Computational Geosciences
https://doi.org/10.1007/s10596-024-10285-y

ORIGINAL PAPER

Analysis of the hyperparameter optimisation of four machine learning satellite imagery classification methods

Francisco Alonso-Sarría[1,2] · **Carmen Valdivieso-Ros**[1,2] · **Francisco Gomariz-Castillo**[1,2]

Received: 28 July 2023 / Accepted: 21 March 2024

Abstract
The classification of land use and land cover (LULC) from remotely sensed imagery in semi-arid Mediterranean areas is a challenging task due to the fragmentation of the landscape and the diversity of spatial patterns. Recently, the use of deep learning (DL) for image analysis has increased compared to commonly used machine learning (ML) methods. This paper compares the performance of four algorithms, Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and Convolutional Network (CNN), using multi-source data, applying an exhaustive optimisation process of the hyperparameters. The usual approach in the optimisation process of a LULC classification model is to keep the best model in terms of accuracy without analysing the rest of the results. In this study, we have analysed such results, discovering noteworthy patterns in a space defined by the mean and standard deviation of the validation accuracy estimated in a 10-fold cross validation (CV). The point distributions in such a space do not appear to be completely random, but show clusters of points that facilitate the discovery of hyperparameter values that tend to increase the mean accuracy and decrease its standard deviation. RF is not the most accurate model, but it is the less sensitive to changes in hyperparameters. Neural Networks, tend to increase commission and omission errors of the less represented classes because their optimisation lead the model to learn better the most frequent classes. On the other hand, RF and MLP prediction layers are the most accurate from a general qualitative point of view.

Keywords Machine learning · LULC · Convolutional neuronal networks · Random forest · Support vector machines · Hyperparameter optimisation

Mathematics Subject Classification (2010) 62P12 · 62P25 · 91D20 · 68T07

1 Introduction

The classification of land use and land cover (LULC) from remotely sensed imagery is crucial for several aspects of land management [16, 42], yet it may be a challenging task in some cases [42]. Some areas, such as the semi-arid Mediterranean region of south-eastern Spain, can be challenging due to their socio-economic and physical characteristics, which create a particularly fragmented landscape [7, 22] with a high diversity of spatial patterns, implying not only a wide range of different uses, but also vegetation covers at the regional scale [7]. The mixture of lithology and spectral properties of soils and the diversity of biophysical characteristics of plant species jeopardise the distinction between crops and natural vegetation, rainfed or irrigated crops, or between some anthropogenic surfaces, among others.

Traditionally, LULC classification in remote sensing has been carried out using Machine Learning (ML) algorithms such as Random Forest (RF) or Support Vector Machines (SVM), which have achieved more than satisfactory accuracy rates in the last decades [46], thanks to projects and space missions developed by different government agencies

✉ Francisco Gomariz-Castillo
fjgomariz@um.es

Francisco Alonso-Sarría
alonsarp@um.es

Carmen Valdivieso-Ros
mcarmen.valdivieso@um.es

[1] Departamento de Geografía, Universidad de Murcia, Campus La Merced, Murcia 30001, Murcia, Spain

[2] Instituto Universitario del Agua y del Medio Ambiente, Universidad de Murcia, Campus de Espinardo, Murcia 30100, Murcia, Spain

Published online: 05 April 2024

around the world to collect information for Earth monitoring, such as the NASA Landsat programme or the European Spatial Agency (ESA) Sentinel programme. The different constellations of satellites that make up the system provide a wide range of different data sources that allow the classification results to be improved [46, 52].

Using multi-source data, RF has been successfully applied to classify small agricultural areas [41], urban areas [18] or forests [39], improving classification results over other ML algorithms as well as decreasing confusion between sparse vegetation and forest. Also using RF as a classifier, [4] tested multi-source feature classification with optical and radar data combined with a DEM to accurately monitor a tropical peatland, [13] used it to effectively map forest above-ground biomass in heterogeneous mountainous regions.

However, when spatial or temporal context is important in the land cover being analysed, there may be contextual features that are difficult to extract from images [46]. For this reason, together with the rapid increase in computing power, the use of deep learning (DL) for image analysis has made great progress. However, there are still several unresolved aspects to the accurate application of ANN architectures in LULC classification, as it is a relatively new field of research. One important underdeveloped aspect is the fusion of multi-source data [45, 46]. While such data has greatly expanded its use as input information for ML algorithms, it remains underexploited in DL, especially in architectures such as Deep Convolutional Neural Networks (DCNN) [37] or Multilayer Perceptrons (MLP).

It is important to consider the quality and quantity of the training data set [1]. Sparsity or imbalance problems in the training data have a negative effect on the results of neural networks [45]. The loss function minimization process carried out in neural network training, can increase accuracy by focusing in classify accurately those classes more frequent in the training data, leading to too high omission or commission errors in the less frequent. Conventional machine learning approaches are still widely used for this reason, as algorithms such as RF or SVM are robust to small training data sets, although not necessarily to imbalanced data. In addition, generalisation capacity, which is related to the size of the training data, is still a critical challenge in DL. The same level of success is not always achieved when good performing models with their own training and test data are applied to other data sets [49]. For these reasons, it is necessary to check not only global accuracy metrics but also per class metrics. In addition, it is necessary to improve not only the predictive performance but also the understanding and interpretability of DL models [46].

In response to these issues, some efforts have been made to compare ML techniques with some DL architectures, as in [37], where a GEOBIA classification method was developed using DCNN as a classifier, among others, and its performance was studied in comparison with RF and SVM. Their results highlight the importance of the number of training samples in the accuracy results, which are similar or even lower for DCNN than for ML methods when the training data set is small. In [36], the applicability of GEOBIA techniques after classification is studied in comparison with GEOBIA, SVM and RF classification. Castelo-Cabay et al. [11] also compares the results obtained using a Deep Neural Network architecture with a pixel based and a GEOBIA classification performed with RF obtaining an overall accuracy of 87%, 43% with CNN and a pixel-based approach, but a 95% with the GEOBIA approach.

As well as comparing how different algorithms perform, other issues have been explored, such as in [28], which, in addition to comparing DCNN with MLP and SVM, proposed a procedure for the automatic construction of the training dataset. Regarding to multi-source input data, [49] compared the performance and the generalisation ability in LULC classification of one, two, and three dimensional DCNN using SAR and optical data, and [2] compared the results obtained with two different composites of medium resolution images as input to a DCNN architecture trained in one semi-arid location and tested in two other semi-arid locations.

In general, RF seems to be more accurate than SVM when no hyperparameter optimisation is performed. However, when the hyperparameters are optimised, SVM tends to perform slightly better. In fact, one of the main advantages of RF is its lack of sensitivity to the values of its hyperparameters [34]. On the other hand, SVM is very sensitive to them. Another question yet to be explored is whether some pattern might be found between the hyperparameter values and the accuracy obtained.

It is important to take into account that the more exhaustive the exploration of the hyperparameter space the more overestimated might the final overall accuracy be. So it is necessary to have a test dataset, independent of those data used to explore the hyperparameters space, to perform a final honest accuracy estimation.

The main objective of this paper is to compare four algorithms, Random Forest, Support Vector machine, Multilayer Perceptron and Convolutional Network. The first two algorithms can be considered clasical machine learning models, whereas the other two are types of deep learning models. Additionally, an exhaustive optimization of the hyperparameters of each model has been carried out in order to guarantee the maximum possible accuracy and also to evaluate the sensitivity of the models to them. Both average and standard deviation of accuracy are taken into account; the results are presented in a hyperparameter space defined by these two statistics. As an exhaustive exploration of the hyperparameter space might produce an overestimated accuracy value, a

test set will be separated from the dataset to obtain a more honest final accuracy estimation. Per class, as well as global accuracy metrics, will be analysed to check for imbalances in class accuracies.

2 Metodology

Four different classification algorithms were used to compare their suitability for application over a semi-arid Mediterranean area in south-east Spain. As Fig. 1 shows, the process starts with the acquisition of images from three different sensors, followed by the pre-process recommended for each different type of sensor, with its resampling to 10 metres and co-registration to a common spatial reference system (SRS). Features derived from each sensor, i.e. indices, textures and lidar metrics, are then extracted to form the dataset. Besides, the summer RGBI image of Sentinel-2 is used as input data for the CNN. Finally, the optimisation process of each individual algorithm takes place before the classification with optimal parameters.

2.1 Study area

The study area is the inland basin of the Mar Menor coastal lagoon in SE Spain (Fig. 2) with 1275 km^2 and a slight slope of less than 10%. It belongs to the domain of the Mediterranean semi-arid climate with irregular and scarce rainfall, usually below 300-350 mm/year. The alternation of extreme droughts and floods is common due to this high spatial and temporal variability of rainfall. Temperatures are warm throughout the year, with an average of 16°C to 18°C and an annual average maximum of more than 42°C.

The Mar Menor is the highest coastal saline lagoon in the western Mediterranean, almost closed by a 22 km long and between 100 and 1,200 m wide sand barrier called *La Manga* of the Mar Menor. The lagoon and its surroundings include the most important protection figures delivered by European laws for its unique ecological values.

The soil features in the inland area of the basin and its climate and orography make the area very well fit for agricultural purposes since ancient times, changing during the last fifty years progressively from rainfed to irrigated cultivation, thanks to the support of water transferred from river Tagus, desalination plants and underground waters. The inland area of the basin is then one of the main agricultural surfaces in Murcia Region. According to regional statistics [10], fields of irrigated grass crops alternate with irrigated dense tree crops in lower slope areas, representing near 38,000 ha. Greenhouses cover more than 1,500 ha in this area, while other types of plastics coverages as nets used to prevent birds and insects from nibbling fruit on trees, and also to prevent hail damages has been considered out of this measure.

Considering natural vegetation, there is a wide range of biodiversity and vegetation heterogeneity, mainly Mediterranean scrubs, although there are also patches of Mediterranean forest. The other main use in this territory is urban; many large urbanised surfaces, whose summer population increase is hard to quantify, can be found along the coastline delimiting the lagoon. The agricultural an residential

