International Journal of Engineering and Geosciences
https://dergipark.org.tr/en/pub/ijeg
e-ISSN 2548-0960

Comparison between random forest and support vector machine algorithms for LULC classification

Cengiz Avci [*1], **Muhammed Budak**[1], **Nur Yagmur**[1], **Filiz Bektas Balcik**[1]

[1]*Istanbul Technical University, Department of Geomatics Engineering, Türkiye*

Keywords
Remote sensing
Supervised Classification
Random Forest
Support Vector Machine
Wetland

Research Article
DOI: 10.26833/ijeg.987605

Received: 26.08.2021
Accepted: 30.11.2021
Published: 13.04.2022

Abstract

Nowadays, machine learning (ML) algorithms have been widely chosen for classifying satellite images for mapping Earth's surface. Support Vector Machine (SVM) and Random Forest (RF) stand out among these algorithms with their accurate results in the literature. The aim of this study is to analyze the performances of these algorithms on land use and land cover (LULC) classification, especially wetlands which have significant ecological functions. For this purpose, Sentinel-2 satellite image, which is freely provided by European Space Agency (ESA), was used to monitor not only the open surface water body but also around Marmara Lake. The performance evaluation was made with the increasing number of the training dataset. 3 different training datasets having 10, 15, and 20 areas of interest (AOI) per class, respectively were used for the classification of the satellite images acquired in 2015 and 2020. The most accurate results were obtained from the classification with RF algorithm and 20 AOIs. According to obtained results, the change detection analysis of Marmara Lake was investigated for possible reasons. Whereas the water body and wetland have decreased more than 50% between 2015 and 2020, crop sites have increased approximately 50%.

1. Introduction

In remote sensing, land use and land cover (LULC) maps are hugely demanded for land management and monitoring of natural resources such as wetlands, forest areas, grasslands [1]. The most significant way to produce LULC maps is to classify remotely sensed images using a variety of classification algorithms [2]. LULC maps, which are used in different fields such as agricultural tracking and city planning, are extremely important in sustainable monitoring and analyzing different ecosystems [3-4]. For the long-term change analysis generally, Landsat satellite images have been used because of providing free satellite images since the 1970s [5]. However, Sentinel-2 MSI satellite images have gained importance in the short-term change analysis [6]. With having 13 spectral band capacity, high spatial resolution between 10-60 meters, and having free access on a global scale, Sentinel-2 MSI provides a great advantage for researchers.

Many different classification methods have been developed and conducted such as unsupervised and supervised classification methods. With the development of machine learning (ML) algorithms, the accuracy and usage of classification methods have been increased. There are multiple ML algorithms used for the LULC classification, but random forest (RF) and support vector machine (SVM) algorithms have become prominent in the last years and have been widely used for that purpose [7-8]. These methods provide superior performance compared to traditional methods, especially in remote sensing applications [9]. However, these algorithms do not give the same results with different satellite data. The quality, number, and distribution of the training and test data sets have huge importance for the performance of the selected algorithm. It is important to compare applications to measure the performance of these algorithms based on the number of training and test points.

Google Earth Engine (GEE) has gained popularity in recent years. GEE makes geospatial analysis easy, enables processing satellite images using mathematical operations and many classification algorithms implemented. GEE provides to make global scale [10] and

* **Corresponding Author**

*(avcice16@itu.edu.tr) ORCID ID 0000-0002-6515-1059
(budakm16@itu.edu.tr) ORCID ID 0000-0003-4493-9936
(yagmurn@itu.edu.tr) ORCID ID 0000-0002-5915-6929
(bektasfi@itu.edu.tr) ORCID ID 0000-0003-3039-6846

Cite this article

Avcı, C. Budak, M. Yagmur, N., & Balcık, F. B. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. International Journal of Engineering and Geosciences, 8(1), 01-10

local scale [11] water bodies and wetland analysis. Wang et al. [12] analyzed long-term surface water dynamics of the Yangtze River Basin using RF classification algorithm via GEE and Amani et al. [13] used GEE to create the wetland inventory of Canada by applying RF algorithm to satellite images.

The main objective of the study is the investigate the performance of machine learning algorithms in LULC classification of wetland by increasing the amount of the training dataset. For this purpose, RF and SVM algorithms were selected and applied to Sentinel-2 satellite images via the GEE platform. In the study, Marmara Lake, which is considered as one of the Nationally Important Wetlands in Turkey, was examined between the years of 2015 and 2020 as a test site and spatio-temporal analysis of the lake and its surroundings was investigated with the most accurate classification algorithm for short term change analysis.

2. Study area

In the study, classification performance analysis and LULC change analysis were applied on Marmara Lake. Marmara Lake is located in the Aegean Region, within the borders of Manisa province, between Salihli, Gölmarmara, and Ahmetli districts. Its depth is 3-4 m, and its altitude is 74 m. The lake, which was included in the wetland category according to the Regulation on the Conservation of Wetlands published in 2002, became a Nationally Important Wetland in 2017. Marmara Lake, which is economically important for the local community with fishing and tourism activities, has 162 bird species, 32 mammal species, and 355 different plant species in its ecosystem according to the report prepared in 2017 [14].

