

Optimizing Feature Selection for Solar Park Classification: Approaches with OBIA and Machine Learning

Claudio Ladisa(✉), Alessandra Capolupo, and Eufemia Tarantino

Department of Civil, Environmental, Land, Construction and Chemistry (DICATECh), Politecnico di Bari, via Orabona n. 4, 70125 Bari, Italy
{claudio.ladisa,alessandra.capolupo,eufemia.tarantino}@poliba.it

Abstract. In the context of climate change mitigation, the crucial role of solar energy as an inexhaustible and renewable source is underlined by the widespread integration of photovoltaic (PV) panels. To keep track of the rapid expansion of extensive PV systems (such as solar parks), it is important to keep monitoring their quantity, distribution, and impact. Despite the many complications involved in mapping PV systems, stemming from the diversity of materials and field layouts, many researchers have focused on ways to improve this process through Remote Sensing techniques aided by Machine Learning and Deep Learning. In this study, multiple object-based image (OBIA) classification models were created for mapping large solar parks, using open-source data (Sentinel-2) and computational tools such as Google Earth Engine (GEE) and RStudio. The case study was conducted in Pavagada Solar Park, India, which is the third largest solar park in the world spanning over 153 km^2. For training, validation, and testing purposes two zones within the park were identified. Data were collected for four seasonal periods; a classification model was developed for each season while validating them. The most accurate model was then implemented in the test area. The selection of features for the Random Forest (RF) classifier was done using the Recursive Feature Elimination (RFE) method, with the main objective of eliminating non-relevant features while preserving only the significant ones. This analysis showed a substantial improvement in accuracy, thereby validating the effectiveness of the RFE algorithm in feature selection. This is particularly true in the autumn when a maximum Overall Accuracy (OA) of 88.72% was achieved.

Keywords: Object-Based Image Analysis · Sentinel 2 · medium-resolution satellite images · photovoltaic panels

1 Introduction

Rising public consciousness about environmental problems as well as technological progress resulted from the increase of installed renewable energy sources

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14819, pp. 286–301, 2024.
https://doi.org/10.1007/978-3-031-65282-0_19

in energy production [1]. To enumerate these, solar energy is among the most potential ones abounding in nature and has little negative impact on nature [2]. Countries and international organizations, like the European Union, China, India, the United States, and Japan, have provided huge funding for the design and implementation of PV systems [3]. Indisputably, India has made good use of its energy opportunities, sprawling across wide regions, and exploiting the many areas with more than 300 sunny days a year. It is not only that green energy is brought to rural areas, but by the same token, it keeps the environment clean and safe from the pollution of fossil fuel production [4]. The development of solar energy has been increasing strongly since 2009, with a compound annual growth rate of 41%, which will lead to an almost tenfold expansion by 2040 [5]. Nevertheless, concerns about the loss of territorial resource surface, reducing the size of arable farmlands, and alterations to the surrounding environment are also raised with the widespread installation of solar PV panels [6]. Consequently, more and more attention is being paid nowadays to the possibility of a process being established that will, as a result, fundamentally affect the animal and plant world and the overall functioning of the natural system [7]. Therefore, there is an urgent need for detailed monitoring and mapping of large solar parks. However, access to detailed national data on the location of PV solar panels is often limited and solar databases based on voluntary contributions may lack completeness [8]. Therefore, refining the update of spatial distribution data and generating dynamic information on PV systems plays a fundamental role in research, encompassing the analysis of impacts resulting from changes in land use in local climates [9]. Regarding the example of industrial solar parks, the usual mapping strategies, i.e. direct ground surveys, are too time-consuming. Therefore, the development of a comprehensive, fast, and non-invasive data collection method is of great importance [10]. The strong development of space programs such as Copernicus and Landsat have produced a significant amount of data on the Earth's surface. This phenomenon has enabled remote sensing to be a key subject and to be used in a wide range of environmental applications [11–13]. Several studies have shown excellent results for mapping PV parks utilizing open-source optical data such as Sentinel-2 (S2) or Landsat 8, rather than high-resolution commercial data, which need significant processing resources and limit its applicability across wide geographical expansions [10,14,15]. In some recent studies, remote sensing techniques have been combined with supervised Machine Learning algorithms or convolutional neural networks (CNN) for PV panel extraction [16,17]. However, from the latest research results, it was noted that CNN requires large training sets in contrast to supervised classification algorithms such as RF, which require less training data but still achieve excellent accuracy values [18]. This research aims to develop and validate an RF-supervised classification model to identify PV systems within the Pavaganda solar panel complex developed in Pavagada, Tumkur district, which is located approximately 180 km north of Bangalore in India. To address one of the recognized "salt and pepper" effects in image classification, the research opts for OBIA rather than pixel-based approaches and seeks to utilize the geometrical information in the range of segments [19]. Data

collection was done individually across four different seasons, which are February, April, August, and November 2022. The multi-temporal approach for this study was used to fully investigate the impact of seasonal changes on the efficiency of different RF classification models. During this study, only open-source platforms and tools, such as GEE and RStudio, were used together with S2 multispectral optical data to develop an easily reproducible methodology. This research differs from others in the choice of variables to be input into classification models. Various spectral indices, already used in previous studies, were collected and through the RFE algorithm within the RStudio environment, only the best were chosen. The algorithm is performed by cumulatively removing weak features, which leads the model to eliminate dependencies and collinearities, thus improving the accuracy of the classification process [20].

