Machine learning approach for peatland delineation using multi-sensor remote sensing data in Ogan Komering Ilir Regency

Dionysius Bryan Sencaki, Hari Prayogi, Siti Arfah, Teguh Arif Pianto
Center for Regional Resource Development Technology– Agency for the Assessment and Application of Technology (BPPT), Geostech 820 Building Kawasan Puspiptek, Tangerang Selatan, Indonesia
e-mail: dionysius.bryan@bppt.go.id

Abstract. Peatland is the region that is composed of accumulated layers of decayed vegetations and organic matters. It has an important role in regulating the hydrology cycle and maintaining global climate stability. Therefore protecting peatland from human intervention is imperative and the availability of accurate delineation map of peatland and non peatland is substantial and indispensable. Landsat 8 OLI and MODIS are proven to be beneficial for terrestrial analysis including peatland. For detecting peatland, the use of spectral bands to generate indexes that sensitive to the detection of Net Ecosystem Exchange (NEE) that related to peat water table, biomass vegetation and surface energy budget is the key. Therefore, several spectral-derived Landsat 8 OLI and MODIS products alongside with and DEM data were utilized. For classification, Machine Learning (ML) and Deep Learning (DL) methods namely Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGB) and Cat Boosting (CatB) were ML method that used alongside with Artificial Neural Network (ANN) of DL to delineate peatland distribution. The best result of the delineation was achieved by CatB algorithm with 75.52 % accuracy based on testing data and 82.61% based on validation data of field survey that held in Ogan Komering Ilir (OKI), South Sumatera Province.

1. Introduction
According to the Ramsar Convention, wetlands are areas in which water is a prominent factor that controls environment continuity. It is occurred in a location that fully covered by shallow water or when water table is at or near at soil surface. One of major wetland types that cover more than 3% of global land layer is peatland. In Indonesia peatland covers more than 14.9 million hectares or about 10% of its total land mass. It is mostly found in Sumatera, Kalimantan and Papua islands. Generally, peatland is found on even surface and formed when the accumulation rate of organic matter from decayed vegetation is higher than the decomposition process. The peat formation process can take for centuries and its growing rate is only 1 mm per year (1). With such a slow succession rate, it is imperative to conserve peatland from further anthropogenic disturbances such as peat conversion or draining for agriculture, farmland, plantation and settlement expansion as beside the restoration process may take thousand years, it can release as much as two gigatons of greenhouse gas emission to atmosphere per year as result of carbon oxidization (2). Considering how important and prone peatland is, the availability of updated and accurate delineation peatland map is indispensable as it helps authorities to precisely measure the extent of conservation plan. By using remote sensing technology, it is possible to achieve that task. Remote sensing technology is a technology that allows us to acquire information based on spectral reflectance from the object without having physical contact with it. Some satellite-based remote sensing technology such as Landsat 8 OLI and MODIS capture spectral reflectance information through multispectral sensor Reflectance data from those sensors have already been primary sources for terrestrial analysis over decades by scientists and proven to deliver accurate analysis. In peatland application, the potential use of multispectral sensors is promising especially in delineation task (3-5).
By using spectral-derived indices data from Landsat 8 OLI and MODIS sensor, it is possible to spatially extract peatland key characteristics such as Net Ecosystem Exchange (NEE) that related with peat water table, biomass from peat surface and rate of surface energy budget between peat surface layer and atmosphere (6). DEM-derived data is also useful to extract slope features as area flatness is another significant peatland characteristic.

To solve the complexity of using data from multiple satellite sensors for delineation, Machine Learning (ML) and Deep Learning (DL) methods were employed. Random Forest (RF) and Gradient Boosting (GB) are the most popular and well known ML algorithms in terms of classification even when classifying data with high complexity (7). RF and GB are also understandably to be flexible and able to handle missing data. In further development, there are Extreme Gradient Boosting (XGB) and CatBoost (CatB) methods as the improved version of RF and GB. XGB and CatB are able to reduce more variance and bias, and can outperform RF or GB in terms of flexibility and performance speed (8). Meanwhile Artificial Neural Network (ANN) of DL is well known for its ability that mimics human neural network and able to provide high performance with a huge amount of data. Those algorithms were all simultaneously employed to get accurate peatland delineation map.