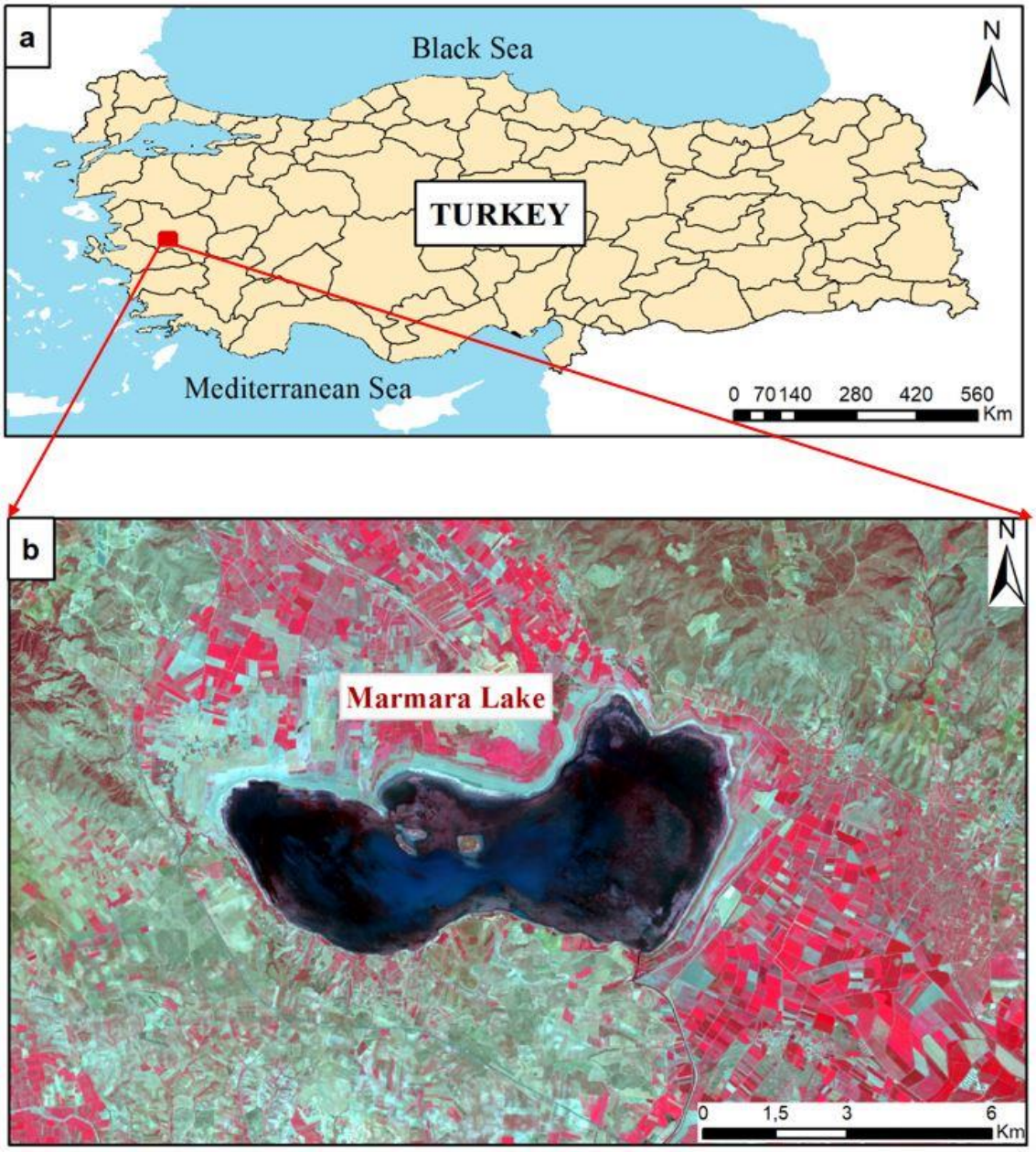

Figure 1. Study area

3. Dataset

Sentinel-2 MSI satellite images were used in the study for 2015 and 2020 years. It is freely available data and provides images since 2015. The Sentinel-2A satellite is a multispectral image with a medium spatial resolution developed by ESA. It has 13 spectral bands and the spatial resolution of them varies between 10 to 60 m. It also provides atmospherically corrected images. The detailed information is given in Table 1. In this study, satellite images acquired in the summer months were selected to use cloud-free images.

Table 1. Sentinel-2A MSI Information

Data	Sentinel-2A MSI
Spatial Resolution	B2-4, B8: 10 m B5-7, B8A, B11-12: 20 m B1, B9: 60 m
Spectral Resolution	B1: 443.9 nm, B2: 496.6 nm, B3: 560 nm, B4: 664.5 nm, B5: 703.9 nm, B6: 740.2 nm, B7: 782.5 nm, B8: 835.1 nm, B8A: 864.8 nm, B9: 945 nm, B11: 1613.7 nm, B12: 2202.4 nm
Radiometric Resolution	12 bits
Temporal Resolution	5 days
Used Dates	2015-07-30 2020-08-02

In general, one of the functions of these satellite sensors is mapping processes regarding LULC. CORINE (Coordination of information on the environment) Level 2 classes were considered in the classification process. CORINE is an inventory of European land cover split into 44 different land cover classes. Corine Level 2 classes are selected for the training data of the three dated images. According to spatial characteristics of the study area, four different classes in the CORINE Level 2 are selected by their order.