2 Study Area

The study area is located in Pavagada taluk, Tumkur district, Karnataka, India. Within this region lies the Pavagada Solar Park, a solar park spanning a total area of 153 km^2, making it the third-largest solar power plant in the world. Known as Shakti Sthala, the park is situated approximately 180 km from Bengaluru. The land on which the park is built was acquired from 2,300 farmers through a 25-year lease agreement [4]. In this work, two distinct areas within the Pavaganda Park extension were utilized. As depicted in Fig. 1, one area was designated as the training and validation zone, while another area was exclusively utilized as the test zone (ground truth).

3 Data Set

3.1 S2 Imagery

S2 satellite images were preferred over Landsat 9 images due to their better geometric resolution. The experience in the past with the use of S2 evidence confirms the good results for both PV panel mapping and other monitoring of land uses, among others [21–24]. The S2 multi-spectral sensor has 13 spectral bands from wavelengths of 0.44–2.2 μ m, including visible, NIR, and shortwave IR regions of the spectrum [25]. The spatial resolution varied from 10 to 60 m across different bands [26]. The S2 Level 2A product, which was already atmospherically corrected, was selected from the GEE platform. Image selection was carried out for February, April, August, and November, to conduct a multi-temporal analysis to identify the most suitable season for developing high-performance classification models. For each month, preference was given to images with minimal cloud coverage, particularly for those with a coverage percentage below 1%. Among the identified images, high-resolution bands of 10 m were used, including blue (band 2, 490 nm), green (band 3, 560 nm), red (band 4, 665 nm), and NIR1 (band 8, 842 nm). Additionally, 20-meter resolution bands were considered, such as red edge 1 (RE1) (band 5, 704 nm), red edge 2 (RE2) (band 6, 740 nm), red edge

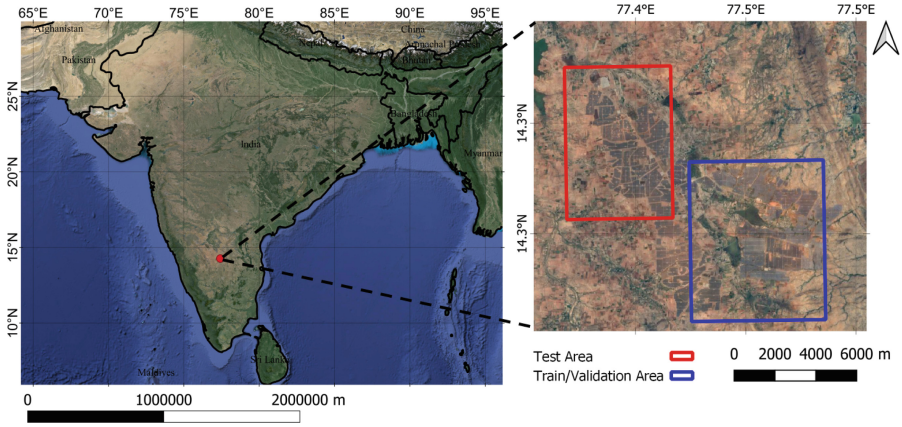

Fig. 1. Train/Validation and Test Area in the Pavaganda Solar Park.

3 (RE3) (band 7, 783 nm), narrow NIR2 (band 8a, 865 nm), SWIR 1 (band 11, 1614 nm), and SWIR 2 (band 12, 2202 nm). Red edge bands, narrow NIR, and SWIR were included, as they are crucial for distinguishing vegetated areas from non-vegetated areas and identifying bare soil from other land cover types [27]. The images were cropped within the designated study areas for training, validation, and testing.

3.2 Methodology

The study's classification model for the solar park was created mostly using OBIA and RF approaches. The initial stage of the work, shown in Fig. 2, was the data pre-processing after S2, as described previously in this paragraph. The next step occurred in segmentation, which is the preliminary phase of the OBIA approach [9]. This was a process that used a Simple Non-Iterative Clustering (SNIC) segmentation method accessible within GEE. The segmentation phase was applied to prepare the data both in the training/validation and test areas. Various features were computed for each object, focusing on the spectral indices previously employed in other studies for PV mapping [15,28,29]. These indices were calculated for each of the four considered months. Then, objects that had been separated were subsequently labeled using Quantum GIS (QGIS) software. Following segmentation, feature extraction, and object labeling, the workflow was continued in the RStudio environment. Before proceeding to the classification phase, feature selection was performed using RFE. This algorithm was applied each month separately. Because the objects classified as PV are much fewer than those assigned as NO-PV, the dataset was rebalanced after selecting the features for each month. The imbalance in the number of class objects, as in this scenario, may produce misleading results by effectively supplying the majority class while failing to identify the minority class [30]. Oversampling was used over subsampling since earlier studies showed that subsampling resulted in

information loss during the training phase [10]. After balancing the objects from the study area designated for training and validation, they were randomly split as follows: 70% of the dataset was set as the training dataset, and the rest of the 30% was used for validating the RF model. For this purpose, a unique RF model was generated for each month. Then plate metrics such OA, Cohen's Kappa (K), Producer's Accuracy (PA), User's Accuracy (UA), and F1 were calculated using different algorithms. Eventually, the model with the best accuracy was applied to the Test area, and the metrics were computed.

