remote sensing

MDPI

Article

Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia

Iban Berganzo-Besga [1], Hector A. Orengo [1,*], Felipe Lumbreras [2], Miguel Carrero-Pazos [3], João Fonte [4] and Benito Vilas-Estévez [5]

[1] Landscape Archaeology Research Group, Catalan Institute of Classical Archaeology, Pl. Rovellat s/n, 43003 Tarragona, Spain; iberganzo@icac.cat
[2] Computer Vision Center, Computer Science Deptartment, Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellaterra, Spain; felipe@cvc.uab.es
[3] Institute of Archaeology, University College London, 31–34 Gordon Square, London WC1H 0PY, UK; miguel.pazos.15@ucl.ac.uk
[4] Department of Archaeology, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK; j.fonte3@exeter.ac.uk
[5] Grupo de Estudos de Arqueoloxía, Antigüidade e Territorio, Facultade de Historia, University of Vigo, As Lagoas, s/n, 32004 Ourense, Spain; benito.vilas@uvigo.es
* Correspondence: horengo@icac.cat

Citation: Berganzo-Besga, I.; Orengo, H.A.; Lumbreras, F.; Carerro-Pazos, M.; Fonte, J.; Vilas-Estévez, B. Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia. *Remote Sens.* **2021**, *13*, 4181. https://doi.org/10.3390/rs13204181

Academic Editor: Timo Balz

Received: 21 September 2021
Accepted: 16 October 2021
Published: 19 October 2021

Abstract: This paper presents an algorithm for large-scale automatic detection of burial mounds, one of the most common types of archaeological sites globally, using LiDAR and multispectral satellite data. Although previous attempts were able to detect a good proportion of the known mounds in a given area, they still presented high numbers of false positives and low precision values. Our proposed approach combines random forest for soil classification using multitemporal multispectral Sentinel-2 data and a deep learning model using YOLOv3 on LiDAR data previously pre-processed using a multi–scale relief model. The resulting algorithm significantly improves previous attempts with a detection rate of 89.5%, an average precision of 66.75%, a recall value of 0.64 and a precision of 0.97, which allowed, with a small set of training data, the detection of 10,527 burial mounds over an area of near 30,000 km^2, the largest in which such an approach has ever been applied. The open code and platforms employed to develop the algorithm allow this method to be applied anywhere LiDAR data or high-resolution digital terrain models are available.

Keywords: tumuli; mounds; archaeology; deep learning; machine learning; Sentinel-2; Google Colaboratory; Google Earth Engine

1. Introduction

During the last 5 years, the use of artificial intelligence (AI) for the detection of archaeological sites and features has increased exponentially [1]. There has been considerable diversity of approaches, which respond to the specific object of study and the sources available for its detection. Classical machine learning (ML) approaches such as random forest (RF) to classify multispectral satellite sources have been used for the detection of mounds in Mesopotamia [2], Pakistan [3] and Jordan [4], but also for the detection of material culture in drone imagery [5]. Deep learning (DL) algorithms, however, have been increasingly popular during the last few years, and they now comprise the bulk of archaeological applications to archaeological site detection. Although DL approaches are also diverse and include the extraction of site locations from historical maps [6] and automated archaeological survey [7], a high proportion of their application has been directed towards

the detection of archaeological mounds and other topographic features in LiDAR datasets (e.g., [1,8–11]).

This is probably due to the common presence of tumular structures of archaeological nature across the globe but also to the simplicity of mound structures. Their characteristic tumular shape has been the primary feature for their identification on the field. They can therefore be easily identified in LiDAR-based topographic reconstructions presented at sufficient resolution. The simple shape of mounds or tumuli is ideal for their detection using DL approaches. DL-based methods usually require large quantities of training data (in the order of thousands of examples) to be able to produce significant results. However, the homogenously semi-hemispherical shape of tumuli, allows the training of usable detectors with a much lower quantity of training data, reducing considerably the effort required to obtain it and the significant computational resources necessary to train a convolutional neural network (CNN) detector. This type of features, however, present an important drawback. Their common, simple, and regular shape is similar to many other non-archaeological features and therefore studies implementing methods for mound detection in LiDAR-derived and other high-resolution datasets are characterised by a very large presence of false positives (FPs) [8,12].

Given the importance of tumuli in the archaeological literature and in that dealing with the implementation of automated detection methods in archaeology, this paper builds up from current approaches, but incorporates a series of innovations, which can be summarised as follows:

1. The use of RF ML classifier to classify Sentinel-2 data into a binary raster depicting areas where archaeological tumuli may be present or not;
2. DL approach using a relatively unexploited DL algorithm in archaeology, YOLOv3, which provides particularly efficient outputs. To boost the efficiency of the shape-detection method a series of innovations were implemented:
    - Pre-treatment of the LiDAR dataset with a multi-scale relief model (MSRM) [13], which, contrary to other methods, is usually employed to improve the visibility of features in LiDAR-based digital terrain models (DTMs), considers the multi-scale nature of mounds;
    - The development of data augmentation (DA) methods to increase the effectivity of the detector. One of them, the training of the CNN from scratch applying own pre-trained models created from simulated data;
    - The use of publicly accessible computing environments, such as Google Earth Engine (GEE) and Colaboratory, which provide the necessary computational resources and assure the method's accessibility, reproducibility and reusability.