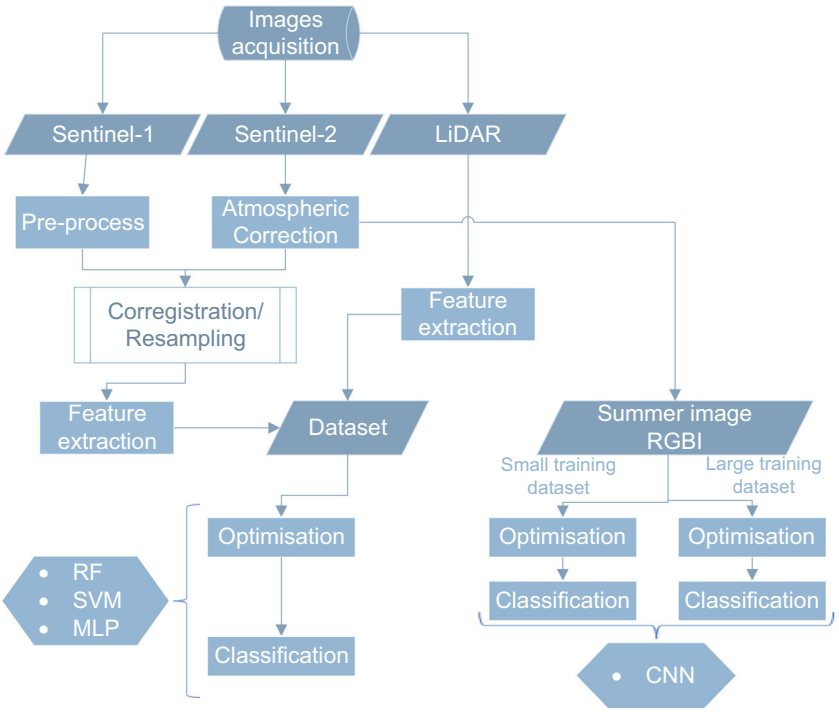

Fig. 1 Flow chart of the methodology used

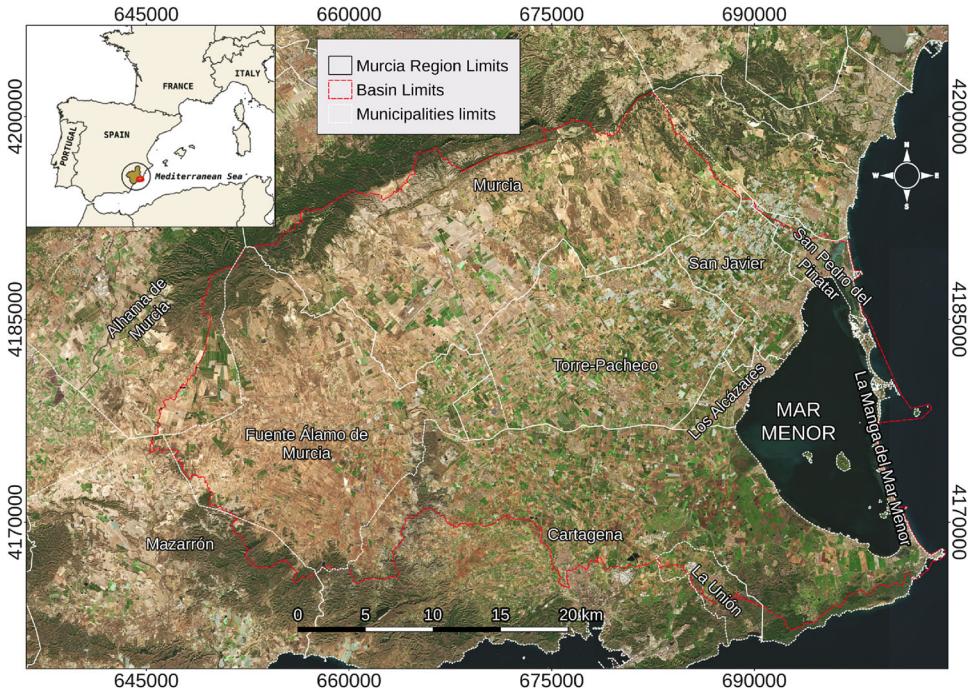

Fig. 2 Study Area: Mar Menor Basin. CRS: ETRS89/UTM zone 30N

developing in the basin have been affecting the marine ecosystem for several decades [21, 40].

2.2 Datasets

Eight SAR and MSI images (four images each) were obtained from the Copernicus Open Access Hub (Table 1). As most of the class separation problems are related to the phenology and temporal evolution of the different crops and natural covers, the dates were selected according to the sowing and harvesting calendar for most of the crops in the area.

2.2.1 Sentinel-2 data

The S2 L1C data were corrected with ACOLITE. It is a generic processor for atmospheric correction and processing for coastal and inland water applications. It supports many sensors, such as Landsat (5/7/8), Sentinel-2 (A/B), Sentinel-3 (A/B), PlanetScope, Pléiades, and WorldView. It performs the atmospheric correction using the "dark spectrum fitting" or the "exponential extrapolation" [53–55], which gave a more accurate classification than others in this study area [51]. The MSI bands from B01 to B12, except B09 and B10, were used and resampled to 10 m resolution using the nearest neighbour method.

Table 1 Sentinel-1 and Sentinel-2 images used in this study

Season	S1 (SAR)	S2 (MSI)
Autumn	2018/11/07	2018/11/08
Winter	2019/02/25	2019/02/24
Early spring	2019/04/11	2019/04/13
Late spring	2019/06/10	2019/05/19

Indices highlight fundamental interactions between spectral variables, so their use to detect biophysical patterns is an effective practice supported by a wide range of studies [9, 17, 26, 33, 43, 57]. Five common indices have been extracted: Normalized Difference Vegetation Index (NDVI, Eq. 1) [48], Tasseled Cap coefficient for Brightness (TCB Eq. 2) [32], Soil-Adjusted Vegetation Index (SAVI Eq. 3) [29], Normalized Difference Built-up Index (NDBI, Eq. 4) [14] and Modified Normalized Difference Water Index (MNDWI, Eq. 5) [56].

$$NDVI = \frac{B_{8A} - B_4}{B_{8A} + B_4} \tag{1}$$

$$\begin{aligned} TCB = {} & (0.3037 \cdot B_2) + (0.2793 \cdot B_3) + (0.4743 \cdot B_4) \\ & + (0.5585 \cdot B_8) + (0.5082 \cdot B_{11}) \\ & + (0.1863 \cdot B_{12}) \end{aligned} \tag{2}$$

$$SAVI = (1 + L)\frac{B_8 - B_4}{B_8 + B_4 + L} \tag{3}$$

$$NDBI = \frac{B_{11} - B_8}{B_{11} + B_8} \tag{4}$$

$$MNDWI = \frac{B_3 - B_{11}}{B_3 + B_{11}} \tag{5}$$

Haralick's GLCM texture metrics [24, 25] recommended in [23] were added as predictors to the dataset, computed over two summary layers per date: the first principal component

of the spectral layers (i.e. an albedo layer) and the NDVI. These metrics distinguish vertical patterns of parallel lines more than one pixel wide, using tonal differences in pairs of pixels within a predefined neighbourhood. Three metrics were used: Angular second moment (ASM, Eq. 6), Contrast (CON, Eq. 7) and Entropy (ENT, Eq. 8).

$$ASM = \sum_{i,j=0}^{N-1} P_{i,j}^2 \tag{6}$$

$$CON = \sum_{i,j=0}^{N-1} (P_{i,j}(i-j)^2) \tag{7}$$

$$ENT = \sum_{i,j=0}^{N-1} (P_{i,j}(-\log P_{i,j}) \tag{8}$$

where $P_{i,j}$ is the probability of i and j values occurring in adjacent pixels; and i and j are the labels of the column and row, respectively, of the GLCM.

2.2.2 Sentinel-1 data

SAR images were selected in the main acquisition mode over the Earth's surface using the Terrain Observations by Progressive Scans SAR (TOPSAR), the Interferometric Wide (IW), with a full swath of 250 km and 5x20 m spatial resolution in a single look. TOPSAR steers the beam from backwards to forwards in the azimuth direction, with sufficient overlap to provide continuous coverage when the three sub-swaths composed of a series of bursts are merged. The angle of incidence in this mode ranges from 29.1° to 46°. The final Ground Range Detected (GRD) product was focused, multi-looked and projected to ground range using an Earth ellipsoid model as it is composed of all bursts and sub-swaths merged and resampled to the common pixel spacing.

The pre-processing of the S1 SAR images was performed in SNAP 7.0 in batch mode and included the following steps: (1) radiometric calibration, (2) speckle filtering (with Lee sigma filter, 5x5 window size, sigma of 0.9 and 3x3 target size window), (3) terrain correction and resampling to 10 m (using the nearest neighbour model with the SRTM 1Sec HGT as the digital elevation model (DEM)), and (4) conversion to dB. The projection of the S1 imagery was set the same as that of the optical imagery for it to be used together. For this study, intensity bands of S1 IW GRD images in co- and cross-polarisation (VV, VH) were used.

In addition, the Dual Polarisation SAR Vegetation Index (DPSVI) Eq. 9 proposed by [44] has been calculated to separate bare ground from vegetation:

$$DPSVI = \frac{(\sigma^0_{VV} + \sigma^0_{VH})}{\sigma^0_{VV}} \tag{9}$$

2.2.3 LiDAR metrics

LiDAR is an active remote sensing system that uses laser pulses in the visible spectrum to record the altitude of several points on the Earth surface [19]. The Spanish *Plan Nacional de Ortofotografía Aerea* (PNOA) [30] of the Spanish Geographical Institute (IGN) includes a globla LiDAR coverage of the whole national territory with a sampling density of 0.5 points per square metre. Data for the study area was obtained from August 2016 to March 2017. The recorded points are pre-classified according to the ASPRS (American Society for Photogrammetry and Remote Sensing) standards with unknown accuracy, but the points have not been reclassified to avoid an over-complicated process. The data is available in the website of the IGN's National Centre for Geographic Information.

First, points not belonging to bare soil, vegetation, buildings or water were filtered out. After that, it was computed the proportion of points of low vegetation (**ppB**), medium size vegetation (**ppM**), high vegetation (**ppA**), buildings (**ppE**) and water (**ppH**) per each 10x10 m cell corresponding to the Sentinel-2 images. The number of medium or high vegetation points whose nearest neighbour is another medium or high vegetation point (**Nvv**) was also computed.

To obtain the heights, the altitude of the terrain extracted from the Spanish oficial DEM with 5 m resolution (also obtained from the IGN website) was substracted from each of the points. Next it is calculated the average height and standard deviation of each of the classes in each 10x10 cell, being set as 0 if the cell has no points of a given class. The resulting layers are the average height of small vegetation (**mZB**), the average height of medium size vegetation (**mZM**), the average height of high vegetation (**mZA**), the average height of building points (**mZE**), the average height of ground points (**mZG**), the standard deviation of small vegetation (**sZB**), the standard deviation of medium size vegetation (**sZM**), the standard deviation of high vegetation (**sZA**), the standard deviation of building points (**sZE**) and the standard deviation of ground points (**sZG**).

The cluster tendency of the dataset has been measured using the Hopkins statistic [27], calculated with the R package clustertend [58]:

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d} \tag{10}$$

where u_i^d is the distance of each point to its nearest neighbour, and w_i^d is the distance of m randomly chosen points to their nearest neighbour. The spatial cluster tendency (**HI**) of medium and high vegetation points was been calculated in each 10x10 cell.

Table 2 Summary of features in dataset

Dataset	Variables
S1	VV, VH
S1 indices	DPSVI
S2	B01, B02, B03, B04, B05, B06, B07, B08, B08A, B11, B12
S2 indices	NDVI, SAVI, NDBI, MNDWI, TCB
S2 texture	PC1, NDVI, Entropy, Contrast, Second angular moment
LiDAR	ppA, ppM, ppB, ppH, ppE, mZG, mZB, mZM, mZH, mZA, mZE, sZG, sZB, sZM, sZA, sZE, sZH, Hv, He, Nk, Nke, Nvv, wCv, wCd, wDv, wDd

It was also calculated the optimum number of clusters of medium and high vegetation points in each 10x10 cell (**nCl**) according to the function available in the NbClust R package [12].

The tendency of points to appear dispersed or forming clusters is measured by the Ripley's K function [38]:

$$K(d) = \frac{A}{n^2} \sum_{i=1} nC_{i,d} \tag{11}$$

where d is a given distance, A is the area of the analysed territory, n is the number of points and $C_{i,d}$ is the number of points whose distance to point i is lower than d. Values of $K(d)$ larger than the expected indicate a clustered point pattern, whereas $K(d)$ values lower than the expected indicate a regular point pattern [38]. Relative K function is calculated with the R package spatstats [6] as:

$$K_r(d) = \frac{K(d) - K_{th}(d)}{K_{th}(d)} \tag{12}$$

where $K_{th}(d)$ is the expected value of $K(d)$ assuming a random point distribution. The function is calculated at several distances to estimate the pattern at different scales.

It has been extracted 4 metrics from the K_r function for medium and high vegetation points: the maximum and minimum values (**wCv** and **wDv**, respectively) and the distances at which they occur (**wCd** and **wDd**).

In summary, the final dataset is composed by a total of 126 multi-source features summarized in Table 2.

2.3 Training areas and classification scheme

The training areas were digitised using aerial photographs from the Spanish *Plan Nacional de Ortofotografía Aerea* (PNOA) [30], acquired in 2016 and 2019 (also available from the IGN website). The representativeness of the set was improved using Isolation Forest with the methodology proposed in [3], resulting in a total of 131 polygons, distributed as shown in Table 3. Seventeen out of the 131 polygons were used for doing a final accuracy testing of the models. Other 114 polygons were used to calibrate and validate the models using a 10-fold Cross-validation approach.

The classification scheme adopted was chosen to group different related coverages in the study area. Netting is a class that consists of covering trees with nets of different mesh sizes to prevent both insects and birds from eating the fruit and to prevent hail damage. Some residual rainfed areas remain in the study area, but they are not included as a separate class in the classification scheme because most of them are in the process of being converted to irrigation or are abandoned and no longer in production. In this case, their spectral signatures are similar to those of bare soil areas and it is preferable to classify them as such.

