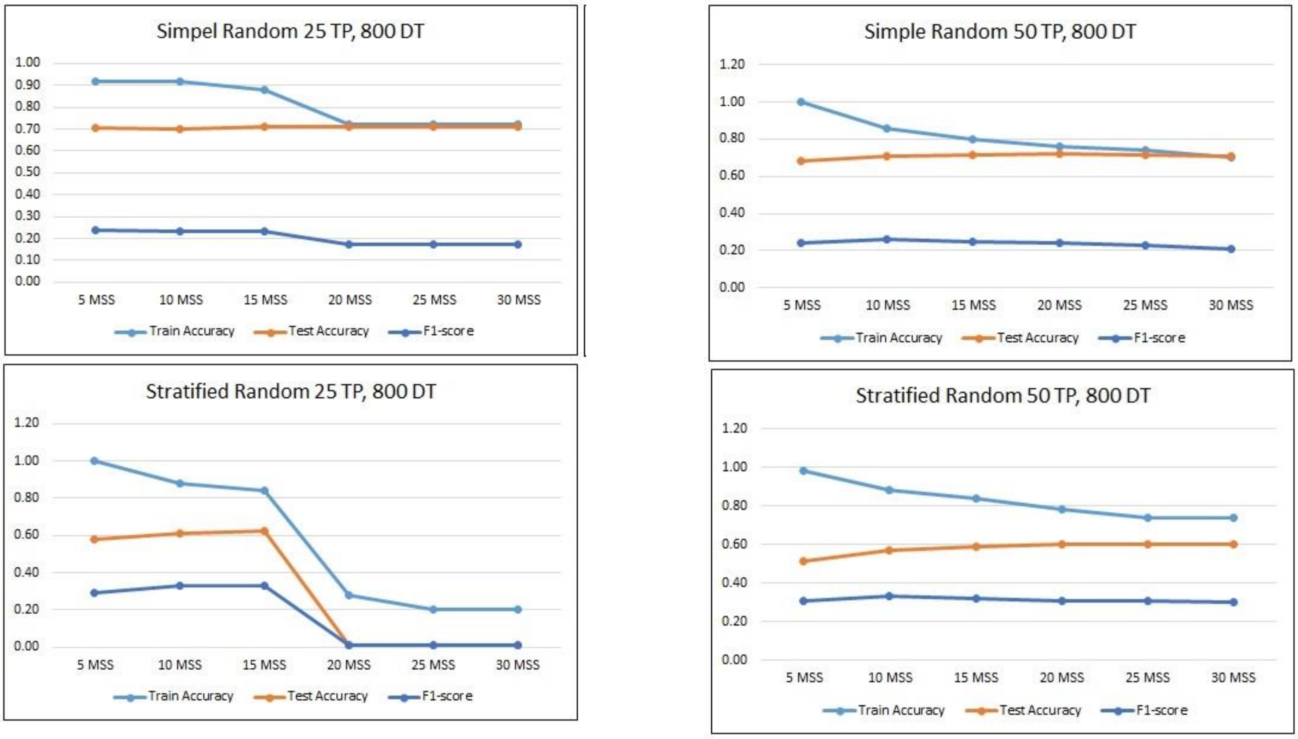

Fig. 7 Effect of the number of minimum sample splits on train accuracy, test accuracy, and F1 score

Notes: TP - training points, DT - decision tree, MSS - minimum sample split

3.3. Effect of the Number of Minimum Sample Splits on Classification Accuracy

The effect of the number of minimum sample splits on train accuracy, test accuracy, and F1 Score is shown by taking 25 and 50 training points as examples, both for simple random and stratified random distributions, with 800 decision trees (Figure 7). Performance of 100 to 500 training points on simple random distribution is similar to 50 training points.

The minimum sample splits significantly affect train accuracy, but the effect is not significant on test accuracy and F1 Score, except for the model with 25 training points with a stratified distribution. The best train accuracy is obtained at several minimum sample splits 5, after which from 10 to 30 it decreases. In fact, on the stratified random distribution of 25 training points, the minimum sample splits number ranging from 20 to 30 result in very low train accuracy, test accuracy, and F1 Score. It is suspected that this happened due to the small number of training points, but the number of samples at the node must be large enough so that the decision tree does not develop properly and results in poor accuracy. This condition does not occur in the stratified random distribution with 50 training points. The graphic shape shown by the simple random distribution with 50 to 500 training points gave almost the same results. Based on the above results, the optimal number of minimum sample splits is in the range of 5-10.

