Satellite-based Mangrove species abundance estimate using Machine Learning ensemble

Hastuadi Harsa, Gathot Winarso, Kuncoro Teguh Setiawan and Wikanti Asriningrum

Remote Sensing Research Center - Research Organization for Aeronautics and Space
Indonesia National Research and Innovation Agency

E-mail: hastuadi.harsa@brin.go.id

Abstract. The mangrove ecosystem is a vital feature in a coastal area, playing a critical role in carbon sequestration beneath the soil. Carbon preservation capacity varies among different species of mangrove. Thus, by quantifying the number of mangrove species in a given area, the volume of carbon sequestered can be estimated. Satellite imagery is highly effective for gathering such data across vast territories. In this study, we present an evaluation of mangrove species abundance across a large coastal area using Landsat satellite imagery. We employed machine learning algorithms to classify species based on spectral field observation data to achieve this. These algorithms were trained individually and ensembled to enhance prediction performance. There are 466 models generated in a two-hour training phase. After assessing these models, we identified that a stacked ensemble consisting of Deep Learning, two Distributed Random Forests, a Generalized Boosting Model, a Generalized Linear Model, and Extreme Gradient Boosting algorithms has the most superior predictive accuracy. The model achieved a mean accuracy value of 95% when tested on observation data. After applying the best model to the satellite data, our results indicate that Rhizophora Apiculata and Excoecaria Agallocha are the two most abundant mangrove species in the study area, covering 17.71% (19502.37 Ha) and 10.49% (11549.79 Ha), respectively.

1. Introduction

Carbon trading serves as a climate change mitigation strategy by creating economic incentives for reducing greenhouse gas emissions [1]. The approach enables industries and countries to trade carbon credits, allowing those exceeding emission reduction targets to sell surplus credits to entities struggling to meet their targets [2, 3]. By valuing carbon emissions and facilitating their market-driven reduction, carbon trading promotes sustainable practices and contributes to overall emission reductions[4].

Mangrove forests, characterized by their unique coastal ecosystems, are known for their exceptional capacity to sequester and store carbon[5]. These ecosystems, comprising mangrove trees and associated vegetation, act as a vital carbon preservers by absorbing atmospheric carbon dioxide (CO2) and retaining it within their biomass and sediments[6]. Mangrove plants possess the ability to sequester carbon at rates several times higher than terrestrial forests, making them crucial allies in climate change mitigation efforts[7].