In order to avoid strong imbalances between the classes, the number of pixels in the training dataset was extracted

Table 3 Classification scheme including the number of polygons for training and validation (N.Pol.), number of pixels per class (N.Pixels), number of randomly selected pixels per polygon as final training dataset (N.RPix), and percentage (Perc.) (DTC: Dense Tree Crops, IGC: Irrigated Grass Crops)

Id	Class	Description	N.Pol.	N.Pix.	N.RPix.	Perc.
1	Forest	Mediterranean forest	10	41616	10000	2.4
2	Scrub	Scrubland	12	10075	1200	11.9
3	DTC	Fruit and citrus trees	18	19969	1800	9.0
4	IGC	Mainly Horticultural crops	10	8658	1000	11.5
5	Impermeable	All artificial surfaces	18	47004	1639	3.5
6	Water	Water bodies	12	45050	1158	2.6
7	Bare soil	Uncovered land or low-vegetation	11	5203	1055	20.3
8	Greenhouses	Irrigated crops under plastics structures	26	145453	2600	1.8
9	Netting	Crops covered by nets	14	9258	1400	15.1

by randomly selecting 100 pixels per polygon from the 114 polygons, selected as initial training dataset with a larger number of pixels per polygon, as specified in the Table 3. If polygons had less than 100 pixels, all pixels of that polygon were selected.

2.4 Clasification algorithms

2.4.1 Random forest

Random Forest (RF) [8] is a non-parametric classification and regression method based on an ensemble of decision trees, typically between 500 and 2000, with two procedures to reduce correlation between trees: 1) each tree is trained with a bootstrapped subsample of the training data, 2) the feature used to split each node of the trees is selected in each split from a different randomly generated subset of features. These changes reduce the correlation between trees, making the whole concept of ensemble learning more meaningful [31]. Once all the trees are calibrated, each contributes with a vote to classify each new pixel. Finally, the pixel is assigned to the class with the most votes.

Using the default values for the required parameters, the number of trees (*ntree* = 500) and the number or features chosen to split the nodes ($mtry = floor(\sqrt{p})$, where p is the number of features), usually achieves high accuracy results [34]. However, in order to obtain the best possible results, the values need to be optimised by the user. In this study, a third parameter was also optimised, *maxsize*, which refers to the maximum depth allowed for trees. All the combinations of the following hyperparameter values are explored to find the most accurate combination with the training polygons:

- *ntrees* = [250,500,750,1000,1250,1500]
- *mtrys* = [5, 7, 9, 11, 13]
- *maxsize* = [2,4,8,16, 32, None]

The default value of *ntree* according to Breiman is 500. Given the number of features, the default value of mtry should be 11. The parameter **maxsize** is the maximum depth to which the trees are allowed to grow. Although in the original definition of Random Forest the depth is not limited, it is usually another hyperparameter to check, so we decided to include it. **None** means no growth limit.

2.4.2 Support vector machine

Support Vector Machine (SVM) is another of the most widely used ML algorithms for LUC classification in remote sensing [35]. It computes optimal nonlinear hyperplanes between classes in the feature space using two parameters to which SVM is extremely sensitive: a *cost* parameter that determines the flexibility of these hyperplanes, leading to underfitting or overfitting of the model if the parameter is not well optimised; and the type of *kernel* transformation to convert the nonlinear boundaries between classes into linear ones. For this study, the parameters were tested and optimised for three different kernels: the Gaussian Radial Basis Function (RBF), polynomial and sigmoid. For all kernels, the *gamma* parameter was optimised, as well as the aforementioned *cost* parameter and a specific parameter required as an independent term in the polinomial and sigmoid functions, the *coef0*. Finally, the *degree* has also been optimised for the polynomial kernel.

Hyperparameter optimisation is quite more complex in this case, not only for its highest sensitivity, but also for the larger number of parameters. In this case we carried out optimisation in three stages, focusing in each stage in the best hyperparameter values detected in the previous one. The hyperparameter combinations explored were:

- First stage:
  - **Kernel rbf**:
    - **gamma** = np.logspace(-9, 3, 13)
    - **cost** = [0.01, 0.05, 0.1, 0.25, 0.5, 1, 2, 4, 8, 16, 32]
  - **Kernel polynomial**:
    - **gamma** = np.logspace(-9, 3, 13)
    - **cost** = [0.01, 0.05, 0.1, 0.25, 0.5, 1, 2, 4, 8, 16, 32]
    - **degree** = [2,3,4,5]
    - **coef0** = [0,0.1,0.2,0.4,0.7,0.9]
  - **Kernel sigmoid**:
    - **gamma** = np.logspace(-9, 3, 13)
    - **cost** = [0.01, 0.05, 0.1, 0.25, 0.5, 1, 2, 4, 8, 16, 32]
    - **coef0** = [0,0.1,0.2,0.4,0.7,0.9]
- Second stage:
  - **Kernel rbf**:
    - **gamma** = [0.0001, 0.0005, 0.001]
    - **cost** = [16, 24, 32, 40, 48, 56, 64, 72]
  - **Kernel polynomial**:
    - **gamma** = [0.0001, 0.0005, 0.001]
    - **cost** = [16, 24, 32, 40, 48, 56, 64, 72]
    - **degree** = [2,3,4,5]
    - **coef0** = [0.7, 0.75, 0.8, 0.85, 0.9, 0.95]
- Third stage:
  - **Kernel polynomial**:
    - **gamma** = [0.001]
    - **cost** = [32, 36, 40, 44, 48, 52, 56, 60, 64]
    - **degree** = [3]
    - **coef0** = [0.7, 0.75, 0.8, 0.85, 0.9, 0.95]

2.4.3 Multilayer perceptron

A Multilayer Perceptron (MLP) consists of several layers with different numbers of neurons. All neurons in each layer are connected to all neurons in the next layer. The first (input) layers have as many neurons as features used to train the model, and the last (output) layer has as many neurons as classes. Each neuron performs a linear combination of its inputs through a weight vector, and the result is transformed through a non-linear activation function to produce the neuron's output. In this case, the hyperparameters optimised were the number of hidden layers (*nLayers*), the proportion of weights automatically set to zero to prune the network as a method of regularisation (*dropRates*), the rate at which the network learns (*learningRates*), and the number of cases analysed to decide the next step in the gradient descent process (*batch size*). It has been tried a MLP with different numbers of hidden layers (Fig. 3).

All combinations of the following hyperparameters are explored:

- *nLayers* = [1,2,3,4]
- *dropRates* = [0.1,0.2,0.3,0.4,0.5]
- *learningRates* = [10**i for i in range(-5,2)]
- *batchSizes* = [8, 16, 24, 32]

nLayers is the number of hidden layers. The *learning rate* is the rate at which the network learns. A small value can lead to suboptimal performance, but too high a value can lead to unstable behaviour. The *drop rate* is the proportion of weights that are automatically set to zero as a regularisation method by pruning the network. The *batch size* is the number of cases analysed to determine the next step in the gradient descent process.

The number of neurons in each layer varies with the number of hidden layers:

- *nLayers* = 1: [130, 260, 9]
- *nLayers* = 2: [130, 260, 134, 9]
- *nLayers* = 3: [130, 260, 176, 93, 9]
- *nLayers* = 4: [130, 260, 197, 134, 72, 9]

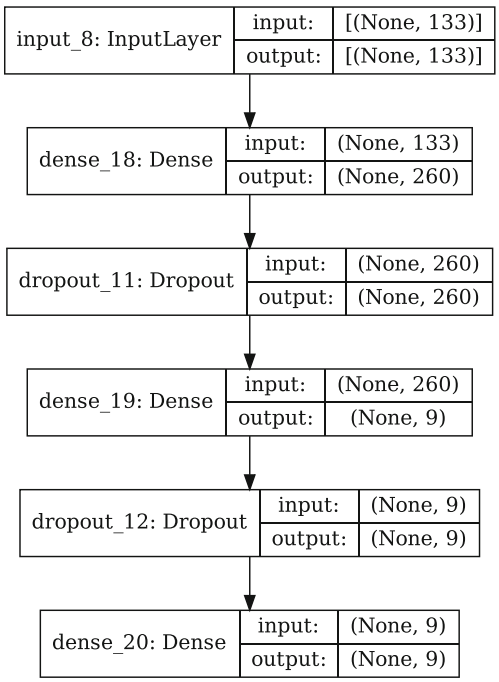

Fig. 3 Architecture of a MLP model with just two layers

2.4.4 Convolutional neural network

Convolutional Neural Networks (CNNs), which have been successfully applied to image classification, are increasingly being used for remote sensing classification. They are composed of several layers that perform different transformations on the input layer using different convolutions, which consist of applying a filter to the input to extract spatial or spectral features, or both. By combining different types of layers, the network learns to assign importance, in the form of weights and biases, to each feature extracted from the image using these transformations. The convolution layers take an image as input and distinguish between objects in the image based on the colour bands in which the image is composed. Using a filter of a given size, they extract high-level features from the image, but not limited to it. It can also extract low-level features such as colour or gradient orientation, or reduce or increase the dimensionality of the image. The convolution layer is followed by a variety of different types of layers, such as pooling, which provide statistics within the window of the image covered by the kernel size. The most commonly used are the maximum (max pooling) and the average (average pooling). Finally, a fully connected layer usually performs the classification task and gives the output layer, with as many neurons as possible classes in the scheme.

For this comparative study, an architecture was tested with 2 convolutional layers (Fig. 4) using the *relu* activation function, with a max-pool layer between them, followed by a fully connected layer of 64 neurons using the same activation function, to finally obtain the output layer using a *softmax* activation. For this experiment, only 4 bands, visible and near infrared, of a single date image, were used.

The hyperparameter optimization was also carried out in two stages, with the second set established based on the results of the first:

- First stage
  - *learning rate*: [0.01,0.005,0.001]
  - *drop rate*: [0,0.25,0.5]

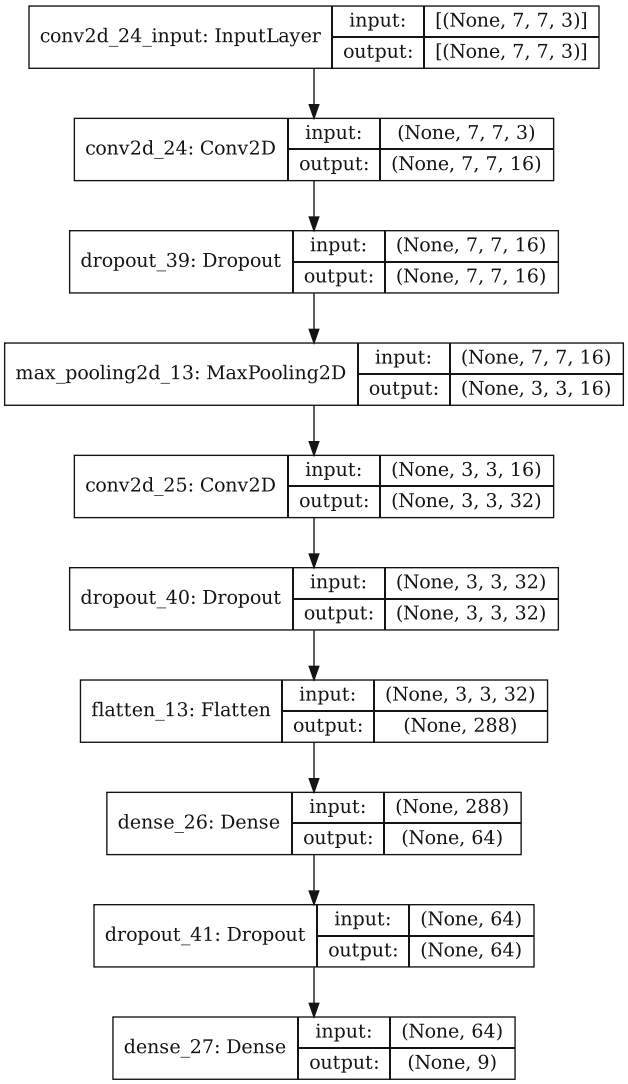

Fig. 4 Architecture of a the CNN model

  - *batch size*: [8,16,32]

- Second stage
  - *learning rate*: [0.005, 0.0075, 0.01, 0.0125]
  - *drop rate*: [0, 0.1, 0.2, 0.3]
  - *batch size*: [16,32]

3 Results

3.1 Hyperparameter optimisation

Figure 5 shows the mean and standard deviations of accuracy values obtained with 10-fold CV with RF for different hyperparameter sets. Obviously, in these figures the best combination is the one that maximises the mean and minimises the standard deviation. The figure shows a zoom to the best combinations in the space defined by the mean and the standard deviation. The first conclusion is that accuracy results are not very high. In addition, it can be seen that the model is not very sensitive to *ntree* or *mtry*, as there is a large diversity of values in the best models. On the contrary, the model is sensitive to the maximum size of the trees as the highest accuracy values are reached only with values of 32 or None. The best model achieved an accuracy of 0.868 in the training and cross validation process with values of *ntree* = 750, *mtry* = 5 and *maxsize* = None. It is noteworthy the clustered pattern of the results and that most of the parameter combinations are in the highest accuracy cluster.