Classification algorithms were applied on the GEE platform using Sentinel-2 satellite images. GEE is a web-based system that enables access to a comprehensive catalog of satellite images, analysis, and visualization on a global scale. While the entire surface can be accessed, it provides freely accessible huge datasets including satellite image archive data for scientists and researchers. Provided data can be used in different remote sensing and geospatial analysis.

4. Methodology

Based on the GEE platform, two ML algorithms, SVM and RF, and different training data sets were used and compared. A flowchart of the study is given in Fig. 2 and conducted research steps were explained in the following sections.

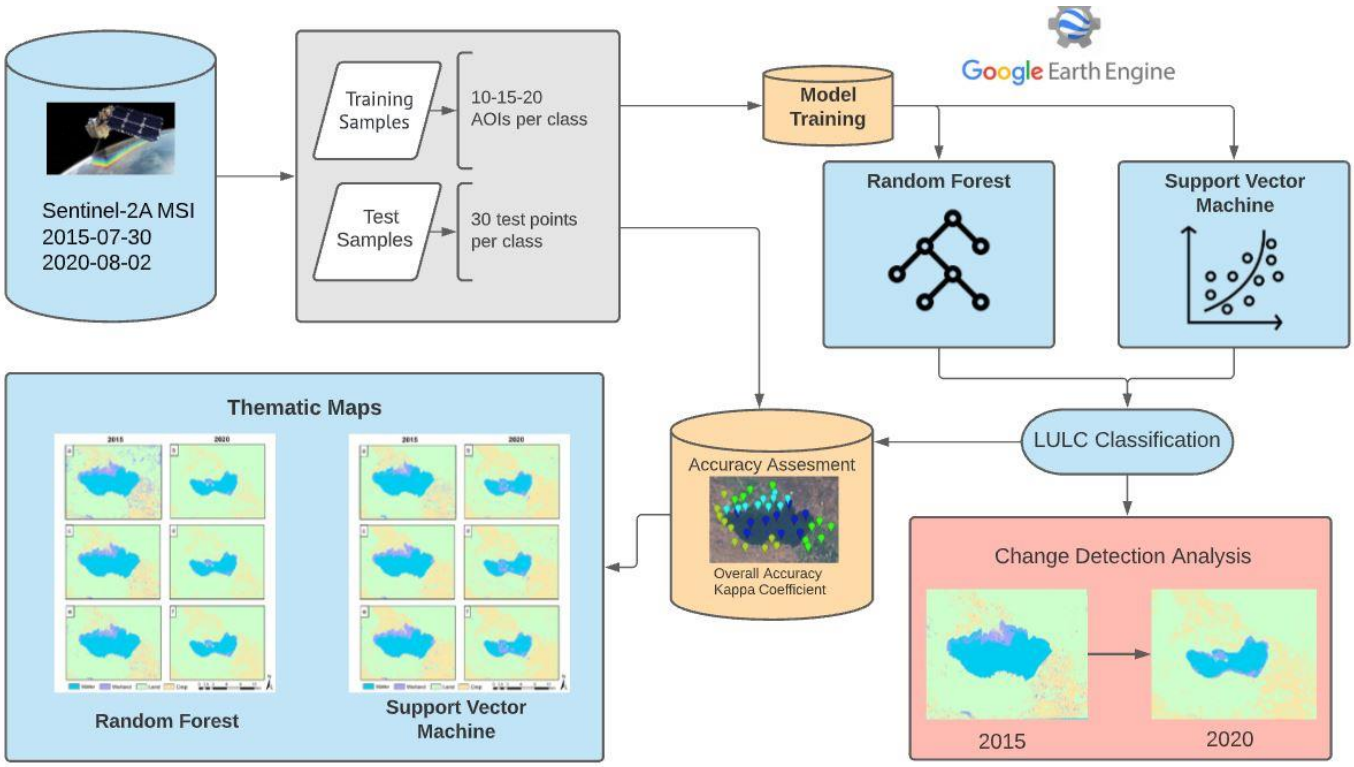

Figure 2. Flowchart of the study

In the first stage, RF classifier was used for mapping Marmara Lake. The RF classifier is a machine learning algorithm that combines many tree classifiers. To classify an input vector, each tree classifier generates a unit vote for the most common class in the tree [15]. RF, which is one of the most applied machine learning algorithms in classification studies [4,16] with a different type of data [17], increases the accuracy of the classification by creating more than one decision tree (Fig. 3). Several studies have shown that the RF classifier is capable of handling high data dimensionality and multi-linearity while still being fast and resistant to overfitting [18]. The

number of active variables in the random subset at each node and the number of trees in the forest are two parameters of RF. The number of active variables was set to the square root of the feature numbers whereas the number of trees was fixed to 25.