3.3 SNIC Algorithm Segmentation

The SNIC algorithm is made available by the GEE platform. The decision to use this segmentation technique was driven by the successful outcomes of previous studies employing S2 multispectral images, in the fields of PV plant mapping and other applications [10,30]. Assigning some pixels to centroids on a regular grid within the image serves as the algorithm's initial step. Following that, the distances between each pixel and its centroid in terms of color and spatial coordinates within a 5-dimensional space were computed [10,30,31]. It is a distance measurement that combines normalized spectral and spatial information leading to the generation of visually homogeneous, compact superpixels [32]. For expanding the superpixel cluster, SNIC chooses four or eight immediately connected candidate pixels from the priority queue without considering them. These candidate pixels are selected based on their minimum distance from the centroid [33]. Hence both memory use and processing time are better than in similar clustering methods like Simple Linear Iterative Clustering. The reason behind this outcome can be attributed to the implementation of connectivity (four or

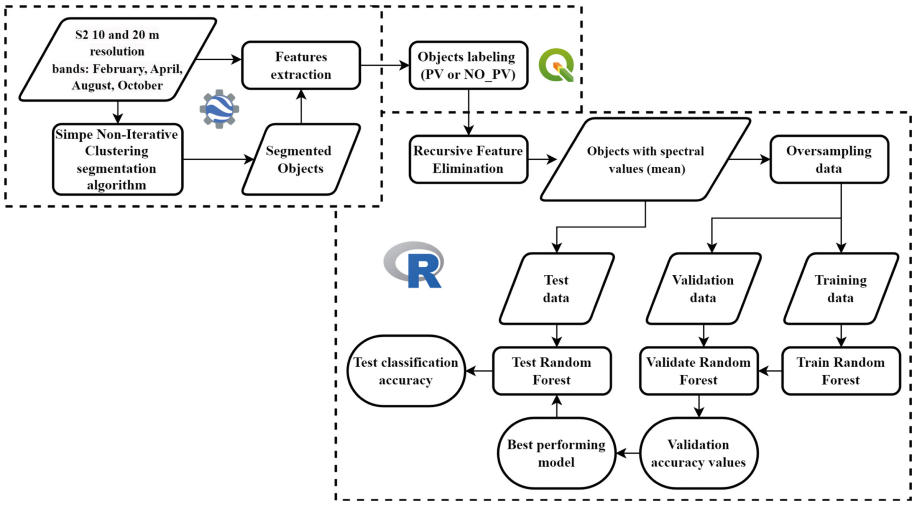

Fig. 2. Operative pipeline of the applied methodology.

eight connected pixels), which led to a reduction in the number of distance calculations during centroid evolution [34]. In the GEE platform, the segmentation parameters used were based on the configurations employed in previous studies [10]. These parameters consisted of (1) compactness, set to 1; (2) connectivity, set to 4; (3) neighborhood size, fixed at 64; (4) size, set to 10; (5) seeds, set to 15; (6) kernel, set to 3; and (7) scale, set to 10. For both the training/validation area and the test area, the multispectral image chosen for segmentation was the one captured in February, as, through visual interpretation, it was the one in which the PV plants were visually most separated from the adjacent land covers.

3.4 Features Extraction

Following the segmentation of elements, computations were done using GEE for intensities of specific S2 bands for each item, as well as many spectral indices. The indices described in Table 1 have shown useful in previous studies, playing an important role in the detection of solar systems.

3.5 Objects Labeling

After calculating all the indices listed in Table 1, the segmented objects from both the training, validation, and test areas were exported as shapefiles from GEE. These items were then separated into two classes in QGIS: PV and NO-PV. The classification technique was based on visual interpretation. Objects with more than half of the pixels associated with PV systems were classified as PV. Notably, this study focused just on PV panels produced from specialist PV systems, eliminating smaller panels attached to buildings, which were absent from both the test, validation, and test regions.

3.6 RFE

This study employed a feature selection method known as RFE. The widely adopted RFE algorithm was utilized to identify the most relevant features in predicting the target variable, whether in regression or classification-oriented models [35]. The RFE technique uses an iterative backward selection process to find the best feature combination by starting with all of the variables and progressively eliminating those that are incorrect for classification accuracy [36]. In this analysis, the implementation of RFE was conducted through the "rfFuncs" function within the R caret framework. With this specific function, the algorithm relies on assigning importance to variables and is internally evaluated using RF classifiers [37]. This process involves multiple rounds of classification, in which a new RF classification model is trained. Its accuracy is evaluated through cross-validation and the importance of parameters for each feature are analyzed. The feature set is updated for the next iteration [38]. Cross-validation was performed for ten cycles in this case. The feature-selection algorithm was repeated for the four seasonal training datasets. The best features obtained from the RFE algorithm were used as input variables successively for the various RF classification models.

Table 1. Spectral indices used in literature.