Figure 6 shows the results of the best parameter combinations for SVM, where the highest mean of accuracies is achieved with a polynomial kernel of *degree* = 3, a *cost* = 56, a *gamma* = 0.001 and a *Coef0* = 0.70. In this case there is an interesting pattern with most of the models in a cluster of results around accuracy = 0.3 and the other around accuracy = 0.8. The intermediate mean accuracy values are accompanied by larger standard deviations. There is no clear pattern of hyperparameter values producing good or bad results, which difficulties the optimization of this classification model. The polynomial kernel seems to produce the best results, that is the reason why the 3 stage optimisation process ends focusing on it. However, it is the kernel with more parameters, which gives it an advantage and also makes it the most represented. The bottom plot in Fig. 6 shows a zoom to the highest mean accuracy area. In general, low values of cost and coef0 seems to work better.

Figure 7 shows the results for a MLP. The accuracy is clearly higher than with Random Forest or Support Vector Machines. The best model obtained a mean accuracy of 0.923 and a standard deviation of 0.051. Such model had 2 hidden layers a learning rate of 0.01, a drop rate of 0.1 and batch size of 16. However, there is no clear hyperspace values giving the best results.

Figure 8 shows the results obtained with the CNN. The highest accuracy (0.951) was obtained with a learning rate of 0.005, a drop rate of 0.25 and batch size 32. Once again, there is no clear hyperspace values giving the best results.

3.2 Test data

The hyperparameter optimisation carried out with the models might lead to an overestimation of the accuracy values obtained by the cross-validation process. Test data polygons were randomly selected from the labelled data set prior to the cross validation. Table 4 shows overall accuracies and kappa indices obtained by each algorithm with the test data; the standard deviations appear between parentheses. These standars deviations are calculated from the confusion matrix following [47], and are smaller than those obtained during the optimisation. The accuracy results are higher than those

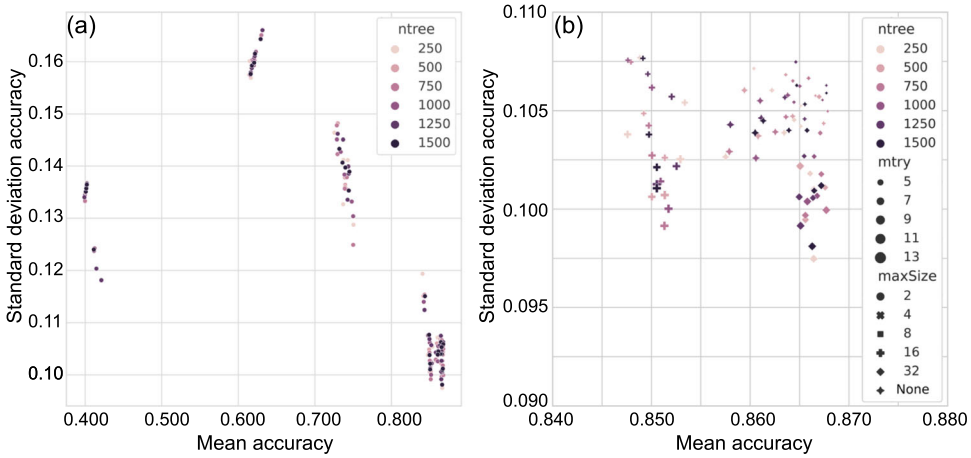

Fig. 5 Accuracy mean and standard deviations of accuracy values obtained with 10-fold CV with RF for different hyperparameter sets for (a) per each parameter combination and (b) detailed zoom to the most accurate models

obtained with cross-validation. Such a result is not usually to be expected, but can happen if the randomly selected test happens to be particularly low-noise. All the obtained metrics are similar in the case of RF, SVM and MLP, being the latter slightly better than the others. It is noteworthy that CNN shows a substantial decrease in test accuracy although validation accuracy is quite high.

Tables 5, 6, 7, 8, and 9 show the confusion matrices and omission and commission errors of each model. Figures 9, 10, 11, 12, and 13 show graphically the different errors that appear in the confusion matrices.

Classification of the test data with the best RF model achieved an accuracy of 0.932 and a kappa index of 0.923. The graph from Fig. 10 shows the main omission errors between impermeable (class 5) with scrub (class 2), bare soil (class 7) and netting (class 9), and between classes of plastic covering of crops, greenhouses and netting (classes 8 and 9 respectively) with themselves or, to a lesser extent, with impermeable (class 5). Almost every other class was well classified, but the traditionally less separable classes, dense tree crops, irrigated grass crops, were particularly well separated, as shown in Table 10, with accuracy metrics by class.

On the test data, the SVM model obtained an accuracy of 0.931 and a kappa index of 0.919. The proportion of omission and commission errors (Fig. 10 and Table 6) is mostly low,

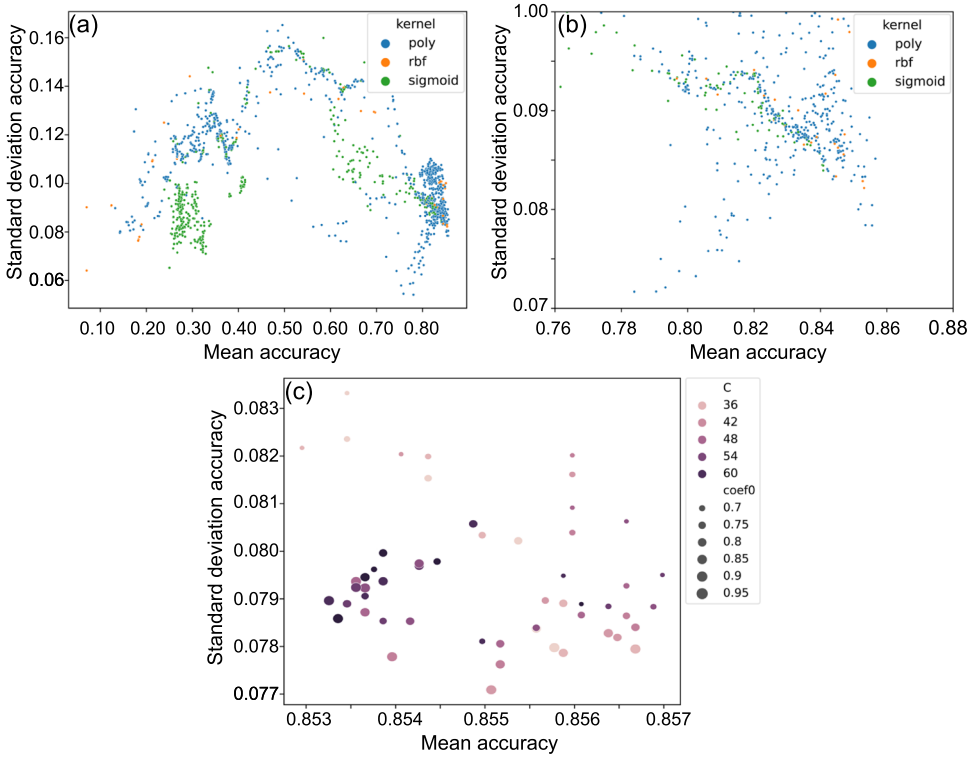

Fig. 6 Accuracy mean and standard deviations of accuracy values obtained with 10-fold CV with SVM in (a) the first optimization stage, (b) detailed zoom to the best parameter combinations and (c) results after the third optimization stage

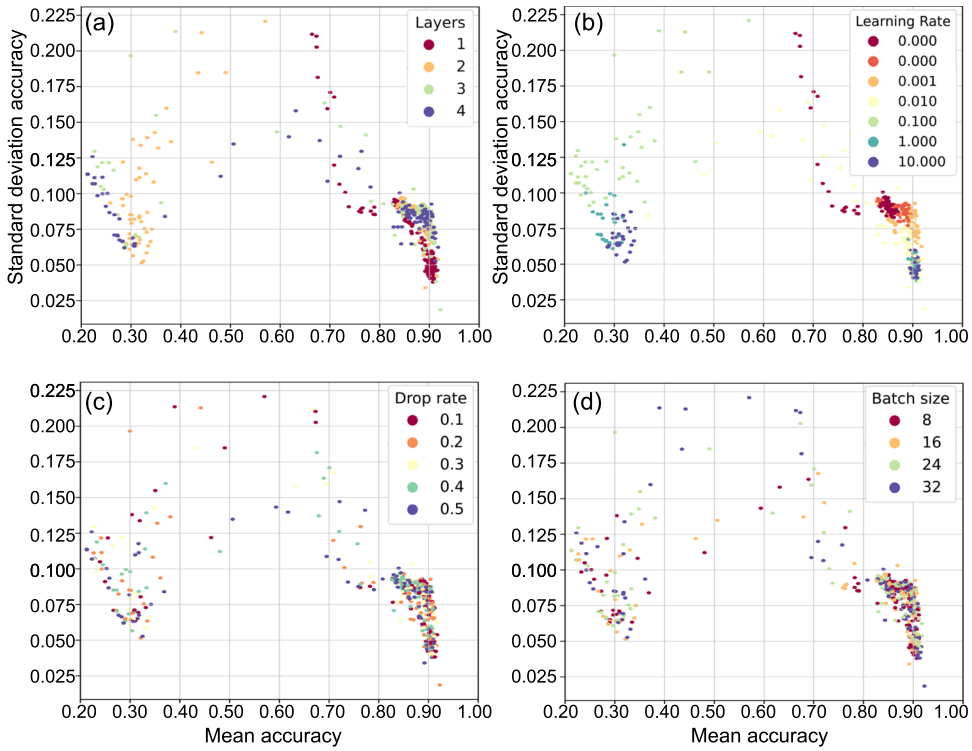

Fig. 7 Accuracy mean and standard deviation of accuracy values obtained with 10-fold CV with MLP per each parameter combination

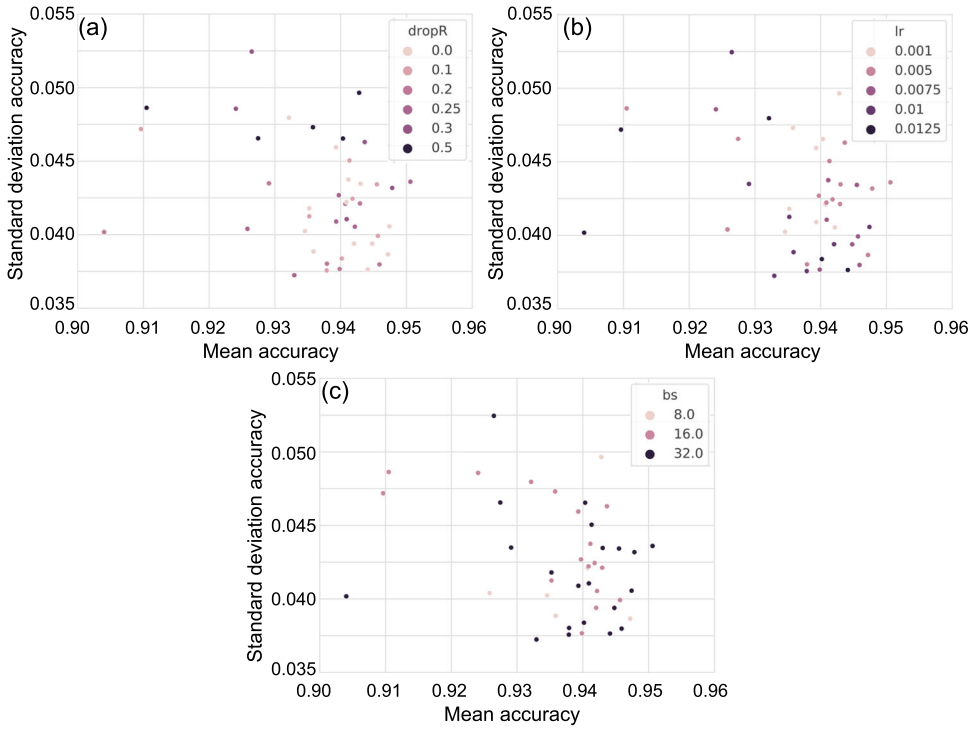

Fig. 8 Accuracy mean and standard deviation of the 10 folds per each parameter combination with CNN. Zoom to the best parameter combinations