We tested this approach in the entire region of Galicia, located in the Northwest of the Iberian Peninsula. Galicia is an ideal testing area due to the following reasons: (1) its size, which allowed us to test the method under a diversity of scenarios at a very large scale (29,574 km^2, 5.8% of Spain), to our knowledge the largest area to which a CNN-based detector of archaeological features has ever been applied; (2) the presence of a very well-known Atlantic burial tradition characterised by the use of mound tombs; and (3) the availability of high-quality training and test data necessary for the successful development of the detector.

Previous research on this area has highlighted a very dense concentration of megalithic sites, mainly comprised by unexcavated mounds covered by vegetation. They present an average size of 15–20 m in diameter, and 1–1.5 m high. In some cases, the mound covers a burial chamber made of granite constituting a dolmen or passage grave [14,15]. The regional government (in Galician *Xunta de Galicia*) has been developing survey works since the 1980s, resulting in an official sites and monuments record. This official catalogue currently has more than 7,000 records for megalithic mounds, although problems regarding its reliability have recently been pointed out [16]. Another issue relates to the archae-

ological detection of those sites during fieldwork. The dense vegetation and forests covering a high percentage of the Galician territory and their subtle topographic nature, which makes many of them virtually invisible to the casual observer, complicates the detection of these structures even for specialised archaeologists. These problems have been identified in other Iberian and European areas [17,18]. The use of automatic detection methods can hugely help to validate and increase heritage catalogues' records, protect those cultural resources, and boost research on the large-scale distribution of the cultural processes that created them. Although visual approaches using LiDAR data have been employed for the detection and analysis of barrows in Galicia [15,19], no automatic detection of megalithic burial mounds has ever been attempted before in the area.

2. Materials and Methods

Most recent research on archaeological feature detection using LiDAR datasets has used algorithms based on region-based CNN (R-CNN). R-CNN is an object detection algorithm based on a combination of classical tools from Computer Vision (CV) and DL that has achieved significant improvements, of more than 30% in some cases, in detection metrics using reference datasets within the CV community [20].

However, the use of single-channel (or single band images) CNN-based approaches for the detection of archaeological tumuli in LiDAR-derived digital surface models (DSMs) has frequently encountered strong limitations, as they cannot readily differentiate between archaeological tumuli and other features of tumular shape, such as roundabouts or rock outcrops. Initial tests solely using an R-CNN-based detection method and a filtered DTM detected hundreds of FPs corresponding to roundabouts, rock outcrops (in mountain and the coastal areas), house roofs, swimming pools but also multiple mounds in quarries, golf courses, shoot ranges, and industrial sites between others. As these presented a tumular shape, they could not have been filtered out to improve the training data without losing a large quantity of archaeological tumuli. This is a common problem in CNN-based mound detection (see, for example, [8]).

To overcome this problem, a workflow combining different data types and ML approaches has been newly developed for this study:

2.1. Digital Terrain Model Pre-Processing

Pre-processing of the DTM is a common practice in DL-based detection. The use of micro-relief visualisation techniques in particular highlights archaeological features that are almost or completely invisible in DTMs [21].

The DTM employed to conduct DL-based shape detection was obtained from the Galician Regional Government Geographical Portal (*Información Xeográfica de Galicia*) [22]. The LiDAR-based DTM (*MDT_1m_h50*) was considered adequate due to its good quality (even in forest-filtered areas), its resolution of 1 m/px and its public availability. The DTM allowed a good visualisation of all mounds used for training data (Figure 1).

In a first approximation to mound detection using DL, we used the DTM data for algorithm training, but, as expected, an average precision (AP) of 21.81% indicated that a pre-processing stage was required on the input data. Three common relief visualization techniques were tested to improve the input data and thus facilitate the detection of burial mounds (Figure 1): 1. MSRM (*fmn* = 1, *fmx* = 19, *x* = 2) [13]; 2. slope gradient [23,24]; and 3. simple local relief model (SLRM) (*radius* = 20), which is a simplified local relief model [25]. These constitute the most used LiDAR pre-processing methods for the detection of small-scale features and those in which the known burial mounds were best observed with the naked eye. The Relief Visualization Toolbox was used to obtain the slope and SLRM raster files [26,27] and GEE Code Editor, Repository and Cloud Computing Platform [28] for the MSRM. The best results were obtained using MSRM (see the results section for details), and therefore it was the one employed for the pre-treatment of the DTM in this study.

