Assessing machine learning techniques for detailing soil map in the semiarid tropical region

D Cahyana1,2, B Barus1, Darmawan1, B Mulyanto1 and Y Sulaeman2

1 Department of Soil Science and Land Resource, IPB University, Bogor, Indonesia
2 Indonesian Center for Agricultural Land Resources Research and Development, Bogor, Indonesia

E-mail: destika_cahyana@yahoo.com

Abstract. The major problem detailing soil map in large tropical country such as Indonesia is high cost and time-consuming. The machine learning technique is one of DSM methodologies that explores spatial patterns to predict soil class and soil attribute. K-nearest neighbours (KNN), random forest (RF) and support vector machine (SVM) are popular for detailing soil map in temperate country, but it is still rare to be applied in a tropical country. This study aimed to assess three machine learning in updating soil map from 1:50,000 to 1:250,000 scale in the semiarid tropical region. The existing soil map was collated and then derived environmental covariates representing soil-forming factors from the digital elevation model. There were 72 training datasets were originating from polygon soil maps used as input for these machine learning to recognize the pattern and predict soil class map in Bikomi Utara Sub District, Timor Tengah Utara Regency, Indonesia. Overall accuracy and kappa coefficient by KNN for the best three predictive soil maps were 74-75% and 0.62-0.63, respectively; and followed by SVM, 71-73% and 0.58-0.60; and the last RF, 69-75% and 0.55-0.63. This research revealed that machine learning of the KNN is potentially for updating soil map in a tropical semiarid area.

1. Introduction
Machine learning application as a branch of artificial intelligence in the digital era is increasingly widespread in various scientific fields including soil science, especially in digital soil mapping (DSM) [1,2]. Machine learning may be used to recognize, identify, and predict patterns in several big data derived from satellite data and its derivative data such as digital elevation models (DEM). These data play a role as controlling variables or environmental covariates that represent soil-forming factors [3].

Currently, derivative data of DEM such as slope, curvature, aspect, convexity, concavity, and geomorphons are getting easier and faster as computer information technology development [4-6]. In soil science, the data can represent environmental covariates, especially topography or relief factors (r). Recently the above environmental covariates are often called part of the 'scorpan' factor [7].

This term is the Jenny's concept development, namely 'clorpt' or often written as 'corpt' that is called as the five factors of soil formation [8,9]. Jenny's concept, was developed from a concept that initiated by V.V. Dokuchaev from Russia. He stated that soil developed from 4 soil-forming factors that work over time [10]. Four other soil-forming factors are: climate (c), organism (o), parent material (p), and time (t) [8].

On the other hand, almost all countries, including Indonesia, have soil map archives created by their national soil research institutes [11]. Generally, soil maps were created by the manual method through a field survey with a course scale. Fortunately, Indonesia has been finished producing soil maps in a...
semi detail scale, 1: 50,000, for the whole area in 511 districts and cities. The soil map archives were created from the previous survey may be referred as soil legacy map.

By utilizing the soil legacy maps, various derivative data of DEM can be useful as predictors for detailed legacy maps, especially to determine and to obtain more accurate information about the position of soil type on toposequence. Conventional soil legacy maps were an invaluable source of information because they represent expert knowledge of soil environment relationships at the landscape in the specific relevant areas [12]. Machine learning, is applied to recognize the patterns in some training data provided, it can be used to predict the overall data [13].

Recently the utilization of machine learning to detail of soil legacy maps by environmental covariates has only been limited in subtropical areas that have more complete of soil legacy maps and satellite data. In subtropical areas, the use of various satellite data has also been reported to be successful due to minimum cloud cover, low and homogeneous of vegetation, and relatively homogeneous of the soil parent material.

Nevertheless, the application of machine learning in tropical regions is still rare. The present study aimed to evaluate the machine learning applications on several environmental covariates for detailed soil legacy maps in the tropics area. The selected covariates are some DEM derivative data and vegetation cover which represent the 'r' and 'o' factor in the 'scorpan' function, while the others covariates are relatively homogenous because the study area is on a small island. This study also aimed to determine the type of machine learning that is the most suitable for application and the best combination of covariates environment data in the tropics area.