Table 4 Overall Accuracy (OA) and Kappa index (K) obtained using test data in classification with each algorithm with optimal hyperparameters

	RF	SVM	MLP	CNN	CNN_{full}
OA	0.932 (0.0044)	0.931 (0.0047)	0.939 (0.0044)	0.852 (0.0065)	0.969 (0.0007)
K	0.923 (0.0052)	0.919 (0.0055)	0.928 (0.0052)	0.826 (0.0077)	0.958 (0.0001)

The standard deviation of the metrics is shown in brackets. CNNa refers to the model calibrated with the same pixels that the others and CNNb refers to the model with the larger dataset

Table 5 Confusion matrix of the optimised RF model (DTC: Dense tree crops, IGC: Irrigated grass crops, Com.Err.: Commission error, Omis.Err.: Omission error)

	Forest	Scrub	DTC	IGC	Imp.	Water	Bare soil	Greenhouses	Netting	Com.Err.
Forest	193	7	0	0	0	0	0	0	0	0
Scrub	0	100	0	0	0	0	0	0	0	0.405
DTC	0	0	499	0	0	0	0	0	1	0.006
IGC	0	0	0	300	0	0	0	0	0	0
Impermeable	0	57	3	0	627	0	33	1	9	0.022
Water	0	0	0	0	0	100	0	0	0	0
Bare Soil	0	4	0	0	0	0	296	0	0	0.106
Greenhouses	0	0	0	0	3	0	2	478	17	0.059
Netting	0	0	0	0	11	0	0	29	160	0.144
Omis.Err.	0.035	0	0.002	0	0.141	0	0.013	0.044	0.2	

Table 6 Confusion matrix of the optimised SVM model (DTC: Dense tree crops, IGC: Irrigated grass crops, Com.Err.: Commission error, Omis.Err.: Omission error)

	Forest	Scrub	DTC	IGC	Imp.	Water	Bare soil	Greenhouses	Netting	Com.Err.
Forest	170	30	0	0	0	0	0	0	0	0.012
Scrub	0	100	0	0	0	0	0	0	0	0.500
DTC	0	0	500	0	0	0	0	0	0	0.006
IGC	0	0	0	300	0	0	0	0	0	0
Impermeable	2	51	2	0	612	3	50	3	7	0.022
Water	0	0	0	0	0	100	0	0	0	0.029
Bare Soil	0	19	1	0	5	0	275	0	0	0.154
Greenhouses	0	0	0	0	7	0	0	490	3	0.039
Netting	0	0	0	0	2	0	0	17	181	0.052
Omis.Err.	0.15	0	0	0	0.162	0	0.083	0.02	0.095	

Table 7 Confusion matrix of the optimised MLP model (DTC: Dense tree crops, IGC: Irrigated grass crops, Com.Err.: Commission error, Omis.Err.: Omission error)

	Forest	Scrub	DTC	IGC	Imp.	Water	Bare soil	Greenhouses	Netting	Com.Err.
Forest	196	4	0	0	0	0	0	0	0	0.005
Scrub	0	100	0	0	0	0	0	0	0	0.315
DTC	1	1	496	0	1	0	0	0	1	0.006
IGC	0	0	0	295	0	0	0	0	5	0.003
Impermeable	0	40	3	1	649	8	15	1	13	0.069
Water	0	0	0	0	0	100	0	0	0	0.099
Bare Soil	0	1	0	0	32	3	264	0	0	0.057
Greenhouses	0	0	0	0	13	0	1	475	11	0.048
Netting	0	0	0	0	2	0	0	23	175	0.146
Omis.Err.	0.02	0	0.008	0.017	0.111	0	0.12	0.05	0.125	

Table 8 Confusion matrix of the optimised CNN model (DTC: Dense tree crops, IGC: Irrigated grass crops, Com.Err.: Commission error, Omis.Err.: Omission error)

	Forest	Scrub	DTC	IGC	Imp.	Water	Bare soil	Greenhouses	Netting	Com.Err.
Forest	133	63	0	0	0	0	0	0	0	0
Scrub	0	98	0	0	0	0	0	0	0	0.553
DTC	0	4	484	0	0	0	0	0	2	0.051
IGC	0	0	22	272	0	0	0	0	0	0.004
Impermeable	0	53	4	1	594	0	49	60	23	0.080
Water	0	0	0	0	0	98	0	0	0	0
Bare Soil	0	1	0	0	10	0	283	0	0	0.148
Greenhouses	0	0	0	0	1	0	0	488	1	0.246
Netting	0	0	0	0	41	0	0	99	56	0.317
Omis.Err.	0.321	0	0.012	0.075	0.242	0	0.037	0.004	0.714	0.321

Table 9 Confusion matrix of the optimised CNN model using the whole sample of training data (CNN_{full}) (DTC: Dense tree crops, IGC: Irrigated grass crops, Com.Err.: Commission error, Omis.Err.: Omission error)

	Forest	Scrub	DTC	IGC	Imp.	Water	Bare soil	Greenhouses	Netting	Com.Err.
Forest	4155	13	0	0	0	0	0	0	0	0
Scrub	0	1520	0	0	0	0	0	0	0	0.202
DTC	0	35	5122	0	7	0	0	4	0	0.018
IGC	0	0	12	3046	0	0	0	4	6	0
Impermeable	0	336	81	0	22059	0	90	164	272	0.029
Water	0	0	0	0	0	14592	0	0	0	0
Bare Soil	0	0	0	0	372	0	987	7	8	0.084
Greenhouses	0	0	0	0	133	0	0	2434	22	0.071
Netting	0	0	0	0	142	0	0	7	326	0.486
Omis.Err.	0.003	0	0.009	0.007	0.041	0	0.282	0.06	0.314	

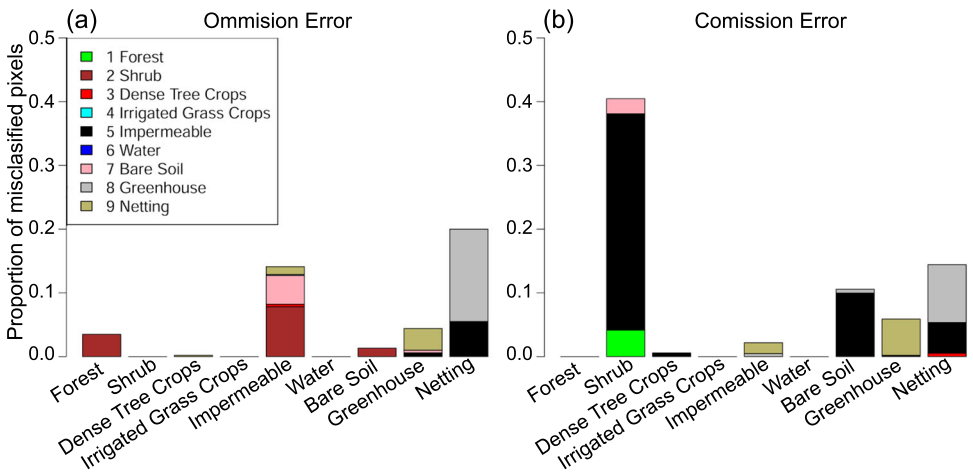

Fig. 9 Omission and Commission errors with the optimised RF model

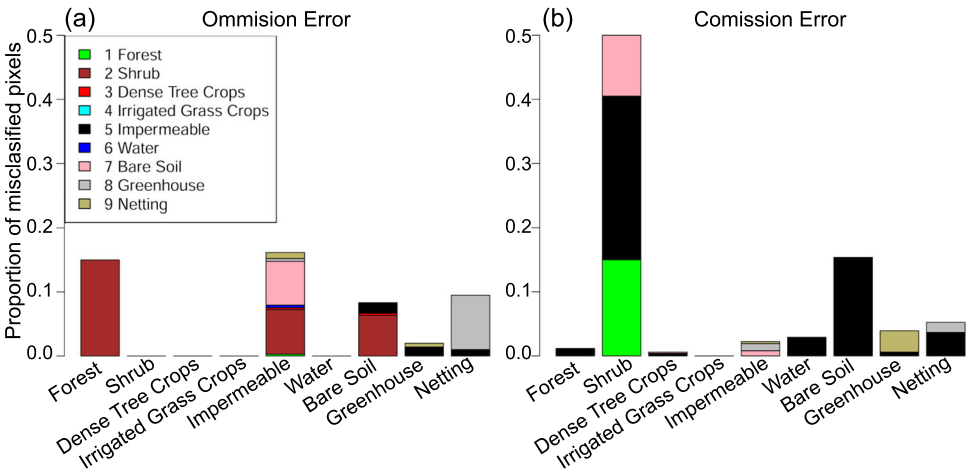

Fig. 10 Omission and commission errors with the optimised SVM model

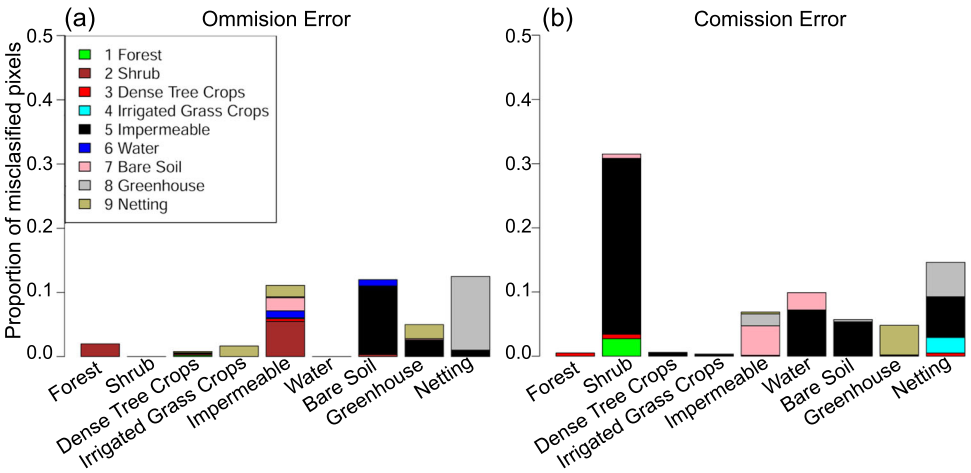

Fig. 11 Omission and commission errors with the optimised MLP model

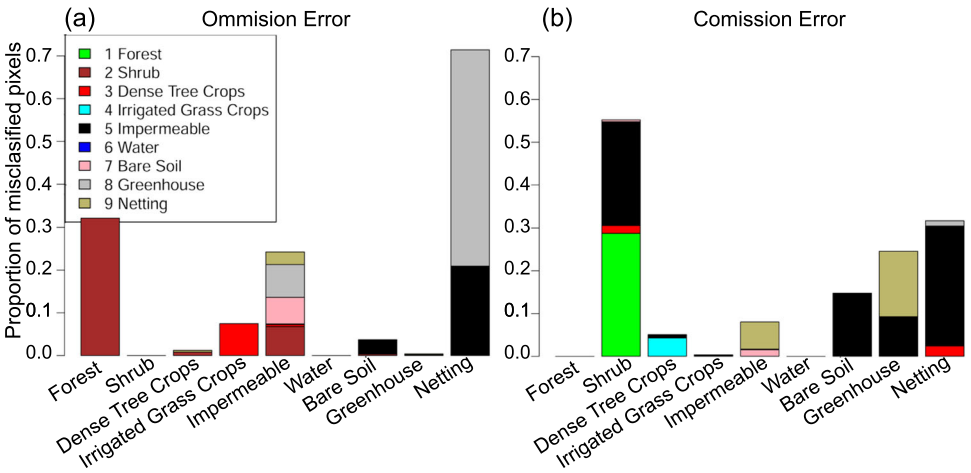

Fig. 12 Omission and commission errors obtained with the optimised CNN model

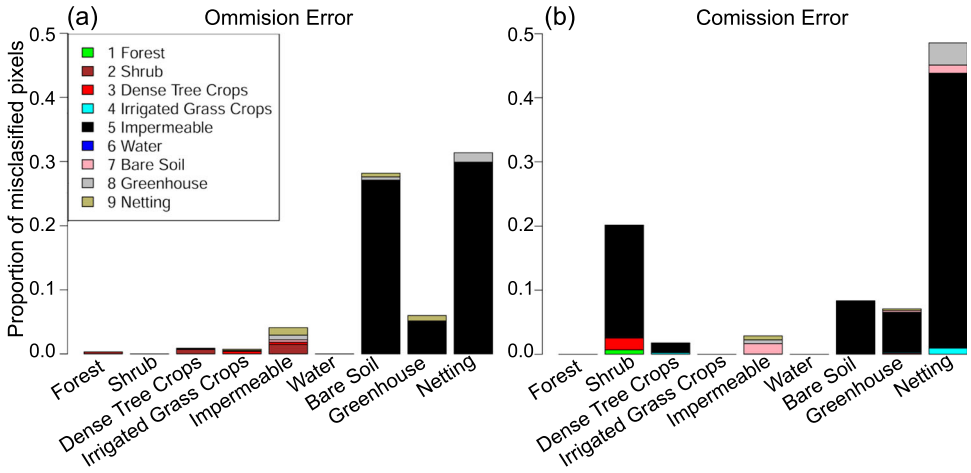

Fig. 13 Omission and commission errors with the optimised CNN model using the whole sample of training data (CNN_{full})

around 0.15, except for the commission error of scrub, where the confusion of this natural vegetation with impermeable and bare soil reaches 0.5. Confusion between netting and greenhouses seems to have been partially solved, as netting is still the most difficult to identify. However, impermeable, bare soil and scrub are still confused with forest and, in the case of impermeable, also with greenhouses and netting. In general, both classes of crops have been well classified, as can be seen in Table 10, with metrics of accuracy per class.