Index Name	Index Formula	Previous Studies
Normalized Difference Vegetation Index (NDVI)	NDVI = (NIR1 − Red)/(NIR+ Red)	[10,18,28,29]
Normalized Difference Water Index (NDWI)	NDWI = (Green − NIR1)/(Green + NIR1)	[10,18,28,29]
Normalized Difference Built-up Index (NDBI)	NDBI = (SWIR1 − NIR1)/(SWIR1 + NIR1)	[15,18,28,29]
Normalized Difference Tillage Index (NDTI)	NDTI = (SWIR1 − SWIR2)/(SWIR1 + SWIR2)	[15,18,28,29]
Enhanced Vegetation Index (EVI)	EVI = [2.5 × (NIR1 × Red)]/(NIR1 + SWIR1 × 4–7.5 × Blue + 1)	[15,18,28,29]
Modified NDWI (MNDWI)	MNDWI = (Green − SWIR1)/(Green + SWIR1)	[15,18,28,29]
Bare Soil Index (BI)	BI = (SWIR1 + Red − NIR1 − Blue)/(SWIR1 + Red + NIR1 + Blue)	[15,18,28,29]

3.7 Oversampling Data

Classification models can drastically change their performance if the learning dataset has unbalanced classes. This happens because models have difficulty effectively learning minority class characteristics. After all, they will receive fewer examples of that class than the majority class [39]. To solve this problem, an oversampling method was adopted aimed at balancing the number of objects in the two classes in question. Using the "upsample" functionality from the "caret" package in R, it was possible to equalize the number of objects for training the PV and NO-PV classes by generating synthetic objects of the PV class while preserving all the information of the objects of the NO-PV class [40].

3.8 RF Classification Models

There is a lot of interest in remote sensing and other machine-learning applications because of the RF technique for classification [41]. Its major feature is its ability to attain accuracy, even with poor training data [42]. RF is composed of a set of tree-based predictors, with each tree depending on the values of a randomly selected vector independently and with the same distribution as the other trees in the forest [43]. The RF model has two initialization parameters: (1) the number of classification trees or bootstrap iterations (ntree) and (2) the number of input variables picked at random at each node (mtry) [44]. Default settings were used for mtry and ntree, with mtry set to $\checkmark$p (where p denotes the number of predictive variables in a dataset) and ntree set to 500. Following RFE,

the polygons of the training and validation area were split. In this study case, the RF model was trained on 70% of the polygons, and the remaining 30% were utilized for validation. This technique was applied to all four seasonal datasets. The model with the highest accuracy was then applied to the test region data, which had already been categorized into two classes: PV and NO-PV.

3.9 Accuracy Analysis

The confusion matrix was used for model validation in the training and validation zone as well as evaluation in the test zone. The accuracy measures produced from the error matrix are OA, K, PA, UA, and F1. Table 2 shows the equations of the accuracy metrics used.

Table 2. Spectral indices used in literature.

Accuracy Metrics	Equations
Overall Accuracy	OA = (TP + TN)/(TP + TN+ FP+ FN)
Cohen's Kappa	K = (P(A) - P(E))/(1 - P(E))
Producer's Accuracy	PA = TP/(TP + FN)
User's Accuracy	UA = TP/(TP + FP)
F1 score	F1 = (2*TP)/(2*TP + FP +FN)

TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives, P(A): Concordance rate observed, P(E): Concordance rate expected

4 Results

4.1 Segmentation

The segmentation within GEE, using the SNIC algorithm, produced 2341 objects for the training/validation study area and 1599 objects for the test region. As illustrated in Fig. 3, the settings utilized for the SNIC algorithm resulted in objects with geometries that identify PV systems from other land coverings.

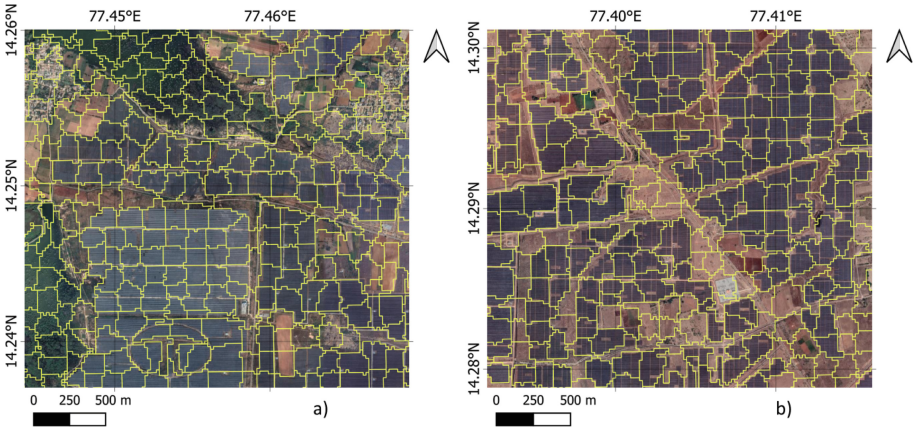

Fig. 3. Example of SNIC segmentation output obtained by processing the February image in the training and validation area (a) and test area (b).

4.2 Object Labeling

The QGIS software was used to carry out the labeling process. A total of 753 objects were identified and labeled as PV, while 1588 objects were categorized as NO-PV within the training and validation area. In the test area, 537 objects received the PV label, while 1062 objects were assigned the NO-PV classification.