3.4. Confusion Matrix

To analyze classification results further using the confusion matrix, we selected nine models as representatives from each category of the number and distribution of training points, as shown in Table 3. The model that is considered the best in each category was determined gradually, firstly based on the value of train accuracy, secondly based on the value of the F1 Score, and thirdly based on the least number of the decision tree. The value of the minimum sample split is not considered in choosing a model because this value is related to the decision tree structure, and its assessment is done using the value of test accuracy value and F1 Score.

Based on Table 3, RF has difficulty classifying since all models show an overfitting classification result. This condition is indicated by the train accuracy value, which is much higher than the test accuracy, with the difference in accuracy reaching 21-51%. The model is very suitable for the training dataset with high train accuracy on all models (91-100%), the model is very suitable for the training dataset, but the model does not work well when using different datasets. Adding a large number of training points to the simple random distribution (100-500) can increase test accuracy by up to 4%, with a maximum test accuracy of 75% on models with 400 and 500 training points. Meanwhile, the effect on increasing precision, recall, and F1 Score was a maximum of 27%, 11%, and 15%, respectively.

Table 3 Values of accuracy, precision, recall, and F1 score on the representative models

Model	Simple Random Training Points	Decision Trees	Minimum Sample Split	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
1	25	100	5	0.92	0.71	0.24	0.24	0.24
2	50	200	5	1.00	0.68	0.25	0.24	0.24
3	100	100	5	0.99	0.72	0.35	0.28	0.30
4	200	200	5	0.99	0.74	0.35	0.26	0.27
5	300	50	5	0.99	0.72	0.36	0.30	0.32
6	400	300	5	1.00	0.75	0.43	0.30	0.33
7	500	200	5	0.99	0.75	0.44	0.33	0.36
	Stratified Random Training Points							
8	25	100	5	1.00	0.56	0.30	0.59	0.31
9	50	400	5	0.98	0.53	0.31	0.59	0.32

Although the overall accuracy based on testing points shows a moderate value (0.53-0.75), the recall is generally low (0.24-0.59). The precision (0.24-0.51) and the F1 scores indicate similar values (0.24-0.39). This data suggests that the built models have poor classification abilities. Models with a stratified random distribution give exciting results. Models 8 and 9 are models with balanced sample data. Compared with the simple random distribution, although it only uses 25 and 50 training points, the F1 Score for both models has the same result as model 3,4,5,6, which operates 100-400 training points. Even though the test accuracy is the worst (0.49-0.58), these two models give the highest recall value (0.54-0.59). This result indicates that the distribution of the sample data that is balanced affects the classification ability of the model.

Based on the overall values shown in Table 3, the models that give optimal results are models 1 and 3, in terms of train accuracy, test accuracy, precision, recall, and F1 Score. These models have good train accuracy (0.92-1.00) and moderate test accuracy (0.71-0.73), although the training points used are only 25 and 100, respectively.

3.5. Lithological Boundaries Comparison

Figure 8 visualizes the RF classification results obtained by each model. The classification models 1 to 9 were overlaid with lithological maps from Mine Serve International to see the differences in lithological boundaries formed. For example, when viewed in the Undifferentiated Porphyry class, the lithological boundaries for this class can be seen clearly and quite solidly, whereas compared to the satellite image of the Komopa region in Fig. 3, it can be seen that this area is an area covered with dense vegetation. Therefore, it can be said that the application of the vegetation suppression technique on the Sentinel satellite imagery combined with ALOS PALSAR and DEM has been quite successful in suppressing the effect of vegetation cover on images in this region. As for the Sedimentary Rocks class, the lithological boundaries are not very clear and not solid.

As shown in Figure 8, all models have described the lithological boundaries quite firmly at several locations. Thus, in general, each model can provide a good depiction of the lithological boundaries. These RF classification results also show that some minor class is almost undetectable, for example, Pseudo Gossan in Model 3 (Figure 8c) and 4 (Figure 8d). However, even the lithology class with a large enough area does not always provide solid classification results exemplified by the Quaternary Alluvium class. This class looks quite solid only in models 3 (Figure 8c), 5 (Figure 8e), 6 (Figure 8f), 7 (Figure 8g), 8 (Figure 8h), and 9 (Figure 8i). On the other hand, models 8 and 9 have balanced training points with a stratified random distribution (Figure 8h and 8i respectively), have reasonably solid classification results, and lithological boundaries are visible and fairly evenly distributed across all classes. The two models can well identify a minor class like Pseudo Gossan.