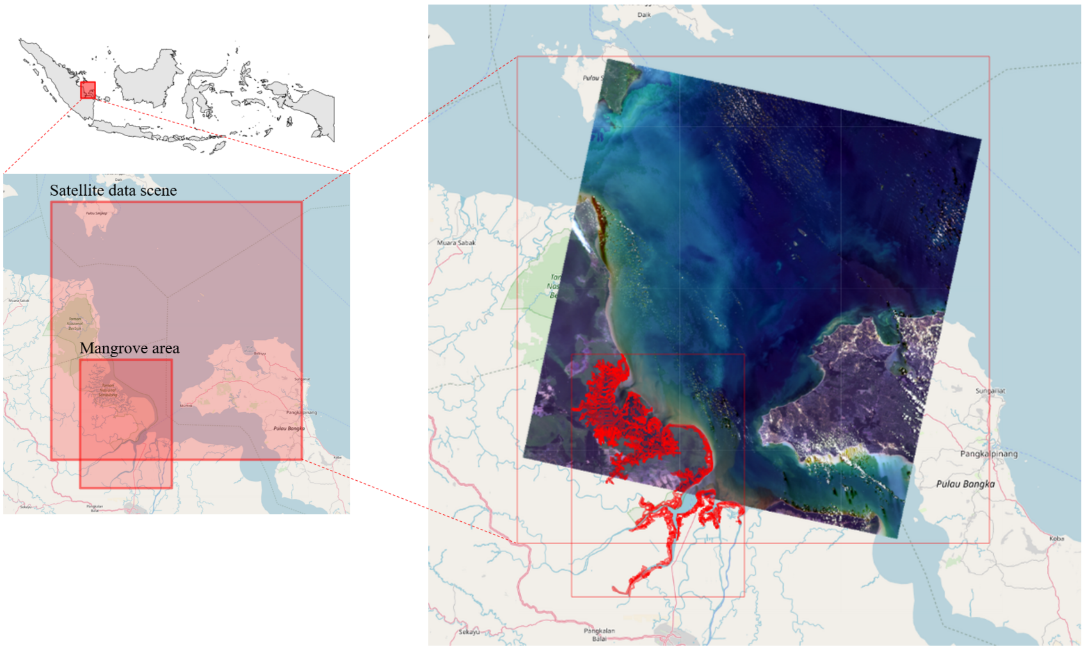

Figure 1. The study area, located in the western part of Sumatra, Indonesia, was within the intersection of the available satellite data scene and the geospatial mangrove vector area.

However, accurately estimating the carbon stock of mangrove forests requires understanding the different carbon sequestration potentials among various mangrove species[8]. Different species exhibit varying levels of biomass and different growth rates, directly impacting their carbon storage capacity[9, 10]. Hence, identifying and quantifying mangrove coverage, accounting for species-level variations, become essential for precise carbon stock assessments[11, 12, 13].

In recent years, remote sensing techniques, particularly satellite imagery, have revolutionized ecological studies by providing a cost-effective and efficient means of mapping large areas[14, 15]. Satellite imagery enables the identification and monitoring of mangrove forests at regional and global scales, facilitating comprehensive assessments of carbon stock estimates[16, 17]. Integrating satellite imagery with Machine Learning (ML) algorithms allows for advanced analysis of spectral bands to predict mangrove species, aiding in species-specific carbon stock calculations[18].

This paper describes the use of ML for predicting mangrove species from Landsat satellite imagery bands. Therefore, the carbon stock estimate can be measured subsequently using the information from the predicted mangrove species abundance.

2. Data and methods

The study area was located within the Sembilang National Park in Banyuasin Regency, along the east coast of Sumatra, Indonesia. The area covered by a red rectangle at the top-left of Figure 1shows the study area. This study collected three distinct data types: the mangrove area represented as polygons in a geospatial vector format file, field-observation data, and Landsat satellite data. The Indonesia Ministry of Environment and Forestry obtained the geospatial vector data. The data determined the boundary of the mangrove area. The field-observation data were collected within the mangrove area, employing a spectrometer instrument. The instrument displayed the spectral characteristics of mangrove leaves at selected sampling locations. Each observed mangrove leaf was analyzed across seven distinct spectral bands to unveil a comprehensive understanding of its unique attributes. The measurement values the spectrometer gave, and the observed mangrove leaf species were then recorded. Eventually, the observation data contained eight variables, i.e., seven bands of spectrums and their related species. In the measurement process, only the broadest leaf in the neighborhood was selected as the sample data, and there were 21 mangrove species identified in the study area.

The satellite data were obtained from a single Landsat Satellite imagery scene and covered some geospatial vector data. The data were represented in a geospatial raster format. The raster had a homogeneous spatial resolution of approximately 30 meters. The superposition of satellite data and the geospatial vector data are displayed at the bottom-left of Figure 1. The satellite data also contained seven bands, as in the field observation data. Both observation and satellite data bands were referred to as B1 to B7, as the abbreviation of a band and its corresponding satellite sensor. The distinction between the observed and satellite data was that the observed data contained eight variables, while the satellite data contained only seven. The eighth variable of the observation data was labeled as 'class', denoting the species of each observed mangrove leaf. A composition of Red-Green-Blue (RGB) bands of satellite raster data is displayed on the right side of Figure 1. This image illustrates a true color composition as a visible color spectrum. The RGB values were denoted by B1, B2, and B3 for Blue, Green, and Red colors, respectively. The red polygon reveals the overlapping area between the mangrove geospatial vector and satellite raster data.

The main effort in estimating the mangrove species' abundance was presenting an absent variable to the satellite data, which was present in the observation data. The absent variable was the eighth variable, i.e., class. Since the class variable was the dependent variable of each datum in the observation data, it was essential to extract the underlying interaction optimally among all band variables as the independent variables. The satellite bands interaction characterized the class variable. In ML terms, this process is known as classification[19, 20, 21]. This study employed some ML algorithms [22]to build models that extract the underlying relationship between the independent and the dependent variables of observation data.