2. Materials and methods

2.1. Study area

The study was conducted in North Bikomi District, North Central Timor, East Nusa Tenggara Province with tropical savanna (Aw) sub-climate as shown in figure 1. The climate is dry that is indicated the soil moisture regime in the study area that is Ustic. Based on soil legacy map at 1: 50,000 scale that produced by the Indonesian Center for Agricultural Land Resources Research and Development, Bogor. The soil in the study area is a landscape dominated by Typic Haplusterts with a combination of Lithic Haplusterts, Typic Haplustalf, Typic Ustorthents, and rock out crops (ROC). The dominant of parent material is marl with a combination of andesite and basalt, and clay and sand deposits.

![Figure 1. Study Area at Sub-Distric of Bikomi Utara, Regency of Timor Tengah Utara](image)

2.2. Datasets

DEMNAS 8 Data from the Geospatial Information Agency, Indonesia has been used to derive various environmental covariates. DEMNAS 8 is processed using the System for Automated Geoscientific Analysis (SAGA) 7.6.3 to produce an 18-layer covariate environment of relief (r). The other datasets are ALOS Palsar 2 with polarization HH and HV that represent the covariate environment of the
organism (o). All covariates are shown in figure 2. The data obtained are stacked as shown in table 1, then selected one by one using the Ranking a Random Feature for Variable and Feature Selection technique based on Gram-Schmidt orthogonalization [14].

Subsequently, based on the soil legacy map, there were 72 polygons collected for training data. Every types of soil training data in a soil map unit (SMU) were collected using the approximation technique of soil position in a landscape based on tacit knowledge of toposequence. Then machine learning used the training data to determine patterns of soil distribution at the facet of the landscape.

![Figure 2](image)

*Figure 2.* Example of covariates from left above to left below (clockwise) are DEMNAS, analytical hill shading, slope, and aspect.

| No | Layer Data                        | Combination of layers |
|----|-----------------------------------|-----------------------|
|    | Covariate of Environment          | RF/ SVM/ KNN | RF07/ SVM07/ KNN07 | RF08/ SVM08/ KNN08 | RF09/ SVM09/ KNN09 | RF10/ SVM10/ KNN10 |
| 1  | Fill Demnas                       | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 2  | Analytical Hillshading            | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 3  | Slope                             | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 4  | Aspect                            | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 5  | Cross sectional curvature         | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 6  | Longitudinal curvature            | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 7  | Convergence index                 | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 8  | Close depression                  | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 9  | Flow accumulation                 | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 10 | Topographic wetness index         | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 11 | LS Factor                         | ✓          | ✓                  | ✓                  | ✓                  | ✓                  |
| 12 | Channel network base level        | ✓          | ✓                  | ✓                  | ✓                  | x                  |
| 13 | Vertical distance to channel network | ✓        | ✓                  | x                  | x                  | x                  |
| 14 | Valley depth                      | ✓          | x                  | x                  | x                  | x                  |
| 15 | Relative slope position           | ✓          | x                  | x                  | x                  | x                  |
| 16 | Profil curvature                  | ✓          | x                  | x                  | x                  | x                  |
| 17 | Plan curvature                    | ✓          | x                  | x                  | x                  | x                  |
| 18 | Convexity                         | ✓          | x                  | x                  | x                  | x                  |
| 19 | Texture                           | ✓          | x                  | x                  | x                  | x                  |
| 20 | HH 1                              | ✓          | x                  | x                  | x                  | x                  |
| 21 | HV 1                              | ✓          | x                  | x                  | x                  | x                  |

*Table 1.* Covariate environments of layer and their combination for prediction soil map.
2.3. Machine learning

Machine learning is a branch of artificial intelligence based on the Alan Turing concept who developed by a simple question: “can machine thinking?” Turing developed computer program to answer his question [15]. In this context, machine learning is computer programs that designed as a human who able to complete certain tasks by learning some data from experience (training data). When the experience of machine learning is increased by the addition of training data, the ability of machine learning will increase.

There are many types of machine learning for digital soil mapping. This research has used three machine learning, namely K-nearest neighborhood (KNN), random forest (RF), and support vector machine (SVM). They were reported to be satisfactory for the application in subtropical areas [1-3]. K-nearest neighborhood (KNN) is machine learning that utilizes the main concept of Tobler's First Law of Geography that is something that is close together is more related than something far away. KNN is a non-parametric machine learning that is classified as a supervised classification [16].