Among the nine models above, model 9 gives the best classification result. Although the test accuracy and F1-score results are relatively low (0.53 and 0.32, respectively), the predictive lithological map resulting from model 9 produced the lithological boundaries closest to the lithological map used as the reference in this study. It should be noted that this model only uses 50 training points with a stratified random distribution (balance data). This number is very small compared to other models, which reach 100-500 training points. Nevertheless, this study shows that the RF classification using balanced sample data can provide good classification results to operate in RPM even though the number is small.

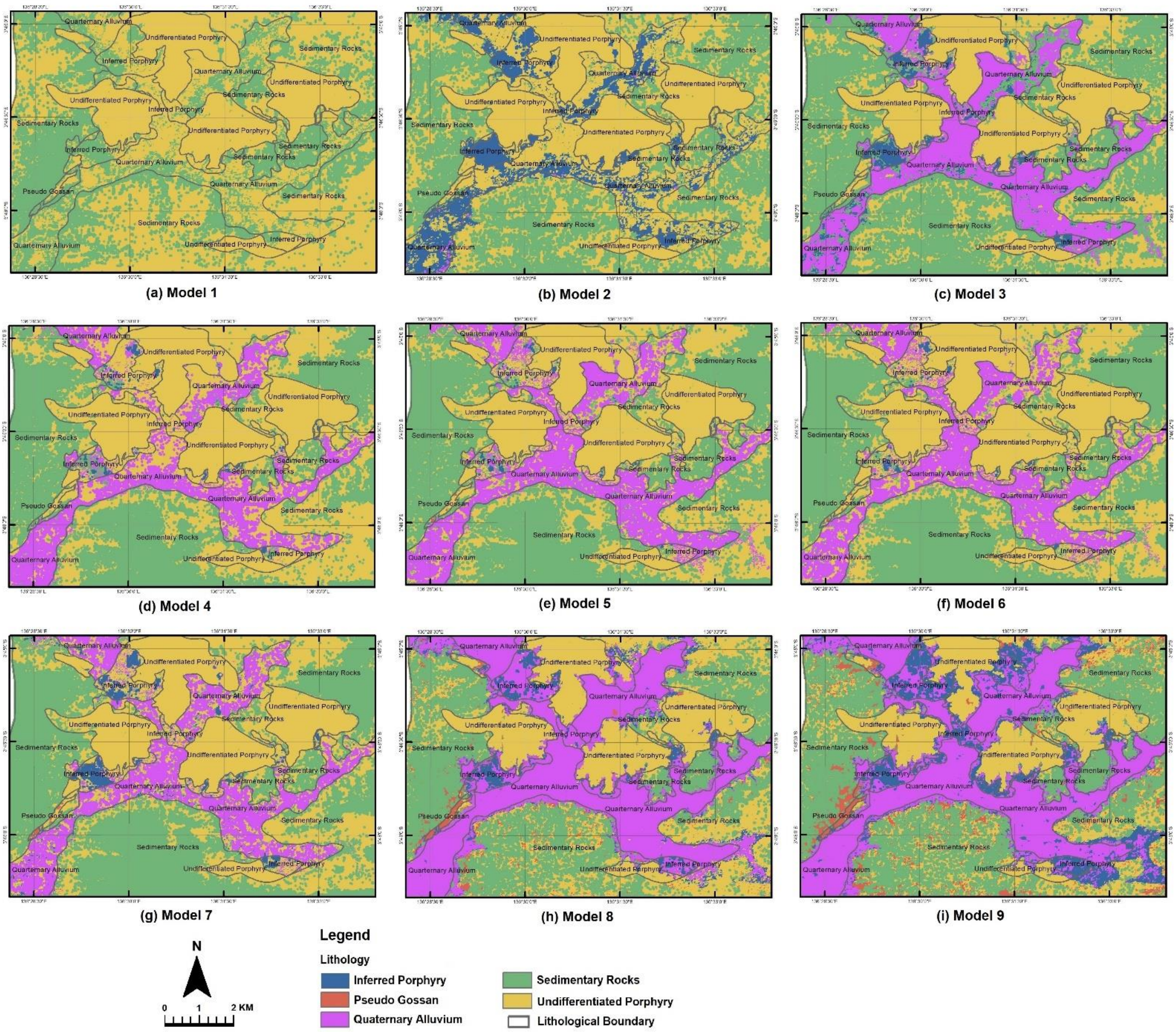

Fig. 8 Visualization of lithological boundaries resulted from 9 models overlaid with a lithological map of Mine Serve International [17]

