A COMPARISON OF TREE-BASED ALGORITHMS FOR COMPLEX WETLAND CLASSIFICATION USING THE GOOGLE EARTH ENGINE

Ali Jamali [1*], Masoud Mahdianpari [2,3], and İsmail Rakıp KARAŞ [4]

[1] Karabük University, Department of Civil Engineering, Faculty of Engineering, Karabük, Turkey, alijamali@karabuk.edu.tr;
[2] Department of Electrical and Computer Engineering, Memorial University of Newfoundland, St. John's, NL A1B3X5, Canada;
[3] C-CORE, 1 Morrissey Rd, St. John's, NL A1B 3X5, Canada, m.mahdianpari@mun.ca;
[4] Karabük University, Department of Computer Engineering, Faculty of Engineering, Karabük, Turkey, ismail.karas@karabuk.edu.tr

KEY WORDS: Wetland Mapping, Big data, Sentinel Imagery, Decision Tree, Random Forest, Extreme Gradient Boosting, Google Earth Engine

ABSTRACT:

Wetlands are endangered ecosystems that are required to be systematically monitored. Wetlands have significant contributions to the well-being of human-being, fauna, and fungi. They provide vital services, including water storage, carbon sequestration, food security, and protecting the shorelines from floods. Remote sensing is preferred over the other conventional earth observation methods such as field surveying. It provides the necessary tools for the systematic and standardized method of large-scale wetland mapping. On the other hand, new cloud computing technologies for the storage and processing of large-scale remote sensing big data such as the Google Earth Engine (GEE) have emerged. As such, for the complex wetland classification in the pilot site of the Avalon, Newfoundland, Canada, we compare the results of three tree-based classifiers of the Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB) available in the GEE code editor using Sentinel-2 images. Based on the results, the XGB classifier with an overall accuracy of 82.58% outperformed the RF (82.52%) and DT (77.62%) classifiers.

1. INTRODUCTION

Wetlands that cover around 3% to 8% of the earth's surface are among the most valuable yet in danger ecosystems. These endangered ecosystems provide significant contributions to the well-being of human-being, as well as the natural resources (Slagter et al., 2020). Wetlands are regarded as the kidney of the earth because bacteria, animals, and plants that are living in wetlands filter the water by trapping nutrients such as phosphorus, which causes the harmful algae blooming in the water bodies (Mahdavi et al., 2018; Tiner, 2015). Services provided by wetlands include water storage, food security, carbon sequestration, shoreline, and flood protection (Board, 2005; Davidson, 2016). As such, developing and proposing new technologies for the systematic and standardized monitoring of these vital ecosystems are essential (Jamali et al., 2021a, 2021b).

For efficient large-scale monitoring of wetlands, remote sensing has been suggested over the conventional techniques such as field surveying as it is regarded as the leading technology for systematic and standardized mapping and monitoring of the earth's surface. Compared to conventional Land Use Land Cover (LULC) mapping, wetland ecosystems' inherent biological and ecological characteristics are among the most complex ecosystems to be mapped. For example, they are not categorized by a specific type of vegetation or land cover; rather, they are united by the amount of water below the vegetation canopy, below, at, or near the surface of earth's ground (Slagter et al., 2020). Moreover, due to the complexity of wetlands in terms of vegetation composition, position, and shape, satellite sensors' capacity for their classification is often insufficient. Consequently, different conventional and advanced machine learning methods for complex wetland classification are developed and proposed (Mahdianpari et al., 2019).

On the other hand, remote sensing has several intrinsic and extrinsic characteristics of big data. The intrinsic characteristics are its dynamic state (i.e., the earth's surface changes continuously), multi-scale (e.g., its spectral range, time interval, resolution, angle, and polarization), and non-linear features (i.e., time-series data are often non-linear and noisy). The extrinsic characteristics of remote sensing can be defined by its multi-source, high-dimensional, and isomer characteristics (Tamiminia et al., 2020). As such several cloud computing platforms, including the Google Earth Engine (GEE) and Sentinel Hub are developed and proposed to address the challenges regarding the geo big data of remote sensing. Specifically, the GEE provides a free-of-charge infrastructure, storage, platform, and software for the processing of large-scale remote sensing images. For instance, the Data Catalog of the GEE contains massive remote sensing data, including Landsat, Sentinel, and MODIS series, as well as the high-resolution images of the US National Agriculture Imagery Program (NAIP). Moreover, the GEE code editor that uses the javascript programming language can be used for the processing of the data provided by the GEE.