The random forest (RF) is a classification method in the form of a decision tree that assumes that there are many trees in a forest, to produce a more stable and accurate classification than a single decision tree [17]. Each tree in the random forest is classified by training data on the sample data. The more trees used, the accuracy of the results increases. Each class selected by each tree is collected in the baskets. The determination of classification class in a random forest is taken from the most votes in each basket of the classification of each tree.

Support vector machine (SVM) is machine learning to analyze data especially for pattern recognition, classification, regression analysis, and detecting recently. The principle of SVM is to find and assign a hyperplane (line or plane) to separate different classes of data [18]. Originally SVM is a linear classification method, but it is developed into a non-linear classification method with high accuracy [19] to recognize objects the images, handwriting, and sound. The purpose of this study was to evaluate the application of three machine learning for detailed soil legacy maps in the semiarid tropics area. This study also aimed to determine the best combination of covariates environment data in the semiarid tropic area.

3. Results and discussion

3.1. Visual assessment

The example of various predictive map patterns were generated from machine learning as shown in figure 4. Each machine learning produces a unique map pattern when used to recognize data from various combinations of environmental covariates.

The principle of the best map prediction is closest to reality and it is produced by minimum combination of layers for easier and faster processing. Therefore, using a large number of datasets is not a recommendation. The initial approach to predict the map resemble to reality is to evaluate the stability level of the pattern from a number of some combinations.

All environmental covariates shown in table 1 consist of 21 layers. The combination of them is based on the eliminated layer one by one. Based on figure 3, the pattern stability level in each machine learning is divided into 2 major patterns. The layer pattern was eliminated by less than 8 layers tends to be different from the pattern that was eliminated by more than 8 layers. In other words, the combination of 13 environmental covariate layers was still resemble to the combination of 21 layers.

The pattern was changed increasingly after eliminated covariates more than 8 layers. However, 4 combinations that showed the transitional patterns of the 2 different patterns. Thus, the 4 combinations were chosen and evaluated them by quantitative comparison, specifically map to map comparison and map to point observation. Both of map and point for reference are legacy data.

This study also evaluated predictors with correlation more than 0.3 and less than -0.3, especially for derivative data of DEMNAS as shown in table 2. The higher value, the more positive or negative it is indicated a strong correlation amount the predictors and the probability map. The predictors were convexity, LS factor, topographic wetness index, slope and plan curvature.
However, they were combined with DEMNAS elevation data and land cover data even though the correlation value was only around 0.1. On the other hand, the total number of data layers was only 8 layers. The probability map generated from the 8 layers for each machine learning can be seen in figure 4.

Figure 3. Workflow of soil map and environment covariates to the update of soil map using three types of machine learning: KNN, RF, and SVM

Figure 4. The example of various combination of layer environment covariate with K-nearest neighbor. The last number refers to amounts of eliminated layers

3.2. Quantitative Assessment
The four combinations of transition patterns for each machine learning were evaluated by the confusion matrix to obtain total accuracy and kappa analysis to obtain kappa values. The results of the analysis is shown in figure 3 (left). In machine learning of KNN, the total accuracy and kappa coefficient for the three best probably maps were 74-75% and 0.62-0.63, respectively. Next followed by machine learning of SVM, 71-73% and 0.58-0.60. Finally, machine learning of RF 69-75% and 0.55-0.63

Reciprocally in the figure 3 (right), the total accuracy and kappa coefficient of machine learning KNN that used predictors with correlation values above 3 and below -3 was 76% and 0.64, respectively. Next SVM was 75% and 0.62, then RF was 73% and 0.6. Based on this range of numbers, the highest
overall accuracy and kappa coefficient maps were consistently produced by KNN machine learning, although in general the quality of probability map produced by three machine learning was moderate.

| Covariate of Environment | Correlation | Covariate of Environment | Correlation |
|--------------------------|-------------|--------------------------|-------------|
| Fill Demnas              | 0.1146704   | Channel network base level | 0.1285487   |
| Analytical Hillshading   | -0.01396308 | Vertical distance to channel network | -0.03302333 |
| Slope                    | -0.2968322  | Valley depth             | 0.001282321 |
| Aspect                   | 0.1235603   | Relative slope position  | -0.05194798 |
| Cross sectional curvature| 0.07968507  | Profil curvature         | -0.02817093 |
| Longitudinal curvature   | 0.0677387   | Plan curvature           | -0.2311643  |
| Convergence index        | 0.08160912  | Convexity                | -0.3452497  |
| Close depression         | -0.01358279 | Texture                  | 0.1160438   |
| Flow accumulation         | 0.1469438   | HH 1                     | 0.15         |
| Topographic wetness index | 0.3030866   | HV 1                     | 0.15         |
| LS Factor                | -0.3151176  |                          |             |