4. Conclusion

Machine learning algorithm such as RF is a useful classification method for producing predictions of lithology. In this study, we have applied RF with hyperparameter tuning on satellite data consisting of Sentinel 2A, ALOS PALSAR, and Digital Elevation

Model (DEM) for lithological classification of the area Komopa by using very limited training points. This study showed that the overall accuracy based on testing points was moderate, but low values on recall, precision, and F1 Score. Meanwhile, the training accuracy achieved by each model is very high (0.92-1.0). Adding many training points to the simple random distribution (100-500) can increase performance accuracy.

In the models with simple random distribution training data, increasing the decision tree number from 25 to 200 will increase train accuracy (2-6%), test accuracy (1-2%) and decreases the F1 score by more than 100 decision trees. Meanwhile, in the models with stratified random distribution training data, increasing the decision tree's number has almost no effect on test accuracy, accuracy, and F1 Score. The number of minimum sample splits significantly affects train accuracy, but the effect is not significant on test accuracy and F1 Score. The optimal number of minimum sample splits is in the range of 5-10.

Compared with the existing lithological map, each model can generally depict the lithological boundaries. Model 9 gives the best classification result since its resulting lithological boundaries are closest to the existing lithological map. This study showed that the RF classification using balanced sample data, even though the number of training points is very small, could provide fairly good classification results in the predictive lithological mapping.

Limited training points and imbalanced conditions used in RF classification are problems that are often encountered in lithological mapping. Therefore, applying solutions at the data level (e.g., oversampling or under-sampling) and algorithm level (e.g., cost-sensitive learning). These solutions can serve as further research directions.

Acknowledgement

This study was partially funded by The Ministry of Education, Culture, Research, and Technology, Republic of Indonesia, through The Independent Campus Competition program. PT Eksplorasi Nusa Jaya supported the research data. This manuscript is part of doctoral theses in the Department of Geodesy and Geomatics, Faculty of Earth Sciences and Technology, Institut Teknologi Bandung, Indonesia.

References

[1] CHOE, B.H., TORNABENE, L.L., OSINSKI, G.R., and NEWMAN, J.D. Remote Predictive Mapping of the Tunnunik Impact Structure in the Canadian Arctic Using Multispectral and Polarimetric SAR Data Fusion. *Canadian Journal of Remote Sensing*, 2018, 44(5): 513–531.

[2] KUHN, S., CRACKNELL, M.J., and READING, A.M. Lithological Mapping via Random Forests : Information Entropy as a Proxy for Inaccuracy Lithological Mapping via Random Forests : Information Entropy as a Proxy for Inaccuracy. *ASEG Extended Abstracts*, 2016, 1: 1–4.

[3] XIE, Y., ZHU, C., ZHOU, W., LI, Z., LIU, X., and TU, M. Evaluation of Machine Learning Methods for Formation Lithology Identification: A Comparison of Tuning Processes and Model Performances. *Journal of Petroleum Science and Engineering*, 2018, 160: 182–193.

[4] REN, Q., CHENG, H., and HAN, H. Research on Machine Learning Framework Based on Random Forest Algorithm. *AIP Conference Proceedings*, 2017, 1820: 1–7.

[5] YING, X. An Overview of Overfitting and Its Solutions. *Journal of Physics: Conference Series*, 2019, 1168(2).

[6] YANG, L., and SHAMI, A. On Hyperparameter Optimization of Machine Learning Algorithms : Theory and Practice. *Neurocomputing*, 2020, 415: 295–316.

[7] LIVING TEXTBOOK. *Spectral reflectance curves*. https://ltb.itc.utwente.nl/page/498/concept/81713.