There are few studies on the comparison of classifiers provided by the GEE code editor. As such, we compare the performance of three available tree-based classifiers of the Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB) in the GEE for the classification of complex wetlands in the Avalon pilot site.

2. METHODS

2.1. STUDY AREA AND REMOTE SENSING DATA

The pilot site is the Avalon, situated in the very eastern portion of Newfoundland, Canada (Figure 1). Wetland habitat and other natural ecosystems, including bog, fen, marsh, swamp, and shallow water, are present in the Avalon. The peatlands (i.e., bog and fen) are the most dominant classes in the pilot site. It is worth mentioning that the ground truth data were collected in the summers of 2015 to 2017 by a group of wetland biologists familiar with the study area. We used the GEE for the processing and classification of the Sentinel-2 surface reflectance of the pilot site of the Avalon. The median values of Sentinel-2 images from 1st June 2021 to 1st July 2021 were used. The selected bands and spectral indices, including the Normalized Difference Vegetation Index (NDVI), Normalized Difference Built-up Index (NDBI), and the Modified Normalized Difference Water Index (MNDWI), are shown in Table 1.

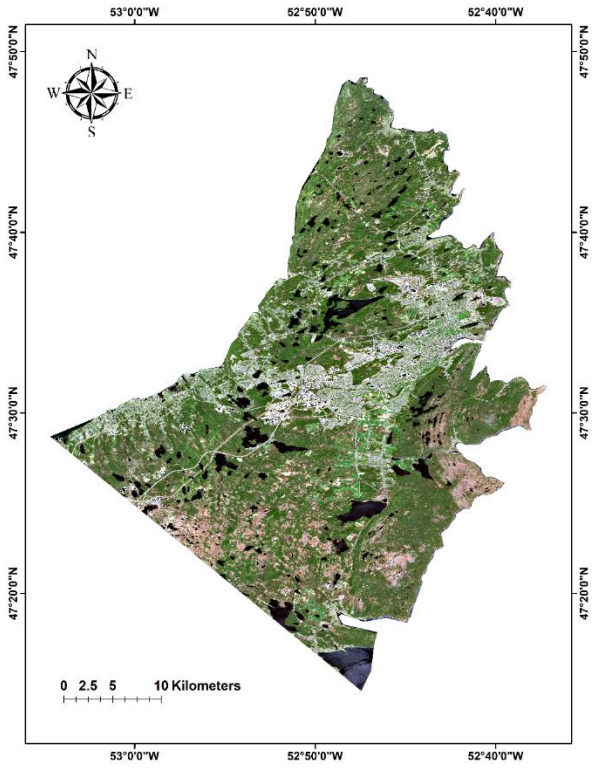

Figure 1. The location of the study area of the Avalon (RGB composite of Sentinel-2 images).

Bands	Spectral Indices
B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12	$NDVI = \frac{B8 - B4}{B8 + B4}$
	$NDBI = \frac{B11 - B8}{B11 + B8}$
	$MNDWI = \frac{B3 - B11}{B3 + B11}$

Table 1. Sentinel-2 bands and spectral indices used in this study in the Avalon pilot site.

The number of training and test pixels is shown in Table 2. It is worth mentioning that we used random sampling where ground truth data was divided into 50% as training and 50% as test data.

Class	Number of Training Pixels	Number of Test Pixels
Bog	3077	3077
Fen	1637	1637
Marsh	867	867
Swamp	1164	1164
Shallow Water	1351	1351
Urban	4866	4866
Deep Water	3369	3369
Upland	9343	9343

Table 2. Training and test pixel sample for the pilot site of the Avalon.

It is worth mentioning that, in the GEE code editor, there is no need for band resampling that has a different spatial resolution. In other words, the band resampling is done by the GEE platform without any extra coding.