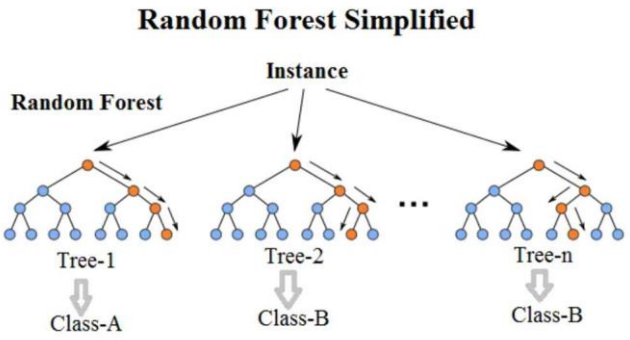

Figure 3. Random Forest Algorithm Architecture [18]

Since SVM is a supervised non-parametric statistical learning method, no assumptions about the underlying data distribution are made [19]. In the SVM algorithm, a hyperplane is created and the data is divided into two classes (Fig. 4). SVM is popular in remote sensing classification studies [20-21] and it is stated that SVM can be dealt with the classification of complex LULC [22]. For the SVM classifier, the radial basis kernel function was selected and the hyperparameters for gamma and cost were selected respectively 0.5 and 10.

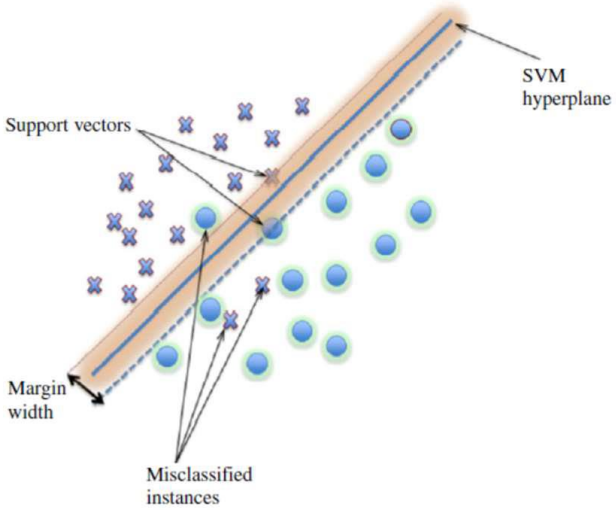

Figure 4. Support Vector Machine Algorithm Example [23]

Training datasets were collected on the GEE as polygons which are shown in Fig 5. The training area of interests (AOIs) for each class were selected as polygons which contain almost homogenous pixels in the area. To estimate the effect of training areas that are used on the classification they were separated into 3 different sized groups. The first group has the lowest area with 10 polygons and the second one has 15 polygons and the last one has the biggest area with 20 polygons. The effects of the increasing number of training data were evaluated. The results of the classification are different than each other.

These results can show us how much training area should be taken into classification to determine the best accuracy. Point data were used for the accuracy assessment (Fig. 5). The number of the training area and accuracy assessment points are given in Table 2 and the number of pixels in three data sets is given in Table 3. Classified satellite images are given in Fig. 6.

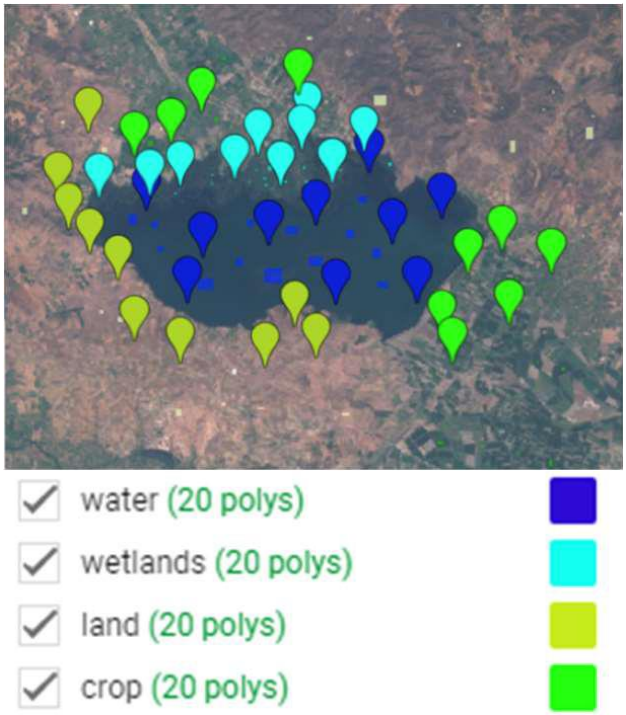

Figure 5. An example of AOIs for each Land Cover class collected with geometry tool in Google Earth Engine map.

Table 2. Detailed information of training and test data

Data	Date	AOI per class	Algorithm	Acc. points per class	Ass. class
Sentinel-2 MSI	2015	10 15 20	RF & SVM	30	
	2020	10 15 20		30	

Table 3. Number of pixels for each class 2015 and 2020

AOI/Class	Water	Wetlands	Land	Crop	Date
10	11159	1063	3714	1891	
15	16961	1141	3958	2177	2015
20	20679	1451	5769	2647	
10	4786	336	2269	740	
15	5483	396	2615	859	2020
20	6379	513	4325	1158	

Accuracy assessment shows the quality of the classification map. Therefore, overall accuracy (OA) and kappa coefficient (K) were calculated for each classification result using Eq.1-2.

$$OA = \frac{S_k}{S} \times 100 \quad (1)$$

where k represents the different AOI on the diagonal of the error matrix, Sk is the number of correctly classified samples, and S is the sample number.

$$K = \frac{p(\text{correct classification}) - p(\text{chance classification})}{1 - p(\text{chance classification})} \quad (2)$$

where p is proportion [24]. K values vary between 0 to 1 and values close to 1 show good classification.