2. Study Location

The study for this research was in Ogan Komering Ilir (OKI), which situated on the Eastern part of South Sumatera Province and adjacent to Bangka Belitung Island. OKI lies between S 2.415 – 4.286 and E 104.636 – 106.22 with Kota Kayu Agung as its administrative capital. OKI is well known as one of regencies in Indonesia with vast wetland and peatland area.

3. Data and Methodology

3.1. Data

Primary data used for this research were Landsat 8 OLI, MODIS Aqua – Terra and DEM. For 30-m Landsat 8 OLI, Level – 2 products of Surface Reflectance for band 5 to band 7 were used coupled with derived data such as PCA band 1 to 3 and several Indexes namely MSAVI, EVI, NDMI, NDVI, MWI and 6 Tasseled Cap bands. For MODIS Aqua – Terra, product MOD16A2 of ET and MOD17A2H of GPP were utilized. This product has the lowest spatial resolution which is 500 m. Meanwhile for DEM, data was acquired from 8-m product of DEMNAS delivered by BIG Indonesia and derived to get Slope feature. Those data were expected to be able to explain and identify the characteristics and physical nature of peatland. Both Landsat and MODIS data were acquired for the whole year 2017 and went through pre processing steps such as clipping over study location, cloud masking, digital number factor scaling and mosaic before preceded to the classification process.

3.2. Methodology
The methodology approach conducted in this paper mainly was to exploit the ability of each ML dan DL algorithms, namely RF, GB, XGB, CatB and ANN – DL, to perform binary classification based on attributes extracted from data derived from remote sensing. There were more than 20 columns and thousands of rows of primary data attribute that were ready to classify. The classification was technically performed on tabular-based data instead of direct classification towards raster-typed imagery as the tabular-based classification process could be delivered in more rapidly and efficiently. The spacing grid of the classification points was matched with the lowest resolution of main data, which is 500 meter, that come from MODIS data. Points for training, testing and validation that come from field survey and peatland data reference were also extracting information from remote sensing.
Figure 2. Simplified Workflow of Classification Process

From figure 2 we can infer that classification performance was solely dependent on how ML and DL model architecture developed. Model architecture was defined by the process called hyperparameter tuning and this is the most important part to build a robust model. Hyperparameter tuning is the process to determine the best possible set of parameter constants for model and to anticipate the potential of...
underfit or overfit conditions. Underfit or overfit condition is the condition where model tries to more predict noises rather than actual data and therefore causing misinterpretation of model performance. During hyperparameter tuning process, monitoring loss function is important to decide whether some parameter constants need to be changed once indication of underfit or overfit to be found. The other significant condition to create a proper model is to have balanced data. Data need to be balanced and not biased towards a certain class to build an optimal model. Even though there were algorithms to tackle data discrepancy such as undersampling, oversampling or SMOTE, having a balanced data from the beginning is important to maintain the nature condition of data (9, 10). Data reference was also important to be included as it gave early insight about how peatland was spatially distributed and also to give remarks for validation survey.

4. Result and Analysis

Results produced by each ML and DL algorithm were assessed by evaluating the accuracy percentage between training – testing data and training – validation data. Training and testing data were produced from splitting combination of field survey and peatland reference data with ratio 70% for training and 30% for testing. The main classification model was developed using training data and first assessed by testing data. Later then, the assessment continued using validation data that gathered from field survey in different amount and taken in different area from training and testing data.