2.2. Decision Tree

DT is a tree-based machine learning algorithm where each node represents a feature, each link (branch) represents a decision (rule), and each leaf node represents an outcome (output class) (see Figure 2; Li and Tang, 2020). Probability analysis is used in this graphical method (Yang et al., 2020). A pattern among attributes and values of features is found in DT models. For example, a recursive top to bottom rule procedure along with a dividing technique is used in DT for image classification. Comprehensibility, simplicity from the mathematical point of view, and the ability to handle mixed-type data in regression and classification tasks are the main advantages of the DT classifier (Su & Zhang, 2006). Despite these benefits, the final solution may not be the optimal model, and overfitting may arise when using one tree as the predictor (Maxwell et al., 2018).

2.3. Random Forest

To overcome problems associated with a single DT, an ensemble tree-based classifier called RF was proposed wherein several DTs are used to reach an optimal global solution (Breiman, 2001). RF is computationally efficient and resistant to the noise in the sample data. In the RF classifier, using the input data (vector), each DT is built by random vectors that are sampled independently. Additionally, a voting procedure is used to find the most popular prediction among various DTs for a given vector data. For instance, majority voting or average values are used to integrate the results of several different DTs to reach an optimal model (Jamali, 2021a, 2021b, 2020; Jamali et al., 2021c; Moayedi et al., 2020).

2.4. Extreme Gradient Boosting

XGB is an advanced ensemble shallow classifier that is constructed based on the gradient boosting framework (Tianqi and Guestrin, 2016). The XGB algorithm can be defined as an improved gradient boosting decision tree (GBDT). It is worth highlighting that by introducing the regularization term into the objective function, the XGB classifier outperforms the GBDT technique in terms of avoiding the overfitting issue. Moreover, the XGB algorithm has a rather more simplified structure compared to the GBDT method. In the objective function of the XGB algorithm, the second-order Taylor expansion is applied. In addition, it simultaneously uses the first-order and second-order derivatives, making the XGB algorithm more accurate and faster than its descent gradient technique (Li et al., 2021, 2020). Based on the previous research, the XGB classifier had better performance compared to the other classifiers such as RF and DT (Buthelezi et al., 2020; Wei and Hsu, 2020a, 2020b).

2.5. Accuracy assessment

The classification results are evaluated in terms of mean overall accuracy (Equation 1).

$$Accuracy = \frac{(True\ positive + True\ negative)}{Total\ number\ of\ pixels} \times 100 \qquad (1)$$

3. RESULTS AND DISCUSSION

Complex wetlands and non-wetlands of the pilot site of the Avalon were classified using three tree-based classifiers of the DT, RF, and XGB in the GEE code editor. The XGB (82.58%) classifier had the best performance over the other two tree-based classifiers of the DT (77.62%) and RF (82.52%) in terms of overall accuracy. Moreover, we evaluated the performance of the RF and XGB classifiers for a different number of trees as well (Figure 2).

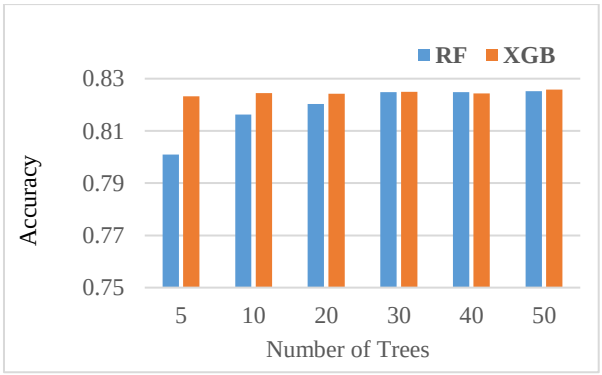

Figure 2. Performance of the Random Forest and Extreme Gradient Boosting for different numbers of trees.

Based on the results, the XGB algorithm slightly outperformed the RF classifier for the complex wetland classification of the pilot site of the Avalon. The best overall accuracy for the RF (82.52%) and XGB (82.58%) classifiers was obtained where we set the number of trees to 50. The XGB classifier had more consistent classification results while increasing the number of trees. On the other hand, increasing the number of trees had a more significant effect on the accuracy of the wetland classification using the RF classifier compared to the XGB algorithm. The reason can be explained by their different method of ensembling. The RF is a bagging method in which each tree is constructed independently, and at the end of the training process, trees are ensembled, while the XGB classifier is a boosting method building one tree at a time.