4.3 RFE Results

Figure 4 shows the results obtained by the RFE algorithm for each season by month in the training and validation area. Table 3 shows the variables chosen by the RFE algorithm for each season, listed according to their predictive significance, from most to least important.

Table 3. Spectral indices used in literature.

Season	Features Selected	Accuracy values
Winter	Blue, BI, NDWI, NDTI, MNDWI, NDBI, EVI, NDVI	0.8729
Spring	NDTI, NDWI, EVI, NDBI, NIR2, NDVI, Blue, MNDWI, NIR1, Green, BI, RE3, RE2, Red	0.8738
Summer	NDTI, MNDWI, NDBI, Blue, SWIR1, BI, SWIR2, NIR1, NDWI, Green, NIR2, RE3, Red, RE2, NDVI	0.8781
Autumn	NDTI, Blue, SWIR1, MNDWI, SWIR2, BI, NDWI, NDBI	0.8742

4.4 RF Model Validation

Table 4 shows accuracy metrics results within the training and validation zone for every RF classification model generated for each season. The table represents

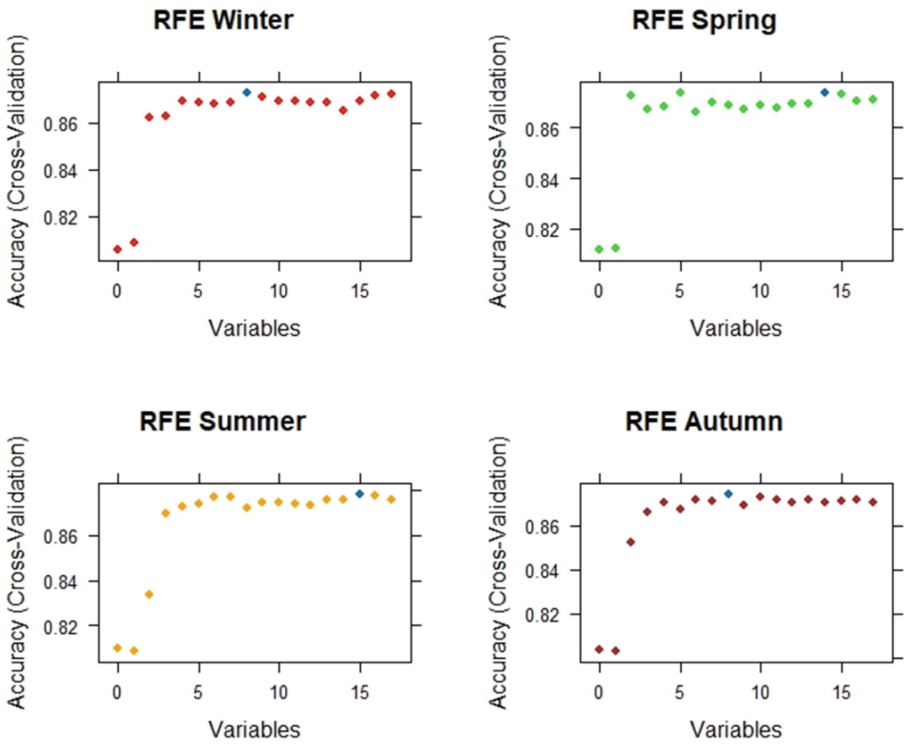

Fig. 4. The count of variables selected for each month using the RFE method, along with their corresponding accuracy values.

Table 4. Spectral indices used in literature.

RF model	OA (%)	K	PA (PV)	UA (PV)	F1 (PV)
Winter RFE	87.87	0.76	90.85	85.74	88.22
Winter	86.17	0.72	84.89	87.12	85.99
Spring RFE	86.89	0.74	85.32	88.52	86.89
Spring	86.49	0.73	86.38	86.57	86.47
Summer RFE	87.98	0.76	89.79	86.65	88.19
Summer	86.06	0.72	90	83.43	86.59
Autumn RFE	**88.72**	**0.78**	**89.57**	**88.08**	**88.82**
Autumn	85.96	0.72	83.62	87.72	85.62

accuracy values for both models using variables derived from RFE and those incorporating all variables, aiming to illustrate the disparities in the outcomes. The RF classification model, utilizing variables selected by the RFE algorithm during the autumn season, demonstrated superior accuracy metric values within

the training/validation domain. Consequently, this model was chosen for classification in the Test area.

4.5 Test Area Classification

Table 5 presents the accuracy metric results calculated using the Autumn RF model, with features selected by the RFE algorithm. Figure 5 visually displays the classification outcomes.

Table 5. Accuracy metrics of the "Autumn RFE" RF model in the Test area.