5. Results and Discussion

The accuracy assessment results of LULC classification have been achieved using error matrices. In the accuracy assessment, four classified categories were considered: water, wetlands, land, and crop. Water defines the open surface area of the water body and the wetland class consists of the inland marshes. The land class covers the sparsely vegetated areas and the crop defines the heterogeneous agricultural area. LULC classes were created based on the CORINE. The overall accuracy (OA), and the kappa coefficient (K) was generated for each classification experiment and are given in Table 4.

In the accuracy assessment analysis, there are multiple methods to determine the number of test points as balanced or unbalanced. In studies with balanced and unbalanced training data sets, it has been observed that as the number of training pixels in each class increases, the model learns better and the classification accuracy increases in both data sets [25-26]. According to Thanh et al. [27] although the performance of balanced and unbalanced training datasets differs for both algorithms, the accuracy between the two models is approximately similar and high if the training sample is sufficient for the model. In our study, a direct link between pixel counts and accuracy was observed in parallel with these articles.

Table 4. Overall accuracy and kappa results of the classifications

AOI\	SVM		RF		Date
Model	OA (%)	Kappa	OA (%)	Kappa	
10	81.0	0.74	82.0	0.76	
15	88.0	0.83	86.0	0.81	2015
20	91.0	0.88	93.0	0.91	
10	80.8	0.74	81.7	0.76	
15	86.7	0.82	86.7	0.82	2020
20	91.7	0.89	92.5	0.90	

Besides the importance of the number of test sets, however, the distribution of the test sets is also important to more accurate analysis. For this purpose, test points should be homogeneously distributed. In this study, balanced test points were used (Table 2) and test points were homogeneously distributed. 30 test points were used per class and according to Story and Congalton [28] at least 30 samples are required to sufficiently populate the error matrix. According to Table 4, an increasing number of the training data improved the classification accuracy in both algorithms. However, the RF algorithm yielded slightly better overall accuracies with higher kappa coefficients (0.91 and 0.90, respectively) on both dates compared to the SVM algorithm. The best classification accuracy in both algorithms was obtained with 20 AOI on both dates. The error matrices of classification with 20 AOI are shown in Table 5. Water and wetland classes were mixed in the classification results.

The main reason is the mixed pixels that contain water and marshes on the coastal side of the lake. Additionally, the other mixed classes are land and crop. Because of having heterogeneous agricultural pattern, crop class mixed with the land class which contains sparsely vegetation. Classification results are shown in Fig. 7 and Fig. 8. Classification maps were produced considering CORINE color standards.

The areal change of the classes was calculated based on the RF classification results that were determined as the most accurate in the study. The areal coverage of the classes on both dates is given in Table 6. Marmara Lake, which is the Nationally Important Wetland, has lost more than half of the water and wetland (50.2% and 62.2%, respectively) since 2015. Most of these lost sites were converted to croplands.

Although the lake has been feeding 3 different sources, the water area has been decreasing in the last years. The first reason for that is the increase in irrigation demand [29]. According to the Turkish Statistical Institute [30], the agricultural area in the Gölmarmara, Ahmetli, and Salihli districts where the lake is located has increased approximately 3.3% between 2015 and 2020. This has caused an increase in water consumption for irrigation purposes.

The second reason is the withdrawal of groundwater sources. Although there is not enough information about groundwater resources, it is known that there are many illegal wells other than licensed wells [31].

The third and significant reason is the effects of the meteorological parameters. According to the Turkish State Meteorological Service [32], the annual precipitation average has decreased from 637.8 mm to 507.6 mm between 2015 and 2020 years on the national scale. In addition to precipitation, the annual temperature average has risen approximately 1 °C in the last 5 years.

6. Conclusion

In a conclusion, training data size and classification algorithms are important to achieve higher classification accuracy. In the study, it can be stated that the RF algorithm showed slightly better classification performance compared with the SVM algorithm. Increasing the number of training datasets improved the classification accuracy. Classification processes were applied to Sentinel-2 satellite images which have high spatial resolution data among the free available satellite images. All classification steps were implemented to the GEE and it makes processing satellite images easy with the increasing number of data and provides to rapidly access the results.

According to change detection analysis from the most accurate classification results, the area of the water body and wetland has dramatically decreased. Although the

lake is fed by different sources, the most likely causes of these significant decreases in the water body and wetland are climate change and unconscious irrigation. Additionally, it is observed that most of the dried lands were converted to the agricultural area in the study site.

This study has generated significant information about the LULC dynamics of Marmara Lake and its transformation during the last five years (2015–2020). It can be used as reference data for the decision-makers to protect water resources and nationally important wetlands. Additionally, this study was conducted for the summer season.

In future studies, water bodies and wetlands can be evaluated with seasonal and integrated with a different type of datasets which can improve the classification accuracy.