While implementing the available tree-based classifiers of the DT, RF, and XGB in the GEE code editor, we encountered several issues. The scale parameter was an important factor for the classification of the complex wetlands of the study area of the Avalon. For instance, we could use smaller scale parameters for the DT and RF classifiers, while we encountered an error of "Computed value is too large." while using the XGB algorithm. The reason can be explained by the higher complexity and computation cost of the XGB classifier compared to the other algorithms of the DT and RF. Another issue was the implementation of the DT classifier, as there was not an example or reference provided by the GEE. Implementation of the DT classifier is shown in Table 3 as a reference for future research.

```
var classifier = ee.Classifier.smileCart().train({
  features: trainingdata,
  classProperty: 'Classes',
  inputProperties: Avalon.bandNames()
});
var treeString = classifier.explain().get('tree');
print(treeString)
```

```
var DT = ee.Classifier.decisionTree(treeString);
```

Table 3. Implementation of the DT in the GEE code editor.

Maps and confusion matrices of the classified landscape of the study area of the Avalon using the tree-based classifiers of the DT, RF, and XGB implemented in the GEE code editor are shown in Figure 3 and Table 4. While the XGB classifier had better performance over the RF and DT classifier, its computation in the GEE was slow, and we needed to use a bigger scale parameter to avoid the error of the high computation cost.