![Accuracy Performance Training - Testing](image)

**Figure 3.** Accuracy Performance between Training and Testing Data

CatB was the best algorithm as shown by figure 3 to predict the distribution of peatland based on assessment using Testing data with accuracy over 75% percent. The least accurate was Random Forest with accuracy slightly over 70%. Even though CatB topped the group with the highest percentage, it might not always be able to achieve the similar accuracy if the model applied with completely separated and different data from training and testing data therefore the model must be assessed again with validation data. If all algorithms accuracy scores showed consistent manner then it could be concluded that the models were optimal and able to depict the general characteristics of peatland.
Figure 4. Final Accuracy Performance between Training and Validation Data

Table 1. Classification Accuracy Metric

| No. | Classifier                | Training – Testing | Training – Validation |
|-----|---------------------------|---------------------|-----------------------|
| 1.  | Random Forest (RF)        | 70.58%              | 69.90%                |
| 2.  | Gradient Boosting (GB)    | 72.65%              | 79.26%                |
| 3.  | Extreme Gradient Boosting (XGB) | 73.96%      | 76.59%                |
| 4.  | Categorical Boosting (CatB) | 75.52%            | 82.61%                |
| 5.  | Deep Learning (DL)        | 73.69%              | 74.92%                |

The latest figure and table above showed how CatB perpetually produced high performance for classification even using validation data that were gathered with different amount and taken different places from training and testing data. CatB with its 75.35% percent of accuracy dominated the chart and followed by GB, XGB and DL later. RF was constantly the worst classifier with an accuracy below 70% almost identical to the previous assessment. The assessments of each algorithm performance were explained further through confusion matrix as Figure 5 displayed below.

Figure 5. Confusion Matrix for Each Algorithm. From upper left to the right are based on score; CatB, GB, XGB, DL, and RF
Summary of accuracy for each algorithm was calculated by adding between TN (True Negative) and TP (True Positive) while loss accuracy or Type I and II error was discovered by adding between FN (False Negative) and FP (False Positive). True Negative scoring was defined by how many class 0 from model prediction was correctly predicted based on class from validation data or field survey while True Positive was the number of correctly predicted of class 1.

| No. | Classifier              | True Positive + True Negative | Type I error + Type II error |
|-----|-------------------------|-------------------------------|-----------------------------|
| 1.  | Random Forest (RF)      | 70%                           | 30%                         |
| 2.  | Gradient Boosting (GB)  | 79%                           | 21%                         |
| 3.  | Extreme Gradient Boosting (XGB) | 77%                         | 23%                         |
| 4.  | Categorical Boosting (CatB) | **82%**                       | **18%**                      |
| 5.  | Deep Learning (DL)      | 74%                           | 26%                         |

Related to confusion matrix in attempt to examine how good the model is, another procedure was needed to get a better view about model performance. Log ROC (Receiver Operating Curve) is a graphical curve to explain model ability to discriminate between classes, in this case, binary classes, 0 and 1. This graph has x and y axes from True Positive Rate (TPR) and False Positive Rate (FPR). TPR is a rate of how many positive results are correctly predicted towards positive data sample during test meanwhile FPR is a rate of how many misclassified of positive data towards negative sample data during the test. The closer the curve to reach value 1 in the plot, in the upper left, the better the model performs to separate classes. For this research, All ML and DL algorithms used for peatland classification were particularly regarded as good model as all models showed ROC curve above red line that represents baseline accuracy. The highest score was surprisingly produced by DL with 0.76, surpassed CatB with 0.74. XGB and GB respectively shared a similar score with 0.73. RF was still with the lowest score of 0.69. Even though RF was seen as the least effective model it was still considered as a good model considering how its ROC curve bent and its metric accuracy score.

Figure 6. Log ROC Curve for Each Algorithm Ordered from Based on Score. Upper Left to Lower Right; DL, CatB, GB, XGB and RF
As seen on table 3, another parameter to analyze how dependable a model is by comparing the value of recall, precision and specificity. TPR is recall, TNR or True Negative Rate is specificity, while PPV or Positive Predictive Value is precision. Combination of high recall and high precision meant that the model was able to describe and predict both class 1 and 0 properly, not making major mistakes in predicting positive values while high specificity indicated how good the model removed false alarms.

For this peatland classification, CatB was once again the best performing model to according to recall, precision and specificity.

5. Conclusion
Performing classification to delineate peatland was possible by utilizing ML and DL algorithms. The knowledge feed to build model was supplied by passive and active remote sensing data from Landsat 8, MODIS and DEM. The best model to identify peatland according to pairs of training – testing and training – validation or confusion matrix performance assessment was achieved by CatB algorithm with accuracy score 75.52% and 82.61% respectively. Meanwhile for Log ROC assessment, DL surpassed CatB in terms of separability performance with score 0.76 but for other parameters such as recall, precision and specificity, CatB was superior with score 0.78, 0.78 and 0.79 respectively. FPR of CatB was also the lowest alongside with GB. Overall, CatB was the best model to predict distribution of peatland.