Table 5. Error matrices of RF and SVM classifications with 20 AOI

Error Matrix of RF 2015						
LULC Classes	Water	Wetlands	Land	Crop	Total	User Acc. (%)
Water	28	1	0	0	29	97.0
Wetlands	2	29	0	0	31	94.0
Land	0	0	27	2	29	93.0
Crop	0	0	3	28	31	90.0
Total	30	30	30	30	120	
Producer Acc. (%)	93.3	97.0	90.0	93.3		**93.0**
Error Matrix of RF 2020						
LULC Classes	Water	Wetlands	Land	Crop	Total	User Acc. (%)
Water	28	4	0	0	32	87.5
Wetlands	2	26	0	0	28	92.9
Land	0	0	28	1	29	96.6
Crop	0	0	2	29	31	93.5
Total	30	30	30	30	120	
Producer Acc. (%)	93.3	87.0	93.3	96.7		**92.5**
Error Matrix of SVM 2015						
LULC Classes	Water	Wetlands	Land	Crop	Total	User Acc. (%)
Water	27	3	0	0	30	90.0
Wetlands	3	27	0	0	30	90.0
Land	0	0	27	2	29	93.0
Crop	0	0	2	28	31	90.0
Total	30	30	30	30	120	
Producer Acc. (%)	90.0	90.0	90.0	93.3		**91.0**
Error Matrix of SVM 2020						
LULC Classes	Water	Wetlands	Land	Crop	Total	User Acc. (%)
Water	27	2	0	0	29	93.1
Wetlands	3	28	0	0	31	90.3
Land	0	0	27	2	29	93.1
Crop	0	0	3	28	31	90.3
Total	30	30	30	30	120	
Producer Acc. (%)	90.0	93.0	90.0	93.3		**91.7**

Figure 6. Sentinel-2A MSI RGB Imageries of 2015-07-30 and 2020-08-02

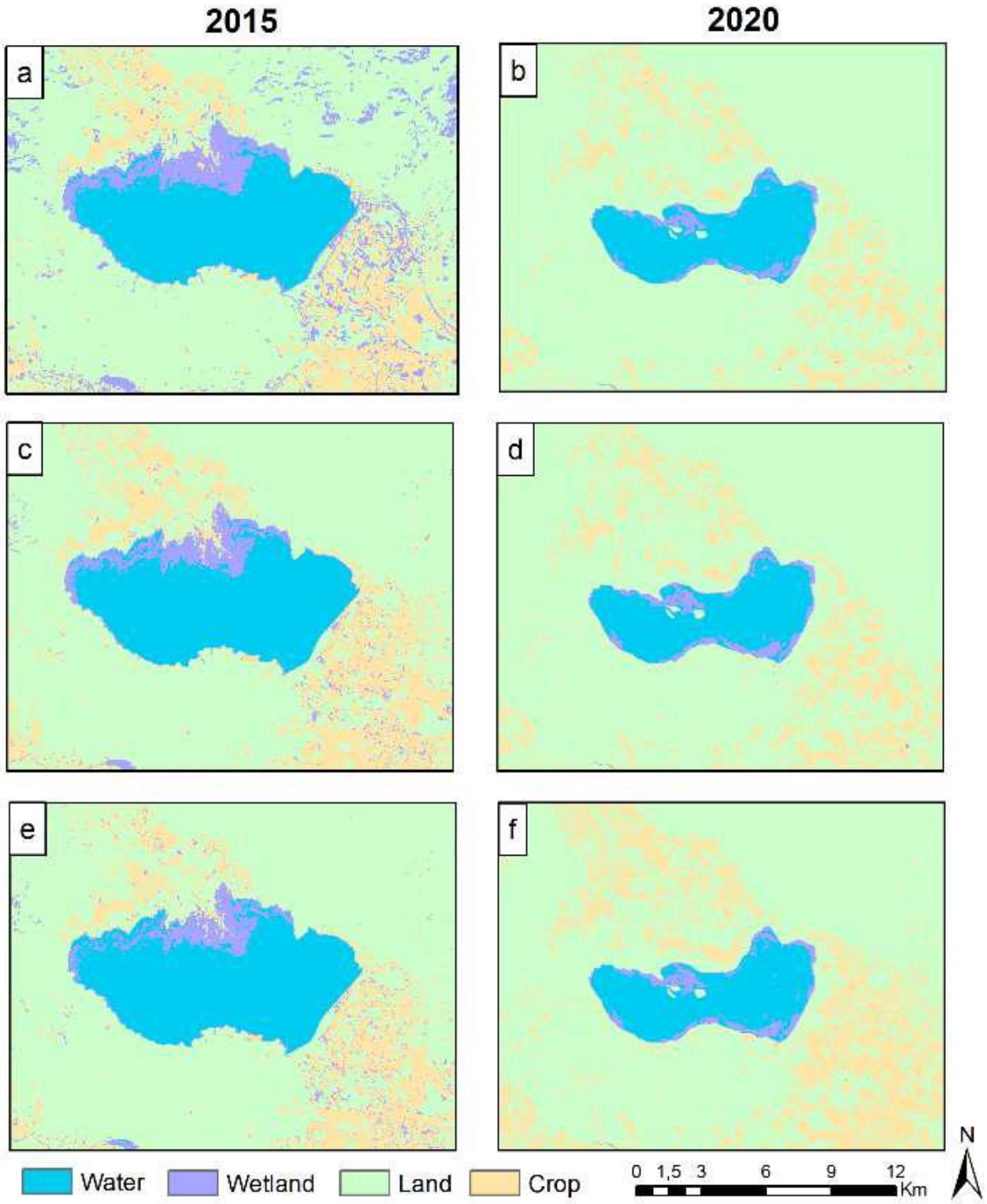

Figure 7. LULC classification results using RF classifier with 10 AOI (a-b), 15 AOI (c-d) and 20 AOI (e-f) for two dates