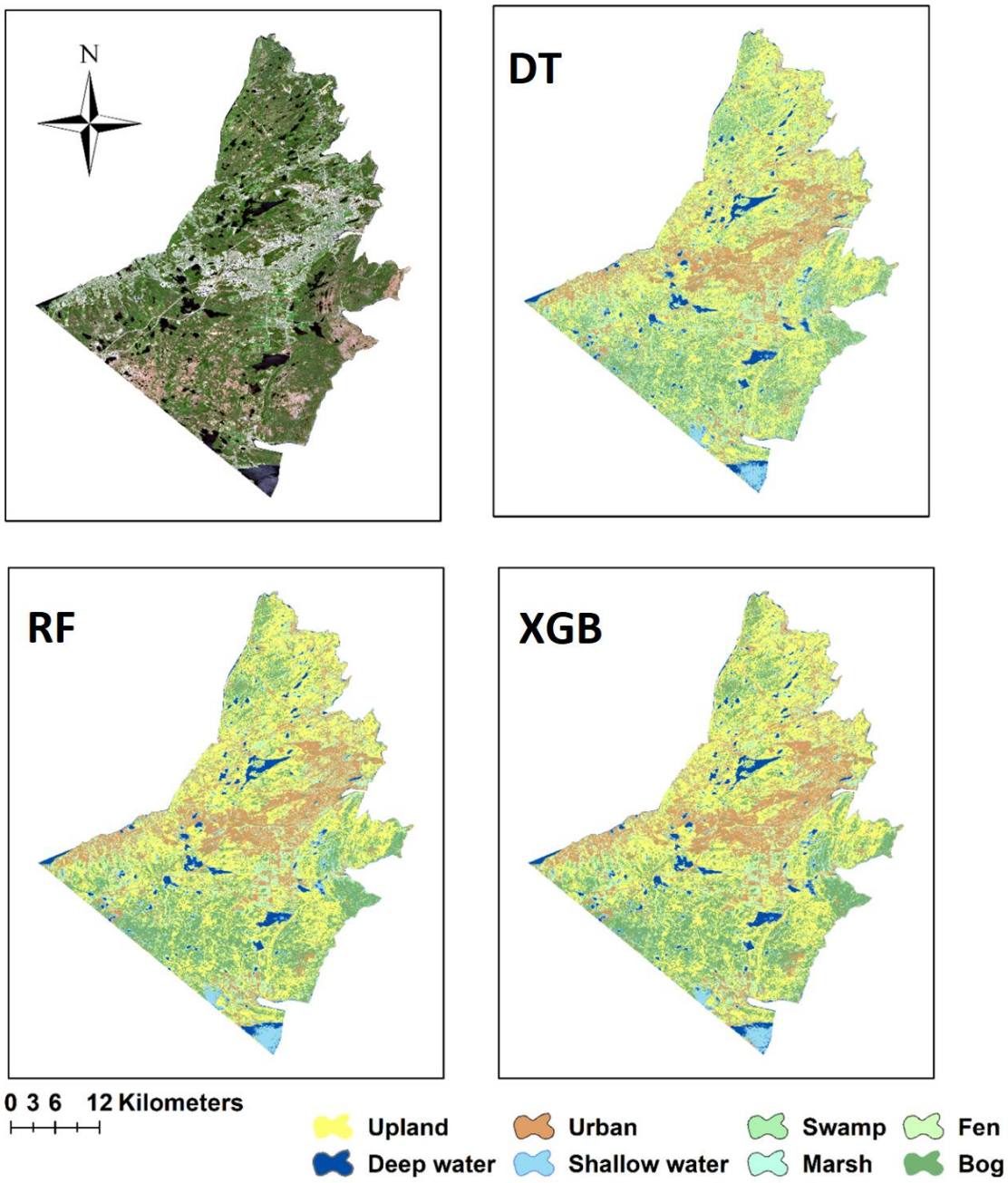

Figure 3. Wetland classification maps of the Avalon pilot site using the Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB) in Google Earth Engine.

DT								
	Bog	**Fen**	**Marsh**	**Swamp**	**Shallow water**	**Urban**	**Deep water**	**Upland**
Bog	2155	708	71	31	27	26	0	59

Fen	608	420	166	128	7	46	0	262
Marsh	71	145	289	65	149	50	0	98
Swamp	304	192	139	175	4	17	0	333
Shallow Water	4	3	84	0	1038	0	220	2
Urban	30	32	10	3	4	4752	0	35
Deep water	0	0	0	0	98	0	3271	0
Upland	339	258	361	281	19	257	0	7828
RF								
	Bog	**Fen**	**Marsh**	**Swamp**	**Shallow water**	**Urban**	**Deep water**	**Upland**
Bog	2722	269	24	3	17	4	0	38
Fen	819	301	90	114	0	21	0	292
Marsh	78	130	327	52	126	30	1	123
Swamp	354	114	90	132	0	1	0	473
Shallow Water	4	0	83	0	1064	1	197	2
Urban	16	7	4	0	0	4829	0	10
Deep water	0	0	0	0	41	0	3328	0
Upland	307	72	85	77	1	269	0	8532
XGB								
	Bog	**Fen**	**Marsh**	**Swamp**	**Shallow water**	**Urban**	**Deep water**	**Upland**
Bog	2657	320	26	6	17	8	0	43
Fen	780	315	77	142	0	26	0	297
Marsh	66	119	307	60	131	41	1	142
Swamp	324	127	85	162	0	4	0	462
Shallow Water	4	0	72	0	1083	3	187	2
Urban	22	6	6	0	0	4821	0	11
Deep water	0	0	0	0	26	0	3343	0
Upland	270	87	90	95	2	285	0	8514

Table 4. Confusion matrices of the wetland classification using the Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB) in Google Earth Engine.

4. CONCLUSIONS

Endangered ecosystems of wetlands provide vital services such as food security, water filtering, and protecting the shorelines from floods are important to be systematically monitored. As the leading technology in global earth observations, remote sensing provides the essential tools and capabilities of the standardized and systematic monitoring and classification of complex wetlands. On the other hand, as remote sensing has the intrinsic and extrinsic characteristics of big data, cloud computing platforms such as the GEE provide the essential tools and infrastructure to address big remote sensing data challenges. As such, in this study, we implemented three tree-based classifiers of the DT, RF, and XGB in the GEE code editor. Based on the results, the XGB classifier with an overall accuracy of 82.58% slightly outperformed the RF algorithm (82.52%) for the complex wetland classification of the pilot site of the Avalon. Besides, the DT had the least performance with an overall accuracy of 77.62%.

REFERENCES

Board, M.A., 2005. Millennium ecosystem assessment. Washington, DC.

Breiman, L., 2001. Random Forests. Machine Learning 54, 5–32.

Buthelezi, M.N.M., Lottering, R.T., Hlatshwayo, S.T., Peerbhay, K., 2020. Comparing rotation forests and extreme gradient boosting for monitoring drought damage on KwaZulu-Natal commercial forests. null 1–24. https://doi.org/10.1080/10106049.2020.1852612

Davidson, N.C., 2016. The Ramsar Convention on Wetlands, in: The Wetland Book I: Structure and Function, Management and Methods. Springer Publishers, Dordrecht.