Before building the model, all values of the independent variables in the observation data were standardized to have zero mean and a standard deviation of one. By standardizing the independent variables, all independent variables would have the same amount of contribution to the model building. In addition, the standardization procedure would exaggerate the interaction pattern among variables. The standardized observation data were then duplicated into 20 folds to provide a cross-validation element. Each fold was separated into two components: training and validation. The ratio of training and validation for all folds was 80% and 20%, respectively. The order of training and validation data were chosen randomly within each fold.

The standardized observation data were then fed to the ML algorithms as training data. The algorithms chosen in this study were Distributed Random Forest (DRF)[23], Generalized Linear Model (GLM)[24], Extreme Gradient Boosting (XGBoost)[25, 26], Generalized Boosting Machine (GBM)[27, 28], and Deep Learning (DL)[29, 30, 31]. This ML algorithm each had different model parameters. All algorithms were trained multiple times using various configurations following the needs of each algorithm's parameter. Each algorithm setup produced a single model. The models were stored for further performance comparison. A total of two-hour training sessions was designated as the maximum training time. Models with good performance were also ensembled to outperform each algorithm's best model.

One final best model was then chosen from the model's collection. This model was adopted to predict mangrove species at pixels of the satellite bands data. The satellite bands data were also standardized first as the training data. The prediction was performed only on the pixels located inside the mangrove polygons. All predicted species were then tabulated to estimate the abundance of species within the mangrove polygons.

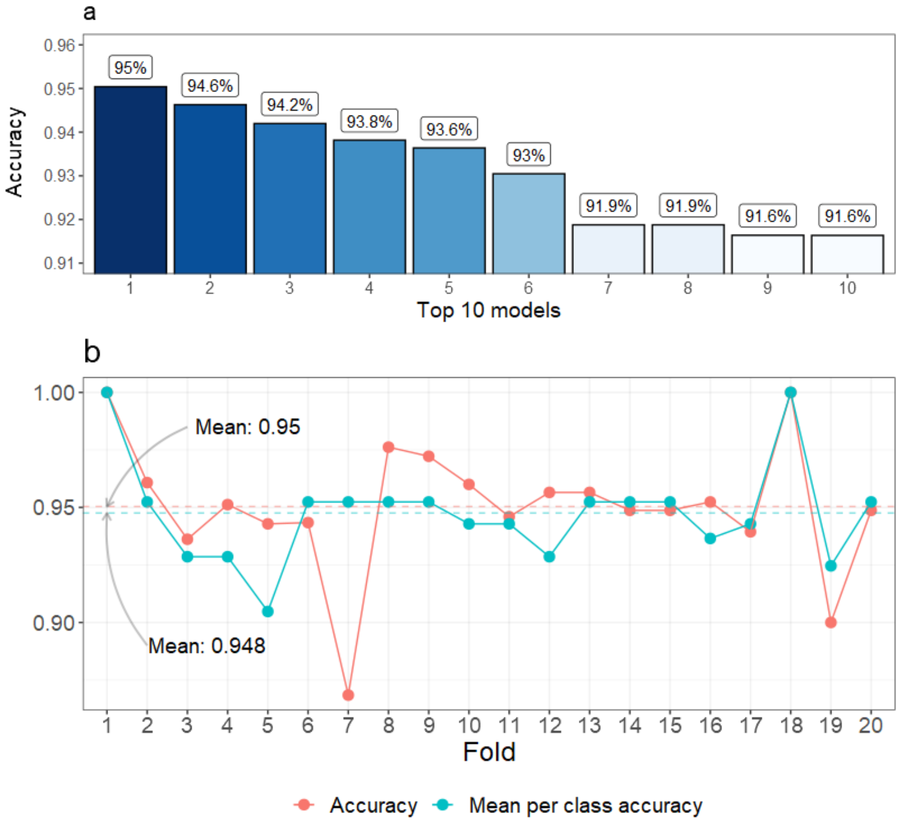

Figure 2. Model performance: (a) the top-10 best models out of 466, sorted by their accuracy, (b) best model accuracy in each fold.