Table 6. Areal change of the lake between 2015 and 2020

Class/Area	2015 (ha)	2020 (ha)	Change (ha)	Change (%)
Water	5399.1	2689.3	-2709.8	- 50.2
Wetlands	1604.4	605.9	-998,5	- 62.2
Land	24199.7	25387.6	+1187,9	+ 4.9
Crop	5049.2	7569.5	+2520,3	+ 49.9

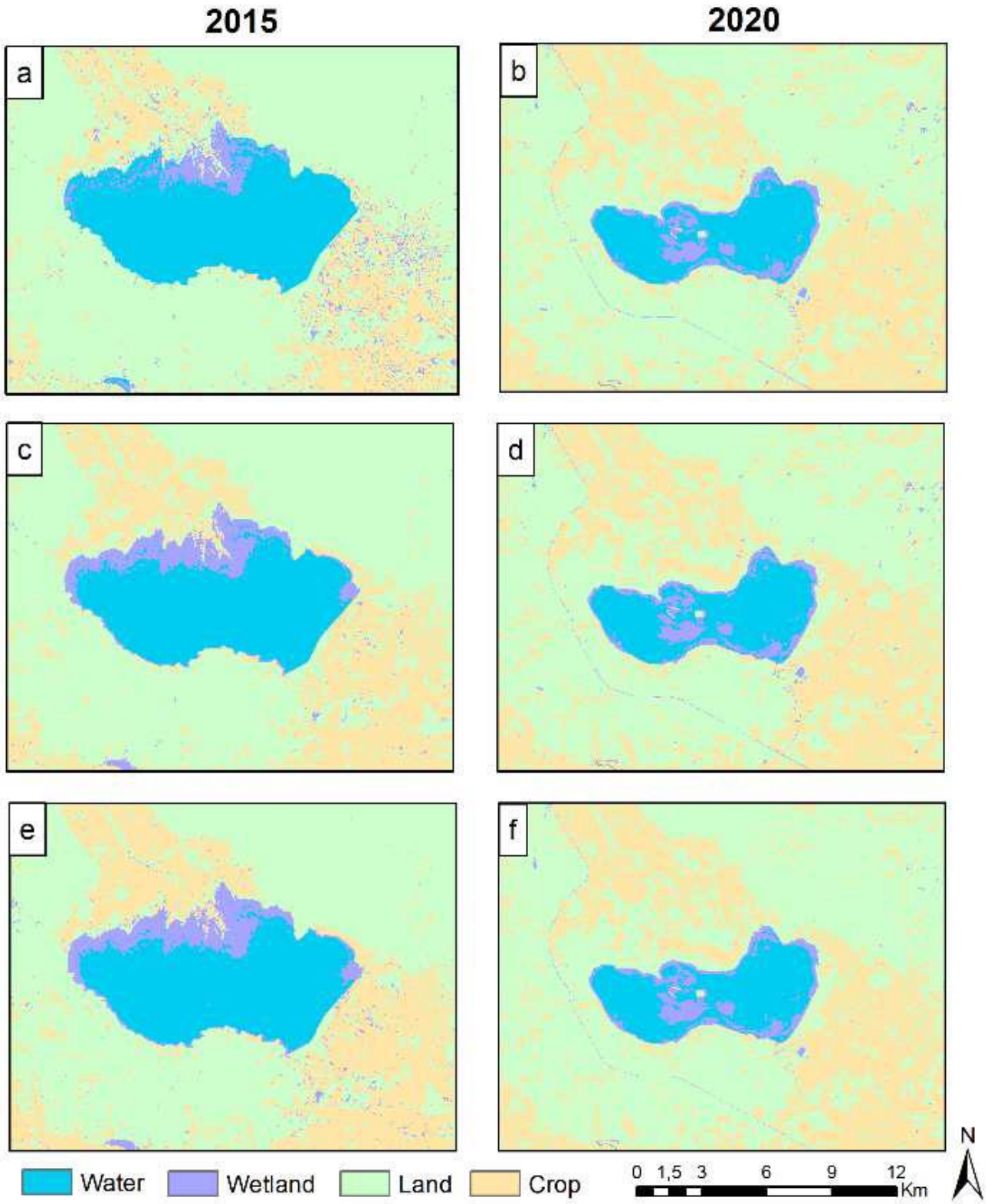

Figure 8. LULC classification results using SVM classifier with 10 AOI (a-b), 15 AOI (c-d) and 20 AOI (e-f) for two dates

Acknowledgement

The authors would like to acknowledge Sentinel-2 imagery from European Space Agency (ESA).

Author contributions

Cengiz Avcı: Methodology, Process, Writing **Muhammed Budak:** Methodology, Process, Writing **Nur Yağmur:** Process, Writing **Filiz Bektas Balcik:** Editing, Consultancy

Conflicts of interest

The authors declare no conflicts of interest.