Jamali, A., 2021a. Improving land use land cover mapping of a neural network with three optimizers of multi-verse optimizer, genetic algorithm, and derivative-free function. The Egyptian Journal of Remote Sensing and Space Science 24, 373–390. https://doi.org/10.1016/j.ejrs.2020.07.001

Jamali, A., 2021b. Land use land cover modeling using optimized machine learning classifiers: a case study of Shiraz, Iran. Modeling Earth Systems and Environment 7, 1539–1550. https://doi.org/10.1007/s40808-020-00859-x

Jamali, A., 2020. Land use land cover mapping using advanced machine learning classifiers: A case study of Shiraz city, Iran. Earth Science Informatics 13, 1015–1030. https://doi.org/10.1007/s12145-020-00475-4

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., Salehi, B., 2021a. Comparing Solo Versus Ensemble Convolutional Neural Networks for Wetland Classification Using Multi-Spectral Satellite Imagery. Remote Sensing 13, 2046. https://doi.org/10.3390/rs13112046

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., Salehi, B., 2021b. Wetland Mapping Using Multi-Spectral Satellite Imagery and Deep Convolutional Neural Networks: A Case Study in Newfoundland and Labrador, Canada. null 47, 243–260. https://doi.org/10.1080/07038992.2021.1901562

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., Salehi, B., 2021c. Deep Forest classifier for wetland mapping using the combination of Sentinel-1 and Sentinel-2 data. null 1–18. https://doi.org/10.1080/15481603.2021.1965399

Li, C., Zhou, L., Xu, W., 2021. Estimating Aboveground Biomass Using Sentinel-2 MSI Data and Ensemble Algorithms for Grassland in the Shengjin Lake Wetland, China. Remote Sensing 13. https://doi.org/10.3390/rs13081595

Li, S., Tang, H., 2020. Classification of Building Damage Triggered by Earthquakes Using Decision Tree. Mathematical Problems in Engineering, 2020. https://doi.org/10.1155/2020/2930515

Li, Y., Li, M., Li, C., Liu, Z., 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Scientific Reports 10, 9952. https://doi.org/10.1038/s41598-020-67024-3

Mahdavi, S., Salehi, B., Granger, J., Amani, M., Brisco, B., Huang, W., 2018. Remote sensing for wetland classification: A comprehensive review. GIScience & Remote Sensing 55, 623–658.

Mahdianpari, M., Salehi, B., Mohammadimanesh, F., Homayouni, S., Gill, E., 2019. The first wetland inventory map of newfoundland at a spatial resolution of 10 m using sentinel-1 and sentinel-2 data on the google earth engine cloud computing platform. Remote Sensing 11.

Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing 39, 2784–2817.

Moayedi, H., Jamali, A., Gibril, M.B.A., Kok Foong, L., Bahiraei, M., 2020. Evaluation of tree-base data mining algorithms in land used/land cover mapping in a semi-arid environment through Landsat 8 OLI image; Shiraz, Iran. Geomatics, Natural Hazards and Risk 11, 724–741.

Slagter, B., Tsendbazar, N.E., Vollrath, A., Reiche, J., 2020. Mapping wetland characteristics using temporally dense Sentinel-1 and Sentinel-2 data: A case study in the St. Lucia wetlands, South Africa. International Journal of Applied Earth Observation and Geoinformation 86.

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B., 2020. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS Journal of Photogrammetry and Remote Sensing 164, 152–170. https://doi.org/10.1016/j.isprsjprs.2020.04.001

Tianqi, C., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Tiner, R.W., 2015. Wetlands: An overview, in: R. W., T., M. W., L., V. V., K. (Eds.), Remote Sensing of Wetlands: Applications and Advances. CRC Press, Boca Raton, FL, pp. 20–35.

Wei, C.-C., Hsu, C.-C., 2020a. Extreme Gradient Boosting Model for Rain Retrieval using Radar Reflectivity from Various Elevation Angles. Remote Sensing 12. https://doi.org/10.3390/rs12142203

Wei, C.-C., Hsu, C.-C., 2020b. Extreme Gradient Boosting Model for Rain Retrieval using Radar Reflectivity from Various Elevation Angles. Remote Sensing 12. https://doi.org/10.3390/rs12142203

Yang, H., Wang, J., Su, B., 2020. Fast processing method of high resolution remote sensing image based on decision tree classification. Aerospace and Electronics 1, 1–11.