OA (%)	K	PA (PV)	UA (PV)	F1 (PV)
87.80	0.74	89.94	77.40	83.20

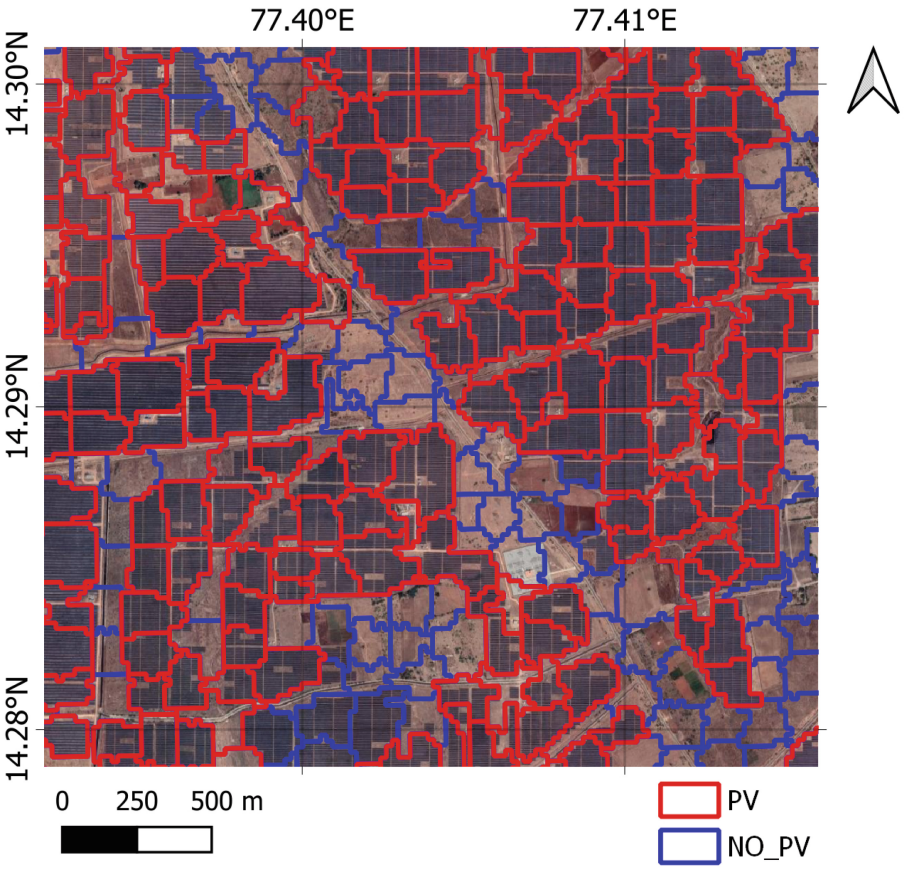

Fig. 5. Classification results obtained from the application of the RF model in the Test area.

5 Discussion and Conclusion

The goal of this study was to develop an RF classification model for mapping PV installations in Pavagada Solar Park in India using open-source data such as S2 and open-source software, including GEE and RStudio. The following conclusions were drawn from the results obtained. Starting with the OBIA approach, specifically from the segmentation phase, the SNIC algorithm within GEE produced geometrically consistent objects, aligning well with the actual geometries of PV installations. This facilitated the manual assignment of the PV and NO-PV classes in the subsequent object labeling phase. Consequently, the parameters of the SNIC algorithm proved effective, confirming the results of previous studies [10]. Some observations arise regarding the RFE algorithm step. First, the predictive ability of the indices chosen wherein from the literature is notable. Table 3 shows that the NDTI has great predictive power for PV objects, ranking first in all seasons except winter, where it remains fourth. This result confirms the exceptional predictive power of the NDTI index for mapping PV systems, as demonstrated by previous studies [15,18,29]. Figure 4 shows that the RFE algorithm selected a varying number of variables for each season while achieving similar accuracy values in predicting the PV class, as shown in Table 3. This is attributed to the seasonal variability of the spectral values corresponding to the PV installations. In any case, Table 4 shows that, as compared to utilizing all variables during all four times studied in the training and validation zone, using the variables chosen by the RFE algorithm produced greater accuracy results. The model developed in Autumn, with variables selected by the RFE algorithm, proved to be the best performant among the others, albeit with a slight difference. Finally, after validation, the same model was applied to the test zone, achieving accurate results, as shown in Table 5, confirming effective "knowledge transfer". This study introduces, for the first time, a methodology for selecting suitable variables for PV mapping. In previous studies, feature selection has relied on the analysis of the spectral signatures of PV installations, or the intrinsic meaning of the spectral indices used. However, the significant seasonal variation in spectral indices, primarily caused by the type of soil beneath the PV panels, did not provide certainty in using certain indices over others. Using the RFE algorithm allowed for the easy extraction of only the best features in terms of the predictive power for each period. This process optimized all classification models both computationally and in terms of accuracy. This work lays the basis for future studies, including the use of the created model in several test areas, the testing and comparison of other types of ML classifiers, and the study of optimizing hyperparameter choices for different classifiers.