[8] HARVEY, A.S., and FOTOPOULOS, G. Geological Mapping Using Machine Learning Algorithms. Int. Arch. Photogramm. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016, 41(July): 423–430.

[9] BROWNLEE, J. Imbalanced Classification with Python: Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning. *Machine Learning Mastery*, 2020.

[10] BREIMAN, L. Random Forests. *Machine Learning Journal*, 2001, 45: 1–33.

[11] BOGNER, C., SEO, B., and ROHNER, D. Classification of Rare Land Cover Types : Distinguishing Annual and Perennial Crops in an Agricultural Catchment in South Korea. *PLoS One*, 2018, 13(1): 1–22.

[12] SCHRATZ, P., MUENCHOW, J., and ITURRITXA, E. Performance Evaluation and Hyperparameter Tuning of Statistical and Machine-Learning Models Using Spatial Data. *Ecological Modelling*, 2018, 406: 109–120.

[13] MUKUNTHU, D., SHAH, P., and TOK, W.H. *Practical Automated Machine Learning on Azure: Using Azure Machine Learning to Build AI Solutions* Quickly. O'Reilly Media, Inc., 2019.

[14] SCIKITLEARN. *Sklearn. Ensemble Random Forest Classifier*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[15] PROBST, P., WRIGHT, M., and BOULESTEIX, A.-L. Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining and Knowledge Discovery*, 2019, 9: 1–19.

[16] GLOVER, J.K. The *Structural and Lithological Setting, Controls of Mineralization and Potential in the Area of the Komopa-Dawagu Prospects*. NBM block II, 1999.

[17] MINE SERVE INTERNATIONAL. *Geological Map Scale of 1:25.000, 2nd Edition*. Komopa, Papua, Indonesia, 2000.

[18] EUROPEAN SPACE AGENCY. *Sentinel 2A*. https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/product- types/level-2a.

[19] L3HARRIS. *Vegetation Suppression*. https://www.l3harrisgeospatial.com/docs/vegetationsuppression.html.

[20] EUROPEAN SPACE AGENCY. *Level-1 Radiometric Calibration*. https://sentinel.esa.int/web/sentinel/radiometric-calibration-of-level-1-products.

[21] EUROPEAN SPACE AGENCY. *Radar Course 2 - Shadow - Course 2 - ERS Radar Courses - ESA Operational EO Missions - Earth Online – ESA.* https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/ers/instruments/sar/applications/radar-

courses/content-2/-
/asset_publisher/qIBc6NYRXfnG/content/radar-course-2-shadow.
[22] OTTINGER, M., and KUENZER, C. Spaceborne L-Band Synthetic Aperture Radar Data for Geoscientific Analyses in Coastal Land Applications: A Review. *Remote Sensors*, 2020, 12(14): 1–36.
[23] BADAN INFORMASI GEOSPASIAL REPUBLIK INDONESIA. *DEMNAS Seamless Digital Elevation Model (DEM) dan Batimetri Nasional*. http://tides.big.go.id/DEMNAS/#Info.
[24] NOORHALIM, N., ALI, A., and SHAMSUDDIN, S.M. Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE. *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. 2019, 19–30.
[25] PEDREGOSA, F., VAROQUAU, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., DUBOURG, V., and VANDERPLAS, J. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2018, 12: 2825-2830.
[26] CAKIR, E., and ULUKAN, Z. *Digitalization on Aviation 4.0: Designing a Scikit-Fuzzy Control System for In-Flight Catering Customer Satisfaction*. Springer, 2022.
[27] EARTHEXPLORER, n.d. https://earthexplorer.usgs.gov/
[28] ALASKA SATELLITE FACILITY, 2021. https://www.asf.alaska.edu/

参考文:

[1] CHOE, B.H., TORNABENE, L.L., OSINSKI, G.R., 和 NEWMAN, J.D. 使用多光譜和極化特区數據融合對加拿大北極地區 圖努尼克撞擊結構進行遠程預測製圖。加拿大遙感雜誌，2018，44(5)：513–531。

[2] KUHN, S.、CRACKNELL, M.J., 和 READING, A.M. 通過隨機森林的岩性測繪：信息熵作為不准確性的代理通過隨機森林的岩性映射：信息熵作為不准確性的代理。擴展摘要，2016，1：1-4。