3. Results and discussion

The modeling of the spectral combination value in composing the mangrove species yielded 466 models. Accuracy metrics determined these models' performance. A single accuracy value was defined as comparing the number of correct predicted species and the number of data. Since there were 20 folds of cross-validation datasets, the accuracy representing a model was determined as the mean of all accuracy of the 20 folds cross-validation dataset. Furthermore, the mean accuracy was taken as the models' sorting variable. After being sorted, the best model was a stacked ensemble model composed of six base models, i.e., DL, two DRFs, GBM, GLM, and XGBoost. The meta-learner algorithm for ensembling these base models was a GLM algorithm. The mean accuracy of the best model was 95.04%, as shown in Figure 2 (a), together with other sorted top-nine models. This accuracy value is qualitatively good. The individual accuracy of the best model for each fold is displayed in Figure 2 (b). There are two groups of accuracy values in each fold presented in Figure 2 (b), i.e., the accuracy for all species in each fold data set and the mean accuracy for each class in each fold. The accuracy for all species in each cross-validation fold ranged between 86% to 100% and 90% to 100% per species.

The spatial distribution of species classification identified by the best models in the satellite data is shown in Figure 3. Most species inhabited a clustered region. There are two important variables related to species identification, i.e., the number and the spread. The greater number of a species would result in an easier identification. On the other hand, the wider the spread of a species would result in a harder identification. If a species scattered over the area, then it must present in a great number for an easier identification. The effect of species spread is