References

1. Heinberg, R., Fridley, D.: Our Renewable Future: Laying the Path for One Hundred Percent Clean Energy. Island Press, Washington, DC (2016)
2. Owusu, P.A., Asumadu-Sarkodie, S.: A review of renewable energy sources, sustainability issues, and climate change mitigation. Cogent Eng. **3**(1), 1167990 (2016). https://doi.org/10.1080/23311916.2016.1167990
3. Omer, A.M.: Energy, environment and sustainable development. Renew. Sustain. Energy Rev. **12**(9), 2265–2300 (2008). https://doi.org/10.1016/j.rser.2007.05.001
4. Kamath, H.G., Ekins-Daukes, N.J., Araki, K., Ramasesha, S.K.: The potential for concentrator photovoltaics: a feasibility study in India. Prog. Photovoltaics Res. Appl. **27**(4), 316–327 (2019). https://doi.org/10.1002/pip.3099
5. Peters, I.M., Liu, H., Reindl, T., Buonassisi, T.: Global prediction of photovoltaic field performance differences using open-source satellite data. Joule **2**(2), 307–322 (2018). https://doi.org/10.1016/j.joule.2017.11.012
6. Ferrara, C., Philipp, D.: Why do PV modules fail? Energy Procedia **15**, 379–387 (2012). https://doi.org/10.1016/j.egypro.2012.02.046
7. Weselek, A., Ehmann, A., Zikeli, S., Lewandowski, I., Schindele, S., Högy, P.: Agrophotovoltaic systems: applications, challenges, and opportunities. A review. Agronomy Sustain. Dev. **39**, 1–20 (2019). https://doi.org/10.1007/s13593-019-0581-3
8. Dunnett, S., Sorichetta, A., Taylor, G., Eigenbrod, F.: Harmonised global datasets of wind and solar farm locations and power. Sci. Data **7**(1), 130 (2020). https://doi.org/10.1038/s41597-020-0469-8
9. Ladisa, C., Capolupo, A., Ripa, M.N., Tarantino, E.: Evaluation of eCognition developer and Orfeo ToolBox performances for segmenting agrophotovoltaic systems from Sentinel-2 images. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds.) International Conference on Computational Science and Its Applications, pp. 466–482. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10545-6-32
10. Plakman, V., Rosier, J., van Vliet, J.: Solar Park detection from publicly available satellite imagery. GIScience Remote Sens. **59**(1), 462–481 (2022). https://doi.org/10.1080/15481603.2022.2036056
11. Tarantino, E., Figorito, B.: Mapping rural areas with widespread plastic-covered vineyards using true color aerial data. Remote Sens. **4**(7), 1913–1928 (2012). https://doi.org/10.3390/rs4071913
12. Jiménez-Lao, R., Aguilar, M.A., Ladisa, C., Aguilar, F.J., Nemmaoui, A.: Multiresolution segmentation for extracting plastic greenhouses from Deimos-2 Imagery. ISPRS Ann. Photogram. Remote Sens. Spatial Inf. Sci. **2**, 251–258 (2022). https://doi.org/10.5194/isprs-annals-V-2-2022-251-2022
13. Novelli, A., Tarantino, E., Caradonna, G., Apollonio, C., Balacco, G., Piccinni, F.: Improving the ANN classification accuracy of Landsat data through spectral indices and linear transformations (PCA and TCT) aimed at LU/LC monitoring of a river basin. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications-ICCSA 2016: 16th International Conference, Beijing, China, 4–7 July 2016, Proceedings, Part II 16, pp. 420–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42108-7-32

14. Ladisa, C., Capolupo, A., Ripa, M.N., Tarantino, E.: Combining OBIA approach and machine learning algorithm to extract photovoltaic panels from Sentinel-2 images automatically. In: Remote Sensing for Agriculture, Ecosystems, and Hydrology XXIV, (vol. 12262, pp. 67–76). SPIE (2022). https://doi.org/10.1117/12.2636451
15. Xia, Z., et al.: Mapping the rapid development of photovoltaic power stations in Northwestern China using remote sensing. Energy Rep. **8**, 4117–4127 (2022). https://doi.org/10.1016/j.egyr.2022.03.039
16. Wang, M., Cui, Q., Sun, Y., Wang, Q.: Photovoltaic panel extraction from very high-resolution aerial imagery using region-line primitive association analysis and template matching. ISPRS J. Photogram. Remote Sens. **141**, 100–111 (2018). https://doi.org/10.1016/j.isprsjprs.2018.04.010
17. Hou, X., Wang, B., Hu, W., Yin, L., Wu, H.: SolarNet: a deep learning framework to map solar power plants in China from satellite imagery. arXiv preprint arXiv:1912.03685 (2019). https://doi.org/10.48550/arXiv.1912.03685
18. Chen, Z., Kang, Y., Sun, Z., Wu, F., Zhang, Q.: Extraction of photovoltaic plants using machine learning methods: a case study of the pilot energy City of Golmud, China. Remote Sens. **14**(11), 2697 (2022). https://doi.org/10.3390/rs14112697
19. Blaschke, T.: Object-based image analysis for remote sensing. ISPRS J. Photogram. Remote Sens. **65**(1), 2–16 (2010). https://doi.org/10.1016/j.isprsjprs.2009.06.004
20. Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F.: Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr. Intell. Lab. Syst. **83**(2), 83–90 (2006). https://doi.org/10.1016/j.chemolab.2006.01.007
21. Novelli, A., Aguilar, M.A., Nemmaoui, A., Aguilar, F.J., Tarantino, E.: Performance evaluation of object-based greenhouse detection from Sentinel-2 MSI and Landsat 8 OLI data: a case study from Almería (Spain). Int. J. Appl. Earth Observ. Geoinf. **52**, 403–411 (2016). https://doi.org/10.1016/j.jag.2016.07.011
22. Yao, Q., Li, H., Gao, P., Guo, H., Zhong, C.: Mapping irregular local climate zones from Sentinel-2 images using deep learning with sequential virtual scenes. Remote Sens. **14**(21), 5564 (2022). https://doi.org/10.3390/rs14215564
23. He, X., et al.: Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. Catena **205**, 105442 (2021). https://doi.org/10.1016/j.catena.2021.105442
24. Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V.R., Murayama, Y., Ranagalage, M.: Sentinel-2 data for land cover/use mapping: a review. Remote Sens. **12**(14), 2291 (2020). https://doi.org/10.3390/rs1214229
25. Drusch, M., et al.: Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. **120**, 25–36 (2012). https://doi.org/10.1016/j.rse.2011.11.026
26. Berger, M., Moreno, J., Johannessen, J.A., Levelt, P.F., Hanssen, R.F.: ESA's Sentinel missions in support of Earth system science. Remote Sens. Environ. **120**, 84–90 (2012). https://doi.org/10.1016/j.rse.2011.07.023
27. Sun, H., Wang, L., Lin, R., Zhang, Z., Zhang, B.: Mapping plastic greenhouses with two-temporal Sentinel-2 images and 1D-CNN deep learning. Remote Sens. **13**(14), 2820, (2021). https://doi.org/10.3390/rs13142820
28. Zhang, H., Tian, P., Zhong, J., Liu, Y., Li, J.: Mapping photovoltaic panels in Coastal China using Sentinel-1 and Sentinel-2 images and Google Earth Engine. Remote Sens. **15**(15), 3712 (2023). https://doi.org/10.3390/rs15153712