Generally, the whole ML and DL models used for this research were relatively optimal models that freed from underfit or overfit conditions as each of models was already scrutinized through hyperparameter tuning process and showed accuracy above model baseline accuracy score with decent Log ROC metric as well. For further research, more training, testing and validation data from field survey are expected to be included in the model as more field data included in the model, the better the model to describe the generalization aspect of peatland. The addition of Synthetic Aperture Radar (SAR) products like L-band or C-band and its derived product such as soil moisture, dielectric constant, or even dual or full polarimetric bands into layer processing for model input is likely to have an impact in increasing the model ability to provide more accurate peatland prediction map.

| No. | Classifier                  | Log ROC | TPR  | TNR  | PPV  | FPR  |
|-----|-----------------------------|---------|------|------|------|------|
| 1.  | Random Forest (RF)          | 0.69    | 0.65 | 0.60 | 0.65 | 0.40 |
| 2.  | Gradient Boosting (GB)      | 0.73    | 0.76 | 0.78 | 0.78 | 0.22 |
| 3.  | Extreme Gradient Boosting (XGB) | 0.73   | 0.74 | 0.73 | 0.74 | 0.27 |
| 4.  | Categorical Boosting (CatB) | 0.74    | 0.78 | 0.78 | 0.79 | 0.22 |
| 5.  | Deep Learning (DL)          | 0.76    | 0.72 | 0.71 | 0.72 | 0.29 |
Figure 7. Peatland Delineation Prediction Map with Categorical Boosting (CatB) Algorithm

6. Acknowledgement
We would like to thank Regional Disaster Prevention Agency of Ogan Komering Ilir (BPBD OKI) for their assistance and guidance during field survey to gather peat sample and providing valuable information to identify peat distribution.

7. Reference
[1] Bourgeau-Chavez L, Endres SL, Graham J, Hribljan JA, Chimner R, Lillieskov E, et al. Mapping peatlands in boreal and tropical ecoregions. In: Liang, S, ed Comprehensive Remote Sensing, vol 6 Oxford, UK: Elsevier: 24–44. 2018:24-44.
[2] Joosten H, Tapio-Biström M-L, Tol S. Peatlands: guidance for climate change mitigation through conservation, rehabilitation and sustainable use: Food and Agriculture Organization of the United Nations; 2012.
[3] Sencaki DB, Muhammad DJ, Sumargana L, Gandharum L, editors. Peatland Delination Using Remote Sensing Data in Sumatera Island. 2018 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS); 2018: IEEE.
[4] Farda N, editor Multi-temporal land use mapping of coastal wetlands area using machine learning in Google earth engine. IOP Conference Series: Earth and Environmental Science; 2017: IOP Publishing.
[5] Salehi B, Mahdianpari M, Amani M, Manesh FM, Granger J, Mahdavi S, et al. A Collection of Novel Algorithms for Wetland Classification with SAR and Optical Data. Wetlands Management-Assessing Risk and Sustainable Solutions: IntechOpen; 2018.
[6] Wu J, Kutzbach L, Jager D, Wille C, Wilmking M. Evapotranspiration dynamics in a boreal peatland and its impact on the water and energy balance. Journal of Geophysical Research: Biogeosciences. 2010;115(G4).

7] Tian S, Zhang X, Tian J, Sun Q. Random forest classification of wetland landcovers from multisensor data in the arid region of Xinjiang, China. Remote Sensing. 2016;8(11):954.

[8] Zhang H, Eziz A, Xiao J, Tao S, Wang S, Tang Z, et al. High-Resolution Vegetation Mapping Using eXtreme Gradient Boosting Based on Extensive Features. Remote Sensing. 2019;11(12):1505.

[9] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6(1):27.

[10] Zhao Y, Wong ZS-Y, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. Journal of healthcare engineering. 2018;2018.