The accuracy and kappa index of the MLP model reach values of 0.939 and 0.928 respectively. Even though it is not the most accurate model, the proportion of omission and commission errors is less than 0.15, except for the commission error for scrub, which is mainly confused with impermeable. Again, the most problematic classes are impermeable, scrub, bare soil, greenhouses and netting, which are mainly confused with each other (Fig. 11 and Table 7). Nevertheless, the accuracy metrics per class (Table 10) show a generally good performance for all classes.

In the case of the CNN model, there is a clear decrease in the accuracy metrics estimated on the test data with respect to the previous models, reaching values of 0.852 and 0.826 for accuracy and kappa index respectively (Figs. 12 and 13 and Tables 8 and 10). The omission and commission errors of class 9, netting, are quite large, with confusion with class 8, greenhouses. In the case of commission errors, class 2 Shurb is noteworthy, with high confusion with classes 1 Forest and 9 Netting.

We thought that the reason might be that a CNN model calibrated with a reduced dataset might not generalise well.

Table 10 Precision, recall and balance accuracy per class obtained with models (DTC: Dense tree crops, IGC: Irrigated grass crops)

		Forest	Scrub	DTC	IGC	Imp.	Water	Bare soil	Greenhouses	Netting
RF	Precision	0.965	1.000	0.998	1	0.859	1	0.987	0.956	0.800
	Recall	1.000	0.595	0.994	1	0.978	1	0.894	0.941	0.856
	Balanced.Accuracy	0.500	0.798	0.872	–	0.549	–	0.896	0.759	0.629
SVM	Precision	0.850	1.00	1.000	1	0.838	1.000	0.917	0.980	0.905
	Recall	0.988	0.50	0.994	1	0.978	0.971	0.846	0.961	0.948
	Balanced.Accuracy	0.525	0.75	0.997	–	0.542	0.985	0.756	0.814	0.646
MLP	Precision	0.980	1.000	0.992	0.983	0.889	1.000	0.880	0.950	0.875
	Recall	0.995	0.685	0.994	0.997	0.931	0.901	0.943	0.952	0.854
	Balanced.Accuracy	0.597	0.842	0.711	0.582	0.652	0.950	0.625	0.721	0.700
CNN	Precision	0.997	1.000	0.991	0.993	0.959	1	0.718	0.940	0.686
	Recall	1.000	0.798	0.982	1.000	0.971	1	0.916	0.929	0.514
	Balanced.Accuracy	0.500	0.899	0.826	0.500	0.690	–	0.553	0.737	0.594
CNN2	Precision	0.679	1.000	0.988	0.925	0.758	1	0.963	0.996	0.286
	Recall	1.000	0.447	0.949	0.996	0.920	1	0.852	0.754	0.683
	Balanced.Accuracy	0.500	0.724	0.881	0.520	0.567		0.835	0.871	0. 420

Therefore, we decided to calibrate the model with the full, albeit unbalanced, dataset. So, there are two different CNN models with two different training datasets:

- *CNN*: CNN using the same number of pixels for training as in the rest of the models (column *N.RPix.* in Table 3).
- CNN_{full}: CNN training with all pixels of 114 polygons, selected as initial training.

The results are shown in Fig. 13 and Tables 9 and 10. The accuracy is quite high, but in both cases, especially in the model calibrated with the smaller dataset, there is a significant increase in the omission and commission errors in some of the classes. The accuracy and kappa index of the CNN_{full} reach values of 0.969 and 0.959 respectively. Although it is higher than the other models, the omission and commission errors of class 9, netting, are notable (Fig. 12 and Table 8), as well as the tendency to confuse impermeable with almost every other class except both crops and forest. Even the accuracy metrics per class are generally worse, as shown in Table 10.

However, performance per class is quite different by algorithm, as Table 10 shown.

Figure 14 shows the predictions obtained with the five models for the whole study area. Figure 15 shows the results in an enlarged area and their comparison with a high resolution orthophoto. RF and MLP give similar results, whereas SVM seems to predict more bare ground than is actually present. Convolutional Neural Networks, on the other hand, seem to predict more urban (impervious) areas than actually

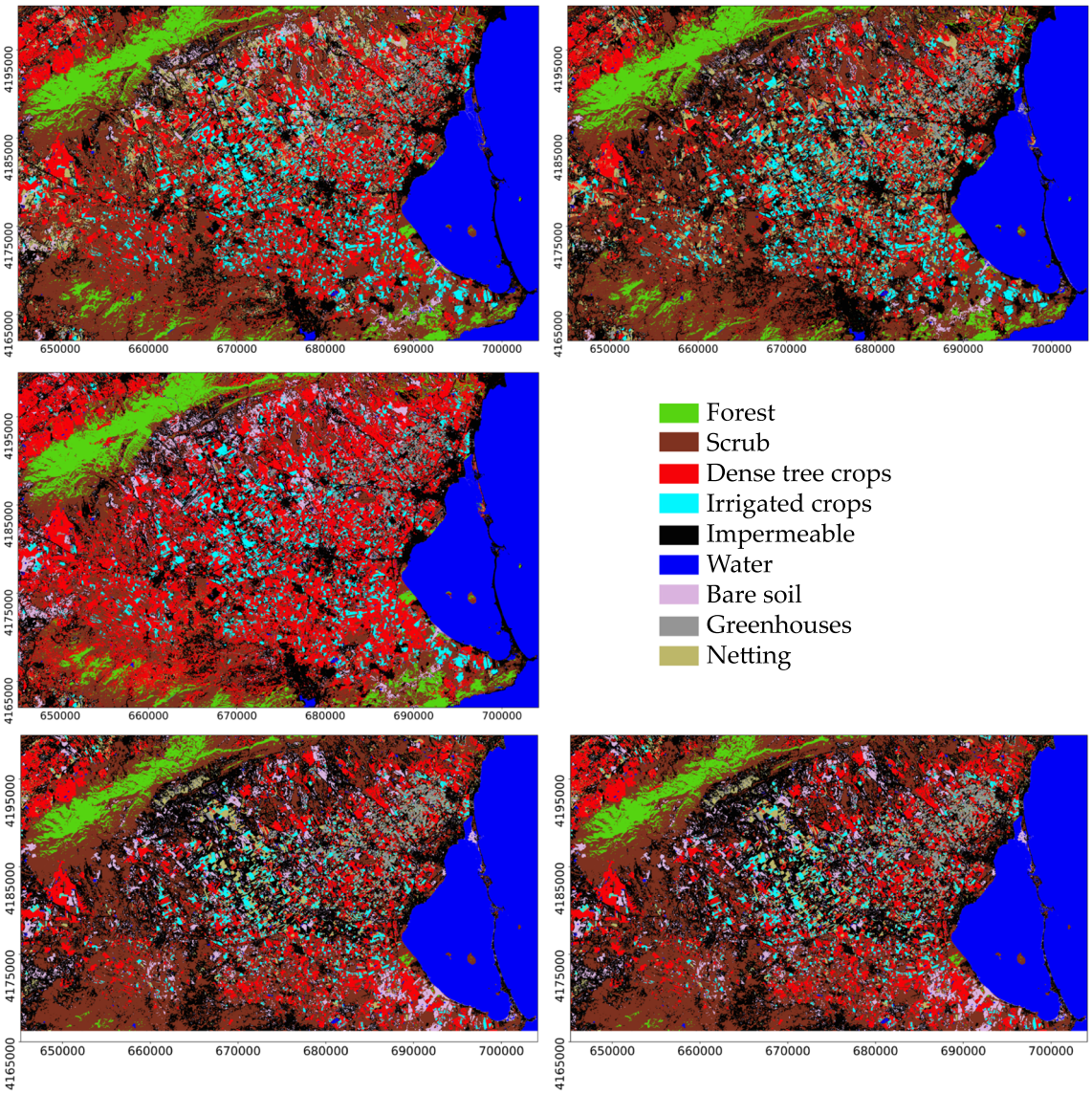

Fig. 14 Prediction of the RF (upper left), SVM (upper right), MLP (middle left), CNN (lower left) and CNN with the full dataset (lower right) models

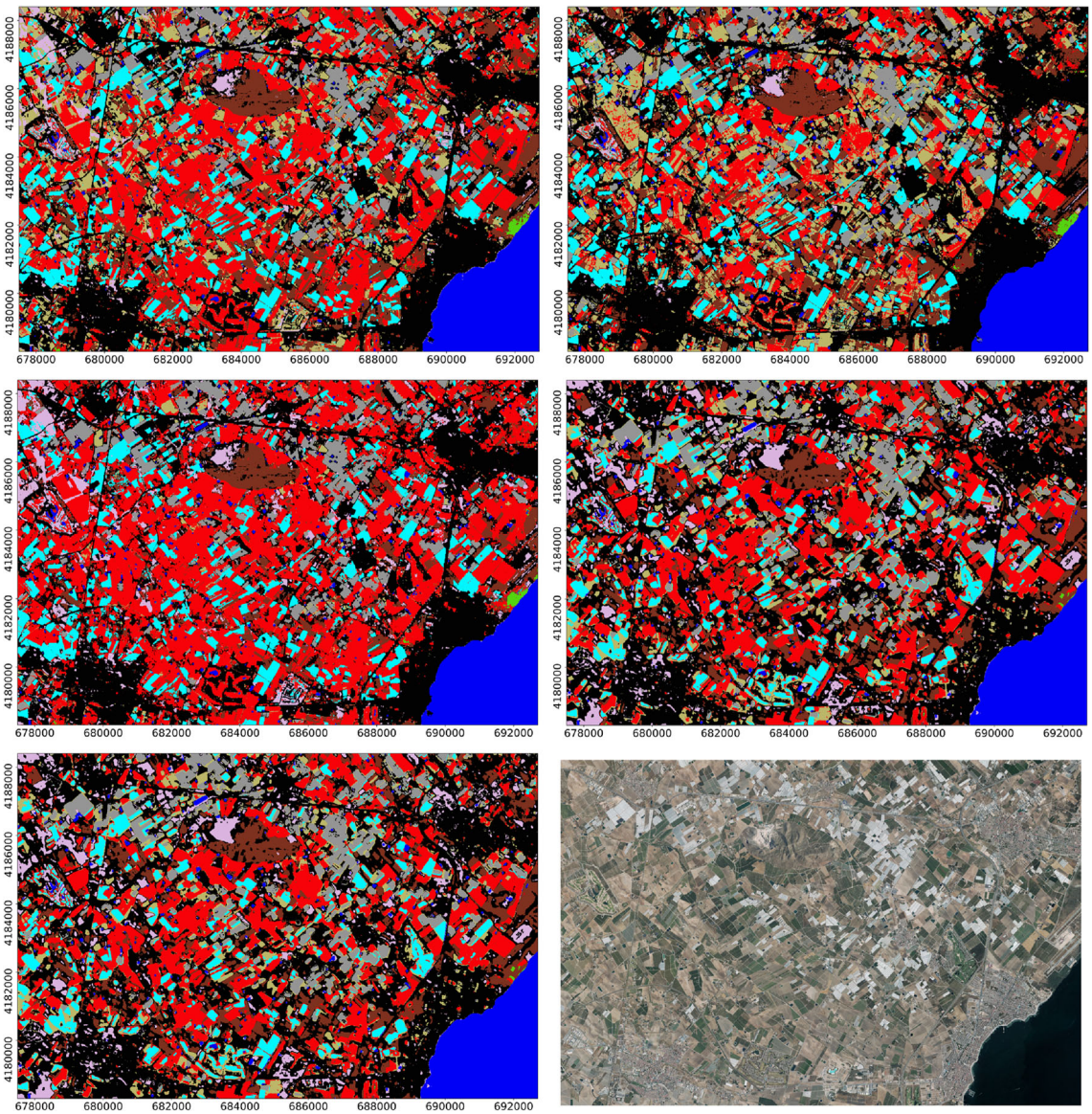

Fig. 15 Prediction of the RF (upper left), SVM (upper right) and MLP (middle left), CNN (middle right) and CNN with the full dataset(lower right) models and comparison with a high resolution ortophotograph (lower right)

exist. The geometry of the plots seem also more rounded with the CNN classifications.