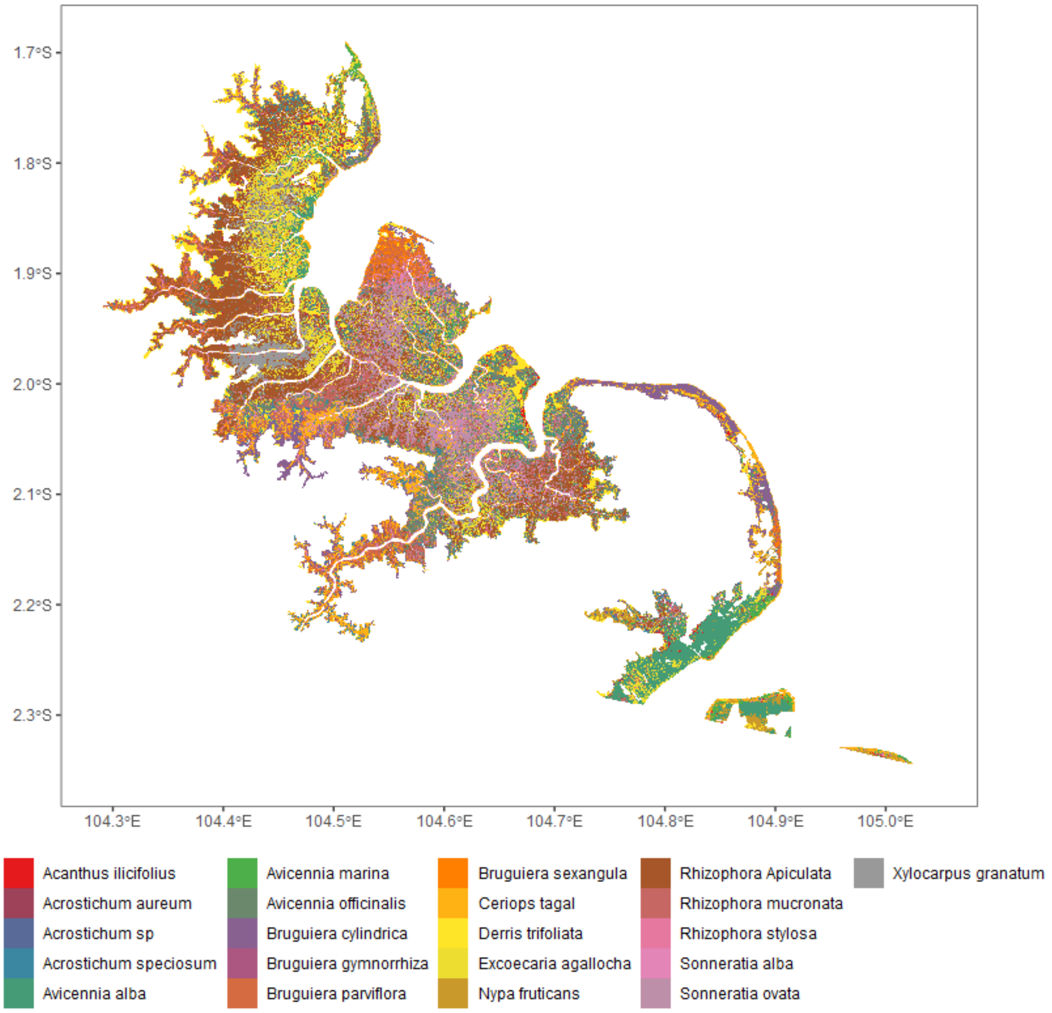

Figure 3. The best model classification output.

shown in Figure 4. The figure emphasizes the location of the five most abundant species for a better location-related analysis. Rhizophora Apiculata, Excoecaria Agallocha, and Sonneratia ovata, the top three most abundant species, are mostly concentrated at a specific location. At the same time, Avicennia Officinalis and Ceriops Tagal are more distributed over the area. A complete list of all species counts is displayed in Table 1.

In Table 1, each pixel of satellite data were tallied by its classification predicted by the model. Rhizophora Apiculata and Excoecaria Agallocha occupied 216693 and 128331 pixels, respectively, or 17.71% and 10.49% in the percentage of the total pixels. The least species is Acrostichum sp, covering 1667 pixels or 0.14% of the total pixels. Since the resolution of Landsat is 30×30 m^2, therefore the coverage area in Hectare (Ha) can be obtained by multiplying the number of pixels with 0.09, resulting in the area coverage of 19502.37 Ha for Rhizophora Apiculata as the most abundant species and 11549.79 Ha for Excoecaria Agallocha as the second most abundant species. A complete list of all species coverage is presented in Table 1.