**Table 2.** Spatial correlation of predictors for each covariates environment

![Figure 5](image1.png)

**Figure 5.** Probability soil maps resulted by random forest (left), support vector machine (center), and K-nearest neighbor (right) using predictors: elevation, convexity, LS factor, topographic wetness index, slope, and plan curvature (relief predictor), and HH and HV polarization (land cover).

![Figure 6](image2.png)

**Figure 6.** Overall accuracy (blue line) and kappa value (orange line) for the most 4 probably maps by eliminated layer of combination (left) and probably map by correlated predictor (right) using 3 types of machine learning.

### 3.3. Soil landscape relations

Based on the soil legacy map that is the source of information in this study, the location of study is the dominated area by young soil of Typic Haplustepts. This soil was developed in dry area with soil moisture regime of Ustic. The dominance of Typic Haplustepts can be shown from each soil map unit (smu) in the study area with 6 smu. There are Typic Haplustepts in each smu with dominant (D), fair (F), and minor (M) proportions.

The study area in Bikomi Utara has dominant to the minor parent material, it is marl, andesite and basalt, and clay and sand deposits, respectively. Based on this setting of environment, the dominant soil...
formed is initial, juvenile, and virile stage because of low rainfall so the rate of weathering is relatively slow. In some narrow smu, soil with continued weathering was found such as Ustalfs. This soil was developed from andesite and basalt parent material that has high to very high base saturation.

The landform setting in the location of the study is tectonic mountains, tectonic hills, and inter-hill plains. Several areas are a volcanic intrusion and flood plains. The relief has a very steep mountainous, very steep hills, steep hills, and partly undulating. This environment produced shallow soils such as Lithic Haplustepts and a large number of rock outcrops (ROC). Based on 1:50,000 scale soil legacy map in Bikomi Utara District, there are 4 subgroups of soil, namely Typic Haplustepts, Lithic Haplustepts, Typic Ustorthents, and Typic Haplustalfs.

A study in Maraharastra, India, that is similar to Bikomi Utara had been conducted in the Mohgaon Area of Nagpur District, Maharashtra, to analyze IRS-ID LISS-III satellite data, to reveal distinct of geomorphological units in the plateau top, isolated mounds, mountains, slopes steep, plateau spurs, subdued plateau, rolling plains, pediments, narrow valleys and main valley floor [20]. The identification of soil profile showed variations in the site and soil morphological characteristics. Moderate soil erosion was occurring on the top of the plateau, isolated mounds, plateau spurs, rolling plains and pediments. Severe erosion was identified on the subdued plateau, while narrow valleys suffer very slight erosion.

Based on these environmental settings, the study was conducted in Bikomi Utara by 3 types of machine learning also revealed the relation between landform and the soil. Very shallow soils were generally found at the tops of mountains and isolated mounds. Shallow soils are found on ridges, steep slopes, and highlands and rolling plains. The shallow and very shallow soils were identified as Lithic Haplustepts or Typic Ustorthents. Moderately deep and deep soils on flat plateaus, pediments, and main valley floors were identified as Typic Haplustepts.

![Figure 7](image)

**Figure 7.** Legacy soil map of scale 1:50,000 (left-above). Soil map reference (left-below) and soil probability map by eliminated predictors and by correlated predictor using KNN (2 of middle map below) and using SVM (2 of right map below).

This study reveals that predictors derived from DEMNAS elevation data will be useful for digital soil mapping based on machine learning, especially when others soil-forming factors such as parent material and climate were relatively homogenous. Machine learning may recognize and classify patterns of landform and soil relationships from some training data provided.