4 Discussion

The results obtained in this research show that all the classification algorithms tested have similar accuracy in this particular study. Previous studies, such as [15], had already shown similar results using different ML algorithms.

Valdivieso-Ros et al. [52] found that the overall classification accuracy reached with RF was more than 0.90 using features from Sentinel-1, Sentinel-2 and LiDAR, outperforming results from SVM and MLP models to classify 9 different types of LULC in a semi-arid Mediterranean region. Similarly, [59] achieved more than 90% overall accuracy over k Nearest Neighbors (kNN), SVM and Artificial Neural Network (ANN) models in a tropical African region to monitor, quantify and map LULC and its changes. Other studies found that SVM achieves better results than RF, e.g. [50], where it was successfully applied for LULC classification using airborne LiDAR and aerial photographs reaching good results in most of the classes included in the classification scheme, or in [5] for LULC classification in an Egyptian governorate over a period of 20 years reached a *kappa index* above 0.91, the highest compared to those obtained using Maximum Likelihood or RF. Ghayour et al. [20] evaluated the performance of SVM, ANN, Maximum Likelihood Classification

(MLC), and Minimum Distance (MD) to produce LULC maps using data from Sentinel-2 and Landsat-8 satellites and concluded that with optimised parameters, SVM achieved the highest accuracy, over 95% with Sentinel-2 data, in an Iranian province with Mediterranean climate. Occasionally, RF and SVM have shown equal results, as in the research carried out in [15] in a large area of semi-arid Mediterranean using Geographic Object Based Analysis (GEOBIA), where the comparison of the performance of RF and SVM with default parameters among five classification algorithms outperformed the classification metrics of the other three with similar values.

Ali and Johnson [2] compared two different composites of medium resolution images as input to a DCNN architecture. They obtained in each test site of the two used accuracies of 91.4% and 94.8%, and kappa values of 0.88 and 0.93 with the four band composite, while using the second composition with ten bands, only achived 85.1% and 88.8% of accuracy and kappa of 0.79 and 0.85. One of the main complications of DL algorithms, as well as with SVM, is the need for a exhaustive optimization of hyperparameters, as final accuracy can be very sensitive to their values.

In [18], RF was compared with Extreme Gradient Boosting, and the latter outperformed the former only in terms of processing time. Results with RF alone were satisfactory in terms of accuracy, integrating multi-temporal optical and SAR data for classification of urban areas, reducing misclassification between vegetation classes as forest or low vegetation, and labelling the water and urban classes almost perfectly. In this study, RF also performed well in labelling the water class. For the Impermeable class, a certain percentage of misclassification with bare soil and scrub still persist, given that urban and peri-urban developments are usually adjacent to fields in various stages of production, or even abandoned, awaiting a change of use.

The results of [41] show how RF outperforms the accuracy obtained with SVM, but not that obtained with ANN, as also happened in this comparison, while [39] used it to classify a heterogeneous Mediterranean forest area, obtaining a high value of overall accuracy (98.13%) with training data, and a low proportion of misclassifications between high and low vegetation classes. In line with these results, also in this study, RF presents less problems to correctly label vegetation classes than others, being the most problematic those where the reflectance of artificial surfaces dominates.

However, a CNN architecture calibrated with four bands of a single date image is slightly higher, about 0.97. Ali and Johnson [2] reached 97.7% in another semi-arid area in Pakistan with the same four combination bands VIS-NIR, which outperformed the accuracy and separability achieved with a 10 combination bands including also red-edge and SWIR bands, which only achived an accuracy of 95.8%, also below those obtained in our research. Unfortunately, due to its sensitivity to imbalanced data, it has had serious issues with the impermeable class, which should be resolved by reviewing the training and validation datasets.

The results obtained in this study also show that, depending on the type of neural network used in the classification process with multitemporal and multi-source data, the results can be quite similar to those obtained with RF or SVM. In this case, the accuracy of MLP is only better when comparing the omission and commission errors of all classes in general, so that the good results obtained with the use of multitemporal and multisource data with RF or SVM, are not outperformed by MLP. Moreover, the use of multi-source data, also considered as feature level fusion by some authors [60] and an underexplored issue [45, 46], does not make a major difference between RF, SVM and MLP in this particular study area, while [60] found worse results with MLP doing a similar comparative study testing different levels of fusion data using optical and SAR data and features extracted from both datasets.

Using CNN, the results show the influence of the quality and quantity of the training data, as pointed out by [45]. With the same training data as the rest, with 100 or less pixels per training polygon, CNN achieves good results in validation, as it does with the larger data set with all possible pixels of training polygons. However, on the test data, the accuracy drops with the small dataset, indicating a lack of good generalisation capacity in the model. Conversely, when the larger data set is used for training (CNN_{full} in this study), the classification accuracy on the test data reaches a value of almost 0.97, which is in line with the results obtained by other researches, such as those presented in [37], comparing patch-based CNN and full CNN. The CNN architecture was tested on a four-band single date image, which [2] found to be more robust in semi-arid locations than other band combinations for separating confusing land cover. The performance of the CNN with both datasets in training and the larger one in test is consistent with the conclusions of [2] for almost all other classes except the impervious class. The research reviewed in [45] pointed out that CNNs are extremely sensitive to an unbalanced dataset, which would affect the results obtained. The problem observed in the opaque class seems to respond to this. The total number of pixels of this class used in the larger training dataset, 47004, would not be sufficiently representative of the total area covered by this class and the intra-class variability it has, as it includes not only the urban class but also transport and industrial infrastructure.

The problem with CNNs is that the training areas were selected in the core of the identified polygons to avoid uncertainties associated with the boundary pixels. A convolutional network will have no problem in predicting such pixels, since the whole window around them will be homogeneous with the reference class. On the contrary, border pixels will have more diversity within their windows. Urban areas (class

impervious) are the most heterogeneous class in the study area, due to the different colours and the presence of parks and gardens within them. We believe this is the reason why there is an overestimation of urban areas in the CNN prediction maps. CNN seems also be acting as a mode filter, generating plots larger and rounder than the real ones.

5 Conclusions

The usual approach in the optimisation process of a LULC classification model is to keep the best model in terms of accuracy without analysing the rest of the results. In this study, we have analysed such results, discovering noteworthy patterns in a space defined by the mean and standard deviation of the validation accuracy estimated in a 10-fold CV. Four clusters appear with RF and 2 with SVM and MLP. In all cases, the largest cluster corresponds to the most accurate models. RF is the model where the worst models have the highest accuracy; it seems that although it is not the most accurate model, it is the most robust.

It is difficult to establish clear relationships between parameter values and accuracy. The only conclusions that can be drawn are that in RF trees should be allowed to grow fully, in SVM the polynomial kernel seems to work better, and in MLP good models seem to reduce the standard deviation when the number of layers is reduced and the learning rate is increased. However, for all models, the standard deviation does not seem to be very large for the most accurate models. This lack of relation difficults but highlights the need for optimisation.

ANN models optimise categorical cross-entropy, a metric that is not accuracy, but can be considered related. Thus, the models have a tendency to increase accuracy by learning to classify particularly well those classes with a larger sample size, leading to an increase in the commission and omission errors of the less represented classes. CNN seems to need a larger sample to achieve good accuracy results. In any case, the CNN models were calibrated with 4 bands from a single summer image, so with much less information a really high accuracy was obtained.

The performance of RF and SVM is similar, although RF offers slightly higher rates of omission and commission errors. However, MLP slightly outperforms both in terms of accuracy, kappa and omission and commission errors. The best accuracy was obtained with CNN, but the problem of misclassification of impermeability needs to be solved.

Having good accuracy metrics is not enough to consider a classification method as the best option. Prediction maps provide additional insight into possible biases of the classification performed.

Acknowledgements We thank the referees and editors for their constructive feedback regarding the initial version of the manuscript.

Author Contributions Conceptualization: Francisco Alonso-Sarría and Carmen Valdivieso-Ros; Methodology: Francisco Alonso-Sarría and Carmen Valdivieso-Ros; Formal analysis and investigation: Francisco Alonso-Sarría, Carmen Valdivieso-Ros and Francisco Gomariz-Castillo; Writing - original draft preparation: Francisco Alonso-Sarría, Carmen Valdivieso-Ros and Francisco Gomariz-Castillo; Writing - review and editing: Francisco Alonso-Sarría, Carmen Valdivieso-Ros and Francisco Gomariz-Castillo; Funding acquisition: Francisco Alonso-Sarría and Francisco Gomariz-Castillo; Resources: Francisco Alonso-Sarría and Francisco Gomariz-Castillo; Software: Francisco Alonso-Sarría and Francisco Gomariz-Castillo; Data curation: Carmen Valdivieso-Ros; Supervision: Francisco Alonso-Sarría. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was supported by the Spanish Agencia Estatal de Investigación (Grant number TED2021-131131B-I00). Carmen Valdivieso-Ros is grateful for the financing of the pre-doctoral research by the Ministerio de Ciencia, Innovación y Universidades from the Government of Spain (FPU18/01447).

Availability of data and materials The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval Not applicable.

Competing interests The authors declare no competing interests.

References

1. Ali, A.M., Abouelghar, M., Belal, A., et al.: Crop yield prediction using multi sensors remote sensing (review article). Egypt. J. Remote Sens. Space Sci. **25**, 711–716 (2022). https://doi.org/10.1016/j.ejrs.2022.04.006, https://linkinghub.elsevier.com/retrieve/pii/S1110982322000527
2. Ali, K., Johnson, B.A.: Land-use and land-cover classification in semi-arid areas from medium-resolution remote-sensing imagery: A deep learning approach. Sensors **22**, 8750 (2022) https://doi.org/10.3390/s22228750, https://www.mdpi.com/1424-8220/22/22/8750