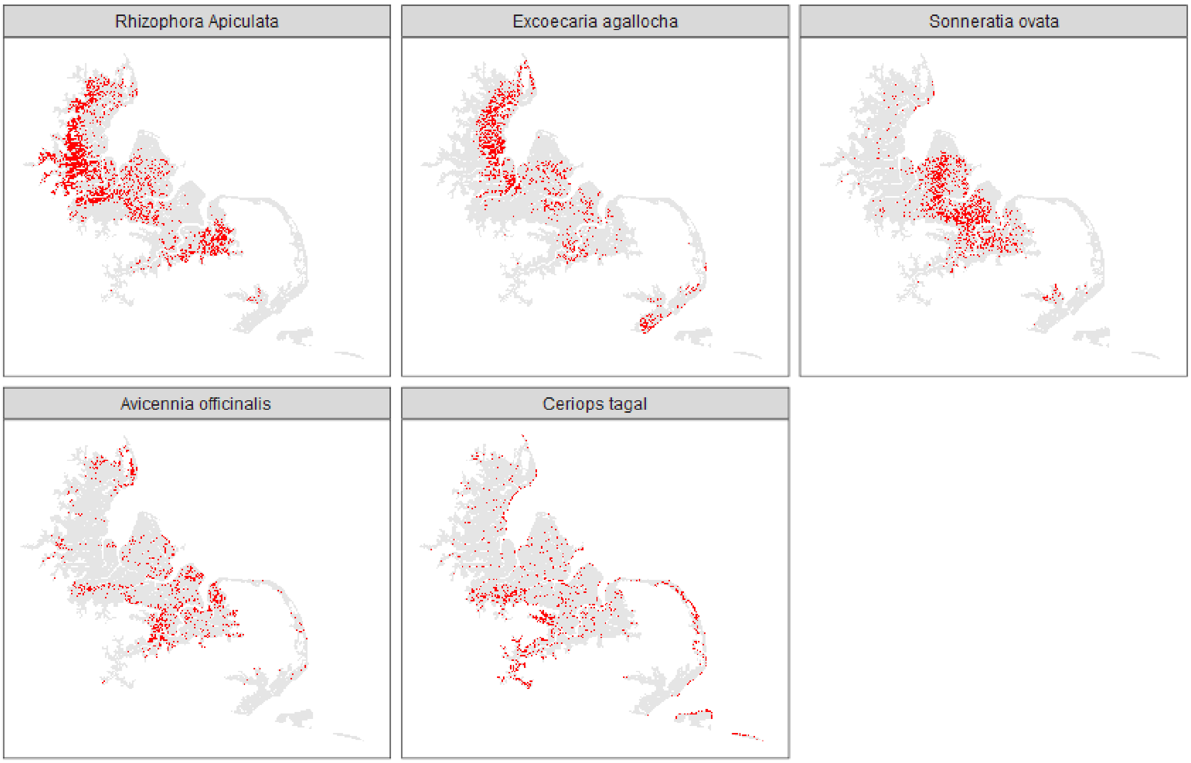

Figure 4. Distribution of five most-abundance mangrove species identified by the best model.

4. Conclusion

An estimate method of mangrove species abundance has been presented in this paper. The method applied some ML Machine Learning (ML) algorithms to provide species classification models. The algorithms used mangrove species spectrum data from field observation as their learning materials. Some best ML models were then ensembled to develop a better model. The best model was then implemented as a classifier of Landsat satellite data. Based on the classification output, the coverage of each mangrove species in the study area can be estimated. Rhizophora Apiculata and Excoecaria agallocha are the two most abundant mangrove species found. By quantifying the abundance of mangrove species, the amount of carbon sequestered can also be computed as a consideration element in carbon trading.

References

[1] Zheng Y, Tan R and Zhang B 2023 *Environmental Research Letters* **18** 045007 URL `https://doi.org/10.1088/1748-9326/acca98`

[2] Xu S, Pan W and Wen D 2023 *Sustainability* **15** 6333 URL `https://doi.org/10.3390/su15086333`

[3] Ke S, Zhang Z and Wang Y 2023 *Ecological Indicators* **148** 110054– URL `https://doi.org/10.1016/j.ecolind.2023.110054`

[4] Cetera K 2022 *Lentera Hukum* **9** 151–176 URL `https://doi.org/10.19184/ejlh.v9i1.29331`

[5] Sondak C F A, Kaligis E Y and Robert A B 2019 *Biodiversitas* **20** 978–986 URL `https://doi.org/10.13057/biodiv/d200407`

[6] Wong C J, James D, Besar N A, Kamlun K U, Tangah J, Tsuyuki S and Phua M H 2020 *Forests* **11** 1018 URL `https://doi.org/10.3390/f11091018`

[7] Nasir S, Muhammad H and Novi S A 2018 *E3S Web of Conferences* **73** 04023 URL `https://doi.org/10.1051/e3sconf/20187304023`

[8] Lassalle G and deSouza Filho C R 2022 *Remote Sensing in Ecology and Conservation* **8** 890–903 URL `https://doi.org/10.1002/rse2.289`

Table 1. Area coverage

Species	Pixel count	Percentage	Coverage (Hectare)
Rhizophora Apiculata	216693	17.71	19502.37
Excoecaria agallocha	128331	10.49	11549.79
Sonneratia ovata	121761	9.95	10958.49
Avicennia officinalis	101432	8.29	9128.88
Ceriops tagal	96629	7.90	8696.61
Avicennia alba	86897	7.10	7820.73
Derris trifoliata	71376	5.83	6423.84
Bruguiera cylindrica	68186	5.57	6136.74
Rhizophora mucronata	64220	5.25	5779.80
Xylocarpus granatum	53867	4.40	4848.03
Bruguiera parviflora	51098	4.18	4598.82
Nypa fruticans	36129	2.95	3251.61
Sonneratia alba	29928	2.45	2693.52
Acrostichum speciosum	29282	2.39	2635.38
Avicennia marina	18256	1.49	1643.04
Bruguiera sexangula	16083	1.31	1447.47
Acrostichum aureum	10411	0.85	936.99
Acanthus ilicifolius	9181	0.75	826.29
Bruguiera gymnorrhiza	6378	0.52	574.02
Rhizophora stylosa	5680	0.46	511.20
Acrostichum sp	1667	0.14	150.03