References

1. DeFries, R. S., Foley, J. A., & Asner, G. P. (2004). Land-use choices: Balancing human needs and ecosystem function. Frontiers in Ecology and the Environment, 2(5), 249-257.
2. Kavzoglu, T., Tonbul, H., Erdemir, M. Y., & Colkesen, I. (2018). Dimensionality reduction and classification of hyperspectral images using object-based image analysis. Journal of the Indian Society of Remote Sensing, 46(8), 1297-1306.
3. Ekumah, B., Armah, F. A., Afrifa, E. K., Aheto, D. W., Odoi, J. O., & Afitiri, A. R. (2020). Assessing land use and land cover change in coastal urban wetlands of international importance in Ghana using Intensity Analysis. Wetlands Ecology and Management, 28(2), 271-284.
4. Basu, T., Das, A., Pham, Q. B., Al-Ansari, N., Linh, N. T. T., & Lagerwall, G. (2021). Development of an integrated peri-urban wetland degradation assessment approach for the Chatra Wetland in eastern India. Scientific reports, 11(1), 1-22.
5. Jamal, S., & Ahmad, W. S. (2020). Assessing land use land cover dynamics of wetland ecosystems using Landsat satellite data. SN Applied Sciences, 2(11), 1-24.
6. Hochreuther, P., Neckel, N., Reimann, N., Humbert, A., & Braun, M. (2021). Fully Automated Detection of Supraglacial Lake Area for Northeast Greenland Using Sentinel-2 Time-Series. Remote Sens. 2021, 13, 205.
7. Shih, H. C., Stow, D. A., & Tsai, Y. H. (2019). Guidance on and comparison of machine learning classifiers for Landsat-based land cover and land use mapping. International Journal of Remote Sensing, 40(4), 1248-1274.
8. Bangira, T., Alfieri, S. M., Menenti, M., & Van Niekerk, A. (2019). Comparing thresholding with machine learning classifiers for mapping complex water. Remote Sensing, 11(11), 1351.
9. Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover classification. Pattern recognition letters, 27(4), 294-300.
10. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202, 18-27.
11. Wang, Y., Ma, J., Xiao, X., Wang, X., Dai, S., & Zhao, B. (2019). Long-term dynamic of poyang lake surface water: a mapping work based on the Google earth engine cloud platform. Remote Sensing, 11(3), 313.
12. Wang, C., Jia, M., Chen, N., & Wang, W. (2018). Long-term surface water dynamics analysis based on Landsat imagery and the Google Earth Engine platform: A case study in the middle Yangtze River Basin. Remote Sensing, 10(10), 1635.
13. Amani, M., Mahdavi, S., Afshar, M., Brisco, B., Huang, W., Mohammad Javad Mirzadeh, S., ... & Hopkinson, C. (2019). Canadian wetland inventory using google earth engine: The first map and preliminary results. Remote Sensing, 11(7), 842.
14. MoAF (Ministry of Agriculture and Forestry) (2018). Wetland Managemant Plan of Marmara Lake. Ankara.
15. Breiman, L. (1999). Random forests. UC Berkeley TR567.
16. Berhane, T. M, Lane, C. R., Wu, Q, Autrey, B. C., Anenkhonov, O. A., Chepinoga, V. V., & Liu, H. (2018). Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. Remote sensing, 10(4), 580.
17. Dubeau, P., King, D. J., Unbushe, D. G., & Rebelo, L. M. (2017). Mapping the Dabus wetlands, Ethiopia, using random forest classification of Landsat, PALSAR and topographic data. Remote Sensing, 9(10), 1056.
18. Jagannath, V. (2020). "Random Forest Template for TIBCO Spotfire® - Wiki Page TIBCO Community." https://community.tibco.com/wiki/random-forest-template-tibco-spotfire
19. Vapnik, V. (1998). The support vector method of function estimation. In Nonlinear modeling (pp. 55-85). Springer, Boston, MA.
20. Qian, Y., Zhou, W., Yan, J., Li, W., & Han, L. (2015). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. Remote Sensing, 7(1), 153-168.
21. Han, X., Pan, J., & Devlin, A. T. (2018). Remote sensing study of wetlands in the Pearl River Delta during 1995–2015 with the support vector machine method. Frontiers of Earth Science, 12(3), 521-531.
22. Pretorius, L., Brown, L. R., Bredenkamp, G. J. & van Huyssteen, C. W. (2016). The ecology and classification of wetland vegetation in the Maputaland Coastal Plain, South Africa. Phytocoenologia, 46(2), 125-139.
23. Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.
24. Canty, M. J. (2014). Image analysis, classification and change detection in remote sensing: with algorithms for ENVI/IDL and Python. Crc Press.
25. Colditz, R. R. (2015). An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. Remote Sensing, 7(8), 9655-9681.
26. Mellor, A., Boukir, S., Haywood, A., & Jones, S. (2015). Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. ISPRS Journal of Photogrammetry and Remote Sensing, 105, 155-168.

27. Thanh, Noi, P., & Kappas, M. (2018). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors, 18(1), 18.
28. Story, M., & Congalton, R. G. (1986). Accuracy assessment: a user's perspective. Photogrammetric Engineering and remote sensing, 52(3), 397-399.
29. Tubitak MAM (2013). Preparation Project of Basin Protection Action Plans, Gediz Basin. Project Report, Kocaeli.
30. TUIK, 2020. https://www.tuik.gov.tr/
31. Korbalta, H. (2019) Marmara Gölü Neden Kuruyor? Kent Akademisi, 12(3), 441-459.
32. MGM (2020). Analysis of meteorological parameters for Turkey. Accessed from: https://www.mgm.gov.tr/veridegerlendirme/il-ve ilceleristatistik.aspx?k=parametrelerinTurkiyeAnaliz i.