29. Wang, J., Liu, J., Li, L.: Detecting photovoltaic installations in diverse landscapes using open multi-source remote sensing data. Remote Sens. **14**(24), 6296 (2022). https://doi.org/10.3390/rs14246296
30. Mei, J., Sun, K., Xu, X.: Combing color index and region growing with simple non-iterative clustering for plant segmentation. In: 2021 6th International Conference on Image, Vision and Computing (ICIVC), pp. 119–123. IEEE (2021). https://doi.org/10.1109/ICIVC52351.2021.9526949
31. Hossain, M.D., Chen, D.: Segmentation for Object-Based Image Analysis (OBIA): a review of algorithms and challenges from remote sensing perspective. ISPRS J. Photogram. Remote Sens. **150**, 115–134 (2019). https://doi.org/10.1016/j.isprsjprs.2019.02.009
32. Achanta, R., Susstrunk, S.: Superpixels and polygons using simple non-iterative clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4660 (2017). https://doi.org/10.1109/CVPR.2017.520
33. Jia, S., et al.: Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification. IEEE Trans. Geosci. Remote Sens. **59**(2), 1437–1452 (2022). https://doi.org/10.1109/TGRS.2020.2996599
34. Wang, G., et al.: Automatic rice early-season mapping based on simple non-iterative clustering and multi-source remote sensing images. Remote Sens. **16**(2), 277 (2024). https://doi.org/10.3390/rs16020277
35. Kuhn, M.: Building predictive models in R using the caret package. J. Stat. Softw. **28**, 1–26 (2008). https://doi.org/10.18637/jss.v028.i05
36. Ramezan, C.A.: Transferability of Recursive Feature Elimination (RFE)-derived feature sets for support vector machine land cover classification. Remote Sens. **14**(24), 6218 (2022). https://doi.org/10.3390/rs14246218
37. Demarchi, L., Kania, A., Ciężkowski, W., Piórkowski, H., Oświecimska-Piasko, Z., Chormański, J.: Recursive feature elimination and random forest classification of Natura 2000 grasslands in lowland river valleys of Poland based on airborne hyperspectral and LiDAR data fusion. Remote Sens. **12**(11), 1842 (2020). https://doi.org/10.3390/rs12111842
38. Misra, P., Yadav, A.S.: Improving the classification accuracy using recursive feature elimination with cross-validation. Int. J. Emerg. Technol **11**(3), 659–665 (2020)
39. Farquad, M.A.H., Bose, I.: Preprocessing unbalanced data using support vector machine. Decis. Support Syst. **53**(1), 226–233 (2012). https://doi.org/10.1016/j.dss.2012.01.016
40. Lee, C.Y., Lee, Z.J.: A novel algorithm applied to classify unbalanced data. Appl. Soft Comput. **12**(8), 2481–2485 (2012). https://doi.org/10.1016/j.asoc.2012.03.051
41. Jhonnerie, R., Siregar, V.P., Nababan, B., Prasetyo, L.B., Wouthuyzen, S.: Random forest classification for mangrove land cover mapping using Landsat 5 TM and ALOS PALSAR imageries. Procedia Environ. Sci. **24**, 215–221 (2015). https://doi.org/10.1016/j.proenv.2015.03.028
42. Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., Homayouni, S.: Support vector machine versus random forest for remote sensing image classification: a meta-analysis and systematic review. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. **13**, 6308–6325 (2020). https://doi.org/10.1109/JSTARS.2020.3026724

43. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
44. Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P.: An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. **67**, 93–104 (2012). https://doi.org/10.1016/j.isprsjprs.2011.11.002