However, the utilization of machine learning for detailing the legacy soil map in this study has some limitations. Machine learning is only adequate to determine the probability of position at a particular landscape for each sub-group soil in the soil legacy map. Machine learning is unable to identify new soil inclusion or minor subgroups of soil that were not recorded on the soil legacy map previously. Machine learning also is unable to identify the deep hierarchical level such as soil families or soil series. A study to identify new subgroups that were not recorded as minor soil and study to detail for deep hierarchical level requires new field surveys with dense of observation.
4. Conclusions
The machine learning of KNN produced total accuracy and kappa coefficient for the three best predictive maps with a value of 74-75% and 0.62-0.63, respectively; and followed by SVM, 71-73%, and 0.58-0.60; and the last, RF produced 69-75% and 0.55-0.63. Consistently KNN produced better predictive maps and followed by SVM and RF. However, they have the same quality that is classified as moderate. The best combination of covariates is 13 layers, they are digital elevation mode, analytical hill shading, slope, aspect, cross-sectional curvature, longitudinal curvature, convergence index, close depression, flow accumulation, topographic wetness index, ls factor, channel network base level and vertical distance to channel network.

Acknowledgments
The authors are grateful to the Indonesian Center for Agricultural Land Resources Research & Development (ICALRRD) that provided some soil legacy map and allows the author to use it. Authors are also grateful to AFACI (Asian Food and Agriculture Cooperation Initiative) for facilitating a variety of literature to be studied from various perspectives. Thanks also to Dr Edi Yatno who helped to provide a number of the legacy soil data and map at the location of the study.

References
[1] Heung B, Chak H, Zhang J, Knudby A, Bulmer C E and Schmidt M G 2016 An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping Geoderma 265 62–77
[2] Minasny B and McBratney A B 2016 Digital soil mapping: A brief history and some lessons Geoderma 264 301–11
[3] Brungard C W, Boettinger J L, Duniway M C, Wills S A and Edwards T C 2015 Machine learning for predicting soil classes in three semi-arid landscapes Geoderma 239 68–83
[4] Zeraatpisheh M, Ayoubi S, Brungard C W and Finke P 2019 Disaggregating and updating a legacy soil map using DSMART, fuzzy c-means and k-means clustering algorithms in Central Iran Geoderma 340 249–58
[5] Zeraatpisheh M, Jafari A, Bagheri Bodaghabadi M, Ayoubi S, Taghizadeh-Mehrjardi R, Toomanian N, Kerry R and Xu M 2020 Conventional and digital soil mapping in Iran: Past, present, and future Catena 188 104424
[6] Jasiewicz J and Stepinski T F 2013 Geomorphons - a pattern recognition approach to classification and mapping of landforms Geomorphology 182 147–56
[7] McBratney A B, Mendonça Santos M L and Minasny B 2003 On digital soil mapping vol 117
[8] Jenny H 1994 Factors of Soil Formation: A System of Quantitative Pedology (Dover Publications)
[9] Zhang G lin, Liu F and Song X dong 2017 Recent progress and future prospect of digital soil mapping: A review J. Integr. Agric. 16 2871–85
[10] Savin I Y, Zhogolev A V, and Prudnikova E Y 2019 Modern Trends and Problems of Soil Mapping Eurasian Soil Sci. 52 471–80
[11] Odgers N P, Sun W, McBratney A B, Minasny B and Clifford D 2014 Disaggregating and harmonising soil map units through resampled classification trees Geoderma 214–215 91–100
[12] Liu X, Zhu A X, Yang L, Pei T, Liu J, Zeng C and Wang D 2020 A graded proportion method of training sample selection for updating conventional soil maps Geoderma 357 1–9
[13] Zhang M, Shi W and Xu Z 2020 Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data Hydrol. Earth Syst. Sci. 24 2505–26
[14] Stoppiglia H, Dreyfus G, Dubois R and Oussar Y 2003 Ranking a random feature for variable and feature selection J. Mach. Learn. Res. 3 1399–414
[15] Turing A M 1950 Computing Machinery Intelligence Mind XX 150–3
[16] Heung B, Ho H C, Zhang J, Knudby A, Bulmer C E and Schmidt M G 2016 An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping Geoderma 265 62–77
[17] Breiman L 2001 *Random Forests* Mach. Learn. pp 5–32
[18] Vapnik V N 1999 *The Nature of Statistical Learning Theory* (Springer)
[19] Han J, Kamber M and Pei J 2012 *Data Mining: Concepts and Techniques* (Morgan Kaufmann, Elsevier)
[20] Raghavendra Reddy M G, Obi Reddy G P, Maji A K and Nageshara Rao K 2003 Landscape analysis for pedo-geomorphological characterization in part of basaltic terrain, Central India using remote sensing and GIS *J. Indian Soc. Remote Sens.* 31 271–82