3. Alonso-Sarria, F., Valdivieso-Ros, C., Gomariz-Castillo, F.: Isolation forests to evaluate class separability and the representativeness of training and validation areas in land cover classification. Remote Sensing **11**, 3000 (2019). https://doi.org/10.3390/rs11243000
4. Amoakoh, A.O., Aplin, P., Awuah, K.T., et al.: Testing the contribution of multi-source remote sensing features for random forest classification of the greater amanzule tropical peatland. Sensors **21**,(2021). https://doi.org/10.3390/s21103399
5. Atef, I., Ahmed, W., Abdel-Maguid, R.H.: Modelling of land use land cover changes using machine learning and gis techniques: a case study in el-fayoum governorate, egypt. Environ. Monit. Assess. **195**, 637 (2023). https://doi.org/10.1007/s10661-023-11224-7, https//link.springer.com/10.1007/s10661-023-11224-7
6. Baddeley, A., Rubak, E., Turner, R.: Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London (2015). https://doi.org/10.1201/b19708
7. Berberoglu, S., Curran, P.J., Lloyd, C.D., et al.: Texture classification of mediterranean land cover. Int. J. Appl. Earth Obs. Geoinf. **9**,(2007). https://doi.org/10.1016/j.jag.2006.11.004
8. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
9. Campos, J.C., Sillero, N., Brito, J.C.: Normalized difference water indexes have dissimilar performances in detecting seasonal and permanent water in the sahara-sahel transition zone. J. Hydrol. 464–465 (2012). https://doi.org/10.1016/j.jhydrol.2012.07.042
10. CARM (2021) Estadística agraria regional. Comunidad Autónoma de la Región de Murcia, accessed: 2021-04-15
11. Castelo-Cabay, M., Piedra-Fernandez, J.A., Ayala, R.: Deep learning for land use and land cover classification from the ecuadorian paramo. Int. J. Digit. Earth **15**, 1001–1017 (2022). https://doi.org/10.1080/17538947.2022.2088872, www.tandfonline.com/doi/full/10.1080/17538947.2022.2088872
12. Charrad M, Ghazzali N, Boiteau V, et al: Nbclust: An r package for determining the relevant number of clusters in a data set. J Stat. Softw. **61** (2014). https://doi.org/10.18637/jss.v061.i06
13. Chen, L., Ren, C., Bao, G., et al.: Improved object-based estimation of forest aboveground biomass by integrating lidar data from gedi and icesat-2 with multi-sensor images in a heterogeneous mountainous region. Remote Sens **14**, 2743 (2022). https://doi.org/10.3390/rs14122743, www.mdpi.com/2072-4292/14/12/2743
14. Chen, X.L., Zhao, H.M., Li, P.X., et al.: Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. Remote Sens. Environ. **104**,(2006). https://doi.org/10.1016/j.rse.2005.11.016
15. Cánovas-García, F., Alonso-Sarría, F.: Optimal combination of classification algorithms and feature ranking methods for object-based classification of submeter resolution z/i-imaging dmc imagery. Remote Sens **7**, 4651–4677 (2015). https://doi.org/10.3390/rs70404651
16. Council NR: Research Strategies for the U.S. Global Change Research Program. The National Academies Press, (1990). https://doi.org/10.17226/1743, https://www.nap.edu/catalog/1743/research-strategies-for-the-us-global-change-research-program
17. Davranche, A., Lefebvre, G., Poulin, B.: Wetland monitoring using classification trees and spot-5 seasonal time series. Remote Sens. Environ. **114**, 552–562 (2010). https://doi.org/10.1016/j.rse.2009.10.009
18. Dobrinić, D., Medak, D., Gašparović, M.: Integration of multitemporal sentinel-1 and sentinel-2 imagery for land-cover classification using machine learning methods. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B1-2020:91–98. (2020).https://doi.org/10.5194/isprs-archives-XLIII-B1-2020-91-2020, https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B1-2020/91/2020/
19. Dong, P., Chen, Q.: LiDAR Remote Sensing and Applications. CRC Press (2017). https://doi.org/10.4324/9781351233354
20. Ghayour, L., Neshat, A., Paryani, S., et al.: Performance evaluation of sentinel-2 and landsat 8 oli data for land cover/use classification using a comparison between machine learning algorithms. Remote Sens **13**, 1349 (2021). https://doi.org/10.3390/rs13071349, https://www.mdpi.com/2072-4292/13/7/1349
21. Giménez-Casalduero, F., Gomariz-Castillo, F., Alonso-Sarría, F., et al.: Pinna nobilis in the mar menor coastal lagoon: a story of colonization and uncertainty. Mar. Ecol. Prog. Ser. **652**, 77–94 (2020). https://doi.org/10.3354/meps13468
22. Gomariz-Castillo, F., Alonso-Sarría, F., Cánovas-García, F.: Improving classification accuracy of multi-temporal landsat images by assessing the use of different algorithms, textural and ancillary information for a mediterranean semiarid area from 2000 to 2015. Remote Sens **9**, 1058 (2017). https://doi.org/10.3390/rs9101058
23. Hall-Beyer, M.: Practical guidelines for choosing glcm textures to use in landscape classification tasks over a range of moderate spatial scales. Int. J. Remote Sens. **38**, 1312–1338 (2017). https://doi.org/10.1080/01431161.2016.1278314
24. Haralick, R.: Statistical and structural approaches to texture. Proc. IEEE **67**, 786–804 (1979). https://doi.org/10.1109/PROC.1979.11328
25. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern **3**, 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314
26. Hong, C., Jin, X., Ren, J., et al.: Satellite data indicates multidimensional variation of agricultural production in land consolidation area. Sci. Total Environ. **653**,(2019). https://doi.org/10.1016/j.scitotenv.2018.10.415
27. Hopkins, B., Skellam, J.G.: A new method for determining the type of distribution of plant individuals. Ann. Bot. **18**, 213–227 (1954). https://doi.org/10.1093/oxfordjournals.aob.a083391
28. Hu, Y., Zhang, Q., Zhang, Y., et al.: A deep convolution neural network method for land cover mapping: A case study of qinhuangdao, china. Remote Sensing **10**, 2053 (2018). https://doi.org/10.3390/rs10122053
29. Huete, A.: A soil-adjusted vegetation index (savi). Remote Sens. Environ. **25**, 295–309 (1988). https://doi.org/10.1016/0034-4257(88)90106-X
30. IGN: Plan Nacional de Ortofotgrafía Aérea. (2023) . https://pnoa.ign.es
31. James, G., Witten, D., Hastie, T., et al: An Introduction to Statistical Learning, vol 103. Springer New York, (2013). https://doi.org/10.1007/978-1-4614-7138-7, http://link.springer.com/10.1007/978-1-4614-7138-7
32. Kauth, R.J., Thomas, G.S.: The Tasselled Cap – a graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT. In: for Applications of Remote Sensing TL (ed) LARS Symposia, vol 159. pp. 4B–41–4B–51. Purdue University, West Lafayette, Indiana, (1976)
33. Klein, I., Gessner, U., Dietz, A.J., et al.: Global waterpack - a 250 m resolution dataset revealing the daily dynamics of global inland water bodies. Remote Sens. Environ. **198**, 345–362 (2017). https://doi.org/10.1016/j.rse.2017.06.045
34. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News 2. (2002)
35. Liaw, A., Yan, J., Li, W., et al: Package 'randomforest'. R news XXXIX. (2014)
36. Liu, S., Qi, Z., Li, X., et al.: Integration of convolutional neural networks and object-based post-classification refinement for land use and land cover mapping with optical and sar data. Remote Sensing **11**, 690 (2019). https://doi.org/10.3390/rs11060690
37. Liu, T., Abd-Elrahman, A., Morton, J., et al.: Comparing fully convolutional networks, random forest, support vector machine, and

patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system. GISci Remote Sens **55**, 243–264 (2018). https://doi.org/10.1080/15481603.2018.1426091
38. Lloyd C (2010) Spatial Data Analysis: An Introduction for GIS Users. Oxford University Press
39. Luca, G.D., Silva, J.M.N., Fazio, S.D., et al.: Integrated use of sentinel-1 and sentinel-2 data and open-source machine learning algorithms for land cover mapping in a mediterranean region. Eur. J Remote Sens. **55**, 52–70 (2022). https://doi.org/10.1080/22797254.2021.2018667, https://www.tandfonline.com/doi/full/10.1080/22797254.2021.2018667
40. Martínez, J., Esteve, M., Martínez-Paz, J., et al: Simulating management options and scenarios to control nutrient load to mar menor, southeast spain. Transitional Waters Monographs TWM, Transit Waters Monogr 1. (2007). https://doi.org/10.1285/i18252273v1n1p53
41. Masiza, W., Chirima, J.G., Hamandawana, H., et al.: Enhanced mapping of a smallholder crop farming landscape through image fusion and model stacking. Int. J. Remote Sens. **41**, 8739–8756 (2020). https://doi.org/10.1080/01431161.2020.1783017
42. Mason, P.J., Manton, M., Harrison, D.E., et al : The second report on the adequacy of the global observing systems for climate in support of the unfccc. GCOS Rep 82. (2003). https://library.wmo.int/doc_num.php?explnum_id=3931
43. Mostafiz, C., Chang, N.B.: Tasseled cap transformation for assessing hurricane landfall impact on a coastal watershed. Int. J. Appl. Earth Obs. Geoinf. **73**, 736–745 (2018). https://doi.org/10.1016/j.jag.2018.08.015
44. Periasamy, S.: Significance of dual polarimetric synthetic aperture radar in biomass retrieval: An attempt on sentinel-1. Remote Sens. Environ. **217**, 537–549 (2018). https://doi.org/10.1016/j.rse.2018.09.003
45. Qin, R., Liu, T.: A review of landcover classification with very-high resolution remotely sensed optical images-analysis unit, model scalability and transferability. Remote Sensing **14**, 646 (2022). https://doi.org/10.3390/rs14030646
46. Reichstein, M., Camps-Valls, G., Stevens, B., et al.: Deep learning and process understanding for data-driven earth system science. Nature **566**, 195–204 (2019). https://doi.org/10.1038/s41586-019-0912-1
47. Rossiter, D.: Statistical methods for accuracy assesment of classified thematic maps. Tech. rep., Department of Earth Systems Analysis International Institute for Geo-information Science & Earth Observation (ITC), Enschede (NL). (2004)
48. Rouse, J.W., Haas, R.H., Schell, J.A., et al: Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. Progress Report RSC 1978-1 (1973) https://ntrs.nasa.gov/citations/19740022555
49. Singh, P.G., Bordu, N., Singh, D., et al.: Permuted spectral and permuted spectral-spatial cnn models for polsar-multispectral data based land cover classification. Int. J. Remote Sens. **42**, 1096–1120 (2021). https://doi.org/10.1080/01431161.2020.1823041, https://www.tandfonline.com/doi/full/10.1080/01431161.2020.1823041
50. Tsai, M.D., Tseng, K.W., Lai, C.C., et al.: Exploring airborne lidar and aerial photographs using machine learning for land cover classification. Remote Sensing **15**, 2280 (2023). https://doi.org/10.3390/rs15092280, https://www.mdpi.com/2072-4292/15/9/2280
51. Valdivieso-Ros, C., Alonso-Sarria, F., Gomariz-Castillo, F.: Effect of different atmospheric correction algorithms on sentinel-2 imagery classification accuracy in a semiarid mediterranean area. Remote Sensing **13**, 1770 (2021). https://doi.org/10.3390/rs13091770, https://www.mdpi.com/2072-4292/13/9/1770
52. Valdivieso-Ros, C., Alonso-Sarria, F., Gomariz-Castillo, F.: Effect of the synergetic use of sentinel-1, sentinel-2, lidar and derived data in land cover classification of a semiarid mediterranean area using machine learning algorithms. Remote Sensing **15**, 312 (2023). https://doi.org/10.3390/rs15020312, https://www.mdpi.com/2072-4292/15/2/312
53. Vanhellemont, Q.: Adaptation of the dark spectrum fitting atmospheric correction for aquatic applications of the landsat and sentinel-2 archives. Remote Sens. Environ. **225**, 175–192 (2019). https://doi.org/10.1016/j.rse.2019.03.010
54. Vanhellemont, Q., Ruddick, K. Acolite for sentinel-2: Aquatic applications of msi imagery. Paper presented at the Living Planet Symposium, Proceedings of the conference held 9-13 May 2016. (2016)
55. Vanhellemont, Q., Ruddick, K.: Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. Remote Sens. Environ. **216**, 586–597 (2018). https://doi.org/10.1016/j.rse.2018.07.015
56. Xu, H.: Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. **27**, 3025–3033 (2006). https://doi.org/10.1080/01431160600589179
57. Yang, X., Qin, Q., Grussenmeyer, P., et al.: Urban surface water body detection with suppressed built-up noise based on water indices from sentinel-2 msi imagery. Remote Sens. Environ. **219**, 259–270 (2018). https://doi.org/10.1016/j.rse.2018.09.016
58. YiLan L, RuTong Z (2022) clustertend: Check the Clustering Tendency. https://CRAN.R-project.org/package=clustertend, r package version 1.4
59. Yuh, Y.G., Tracz, W., Matthews, H.D., et al.: Application of machine learning approaches for land cover monitoring in northern cameroon. Eco. Inform. **74**, 101955 (2023). https://doi.org/10.1016/j.ecoinf.2022.101955, https://linkinghub.elsevier.com/retrieve/pii/S1574954122004058
60. Zhang, H., Xu, R.: Exploring the optimal integration levels between sar and optical data for better urban land cover mapping in the pearl river delta. Int. J. Appl. Earth Obs. Geoinf. **64**, 87–95 (2018). https://doi.org/10.1016/j.jag.2017.08.013