[3] XIE, Y., ZHU, C., ZHOU, W., LI, Z., LIU, X., 和 TU, M. 用於地層岩性識別的機器學習方法評估：調整過程和模型的比較表演。石油科學與工程雜誌，2018，160：182-193。

[4] REN, Q., CHENG, H., 和 HAN, H. 基於隨機森林算法的機器學習框架研究。會議論文集，2017，1820：1-7。

[5] YING, X. 過擬合及其解決方案概述。物理學雜誌：會議系列，2019，1168(2)。

[6] YANG, L., 和 SHAMI, A. 關於機器學習算法的超參數優化：理論與實踐。神經計算，2020，415：295-316。

[7] 活的教科書。光譜反射曲線。https://ltb.itc.utwente.nl/page/498/concept/81713。

[8] HARVEY, A.S., 和 FOTOPOULOS, G. 使用機器學習算法的地質圖。國際。拱。攝影。國際攝影測量檔案，遙感和空間信息科學，2016，41（七月）：423–430。

[9] BROWNLEE, J. 使用 蟒蛇進行不平衡分類：選擇更好的指標、平衡偏斜類和應用成本敏感學習。機器學習精通，2020。

[10] BREIMAN, L. 隨機森林。機器學習雜誌，2001，45：1-33。

[11] BOGNER, C.、SEO, B., 和 ROHNER, D. 稀有土地覆蓋類型分類：區分韓國農業集水區的一年生和多年生作物。 公共科學圖書館一，2018，13（1）：1-22。

[12] SCHRATZ, P.、MUENCHOW, J., 和 ITURRITXA, E. 使用空間數據的統計和機器學習模型的性能評估和超參數調整。生態建模，2018，406：109-120。

[13] MUKUNTHU, D.、SHAH, P., 和 TOK, W.H. 天藍色上的實用自動化機器學習：使用 天藍色機器學習快速構建解決方案。奧萊利媒體公司，2019。

[14] SCIKIT 学习。學習。集成隨機森林分類器。https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html。

[15] PROBST, P., WRIGHT, M., 和 BOULESTEIX, A.-L. 隨機森林的超參數和調整策略。 WIREs 數據挖掘和知識發現，2019，9：1-19。

[16] GLOVER, J.K. 科莫帕-達瓦古勘探區的構造和岩性環境、礦化控制和潛力。 國家广播公司區塊 II，1999。

[17] 礦山服務國際。地質圖比例為 1:25.000，第 2 版。科莫帕，巴布亞，印度尼西亞，2000。

[18] 歐洲航天局。哨兵 2 一种。https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/product-types/level-2a。

[19] 升 3 哈里斯。植被抑制。https://www.l3harrisgeospatial.com/docs/vegetationsuppression.html。

[20] 歐洲航天局。1 級輻射校準。https://sentinel.esa.int/web/sentinel/radiometric-calibration-of-level-1-products。

[21] 歐洲航天局。雷達課程 2 - 陰影 - 課程雷達課程 - 歐空局操作 EO 任務 - 地球在線 - 歐空局。https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/ers/instruments/sar/applications/radar-courses/content-2/-
/asset_publisher/qIBc6NYRXfnG/content/radar- course-2-shadow。

[22] OTTINGER, M., 和 KUENZER, C. 用於沿海土地應用地球科學分析的星載 L 波段合成孔徑雷達數據：綜述。遙感器，2020，12(14)：1-36。

[23] 印度尼西亞巴丹信息地理空間共和國。DEMNAS 無縫數字高程模型和國家水深測量。http://tides.big.go.id/DEMNAS/#Info。

[24] NOORHALIM, N.、ALI, A., 和 SHAMSUDDIN, S.M. 使用處理類不平衡問題的不平衡比率。第三屆計算、數學和統計國際會議論文集。2019, 19-30。

[25] PEDREGOSA, F., VAROQUAU, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., DUBOURG, V., 和 VANDERPLAS, J. 科学学习：蟒蛇中的機器學習。機器學習研究雜誌, 2018, 12：2825-2830。

[26] CAKIR, E., 和 ULUKAN, Z. 航空 4.0 數字化：設計西基特模糊控制系統以提高機上餐飲客戶滿意度。斯普林格, 2022。

[27] 地球探索者，未注明日期 https://earthexplorer.usgs.gov/

[28] 阿拉斯加卫星设施，2021。https://www.asf.alaska.edu/