[9] Wang D, Wan B, Qiu P, Zuo Z, Wang R and Wu X 2019 *Remote Sensing* **11** 2156 URL https://doi.org/10.3390/rs11182156

[10] Navarro J A, Algeet N, FernÃ¡ndez-Landa A, Esteban J, RodrÃguez-Noriega P and GuillÃ©n-Climent M L 2019 *Remote Sensing* **11** 77 URL https://doi.org/10.3390/rs11010077

[11] Maeda Y, Fukushima A, Imai Y, Tanahashi Y, Nakama E, Ohta S, Kawazoe K and Akune N 2016 *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLI-B8** 705–709 ISSN 2194-9034 URL https://doi.org/10.5194/isprs-archives-XLI-B8-705-2016

[12] Zheng Y and Takeuchi W 2022 *Scientific Reports* **12** 1–14 URL https://doi.org/10.1038/s41598-022-06231-6

[13] Jiang X, Zhen J, Miao J, Zhao D, Shen Z, Jiang J, Gao C, Wu G and Wang J 2022 *Ecological Indicators* **140** 108978– URL https://doi.org/10.1016/j.ecolind.2022.108978

[14] Ma C, Ai B, Zhao J, Xu X and Huang W 2019 *Remote Sensing* **11** 921 URL https://doi.org/10.3390/rs11080921

[15] Atmaja T, Fukushi K and Fukushi K 2022 *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **V-3-2022** 517–523 URL https://doi.org/10.5194/isprs-annals-V-3-2022-517-2022

[16] Pham T, Yokoya N, Bui D, Yoshino K and Friess D 2019 *Remote Sensing* **11** 230 URL https://doi.org/10.3390/rs11030230

[17] Gandhi S and Jones T 2019 *Remote Sensing* **11** 728 URL https://doi.org/10.3390/rs11060728

[18] Hsu A J, Kumagai J, Favoretto F, Dorian J, Martinez B G and Aburto-Oropeza O 2020 *Remote Sensing* **12** 3986 URL https://doi.org/10.3390/rs12233986

[19] Boehmke B and Greenwell B M 2020 *Hands-on Machine Learning with R, 1st edition* (Chapman and Hall/CRC) ISBN 9781138495685 URL https://bradleyboehmke.github.io/HOML/

[20] Molnar C 2023 *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable* (Independently Published) ISBN 979-8411463330 URL https://christophm.github.io/interpretable-ml-book/

[21] Scott M 2023 *Machine Learning: Unsupervised and Supervised Learning* URL `https://bookdown.org/content/f097ddae-23f5-4b2d-b360-ad412a6ca36a/`

[22] LeDell E and Poirier S 2020 *7th ICML Workshop on Automated Machine Learning (AutoML)*

[23] Geurts P, Ernst D and Wehenkel L 2006 *Machine Learning* **63** 3–42 URL `https://doi.org/10.1007/s10994-006-6226-1`

[24] Dunn P K and Smyth G K 2005 *Statistics and Computing* **15** 267–280 URL `https://doi.org/10.1007/s11222-005-4070-y`

[25] Mitchell R and Frank E 2017 *PeerJ Preprints* URL `https://doi.org/10.7287/peerj.preprints.2911v1`

[26] Chen T and Guestrin C 2016 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* URL `https://doi.org/10.1145/2939672.2939785`

[27] Click C, Lanford J, Malohlava M, Parmar V and Roark H 2015 *Gradient Boosted Machines with H2O* (H2O.ai, Inc.)

[28] Malohlava M, Candel A, Click C, Roark H and Parmar V 2023 *Gradient Boosting Machine with H2O* (H2O.ai, Inc.)

[29] Candel A and Parmar V 2015 *Deep Learning with H2O* (2307 Leghorn StreetMountain View, CA 94043: H2O.ai, Inc.)

[30] Ghorbani M, Salmasi F, Saggi M, Bhatia A, Kahya E and Norouzi R 2020 *Journal of Hydroinformatics* **22** 1603–1619 URL `https://doi.org/10.2166/hydro.2020.003`

[31] Elsayad A S, Desouky A I E, Salem M and Badawy M 2020 *IEEE Access* **8** 97231–97242 URL `https://doi.org/10.1109/ACCESS.2020.2995790`