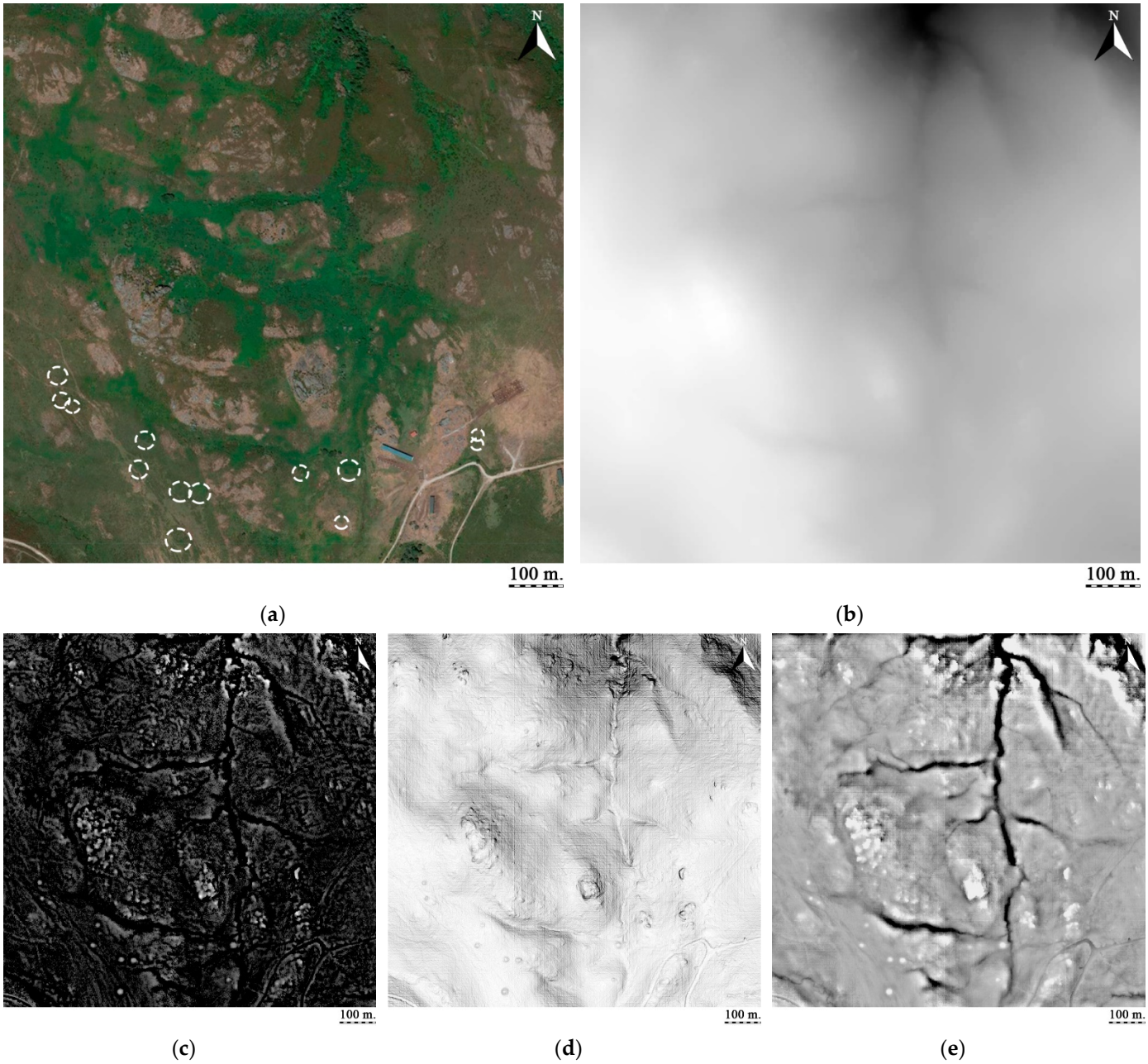


Figure 1. A comparison of a part of the used training data to evaluate which of the visualisation algorithms were better suited for the detection of burial mounds: (**a**) satellite view of that area with the known tumuli marked; (**b**) DTM; (**c**) MSRM; (**d**) slope gradient; (**e**) SLRM.

2.2. Deep Learning Shape Detection

For the DTM-based shape detection we used YOLO [29], an R-CNN-based algorithm previously employed in the field of archaeology for the detection of inscriptions in oracle bones [30]. The YOLOv3 algorithm is faster than other R-CNN methods like Faster R-CNN. Its backbone, Darknet-53, is 1.5 times faster than ResNet-101, working at 78 frames per second [29,31]. YOLOv3 predicts at three different scales, which is similar to what the feature pyramid network does [29,32]. This structure allows the detection of small objects. The bounding boxes are predicted by the anchor boxes generated using k-means clustering with an Intersection over Union (IoU) threshold of 0.5. The class prediction is made using binary cross-entropy loss and independent logistic classifiers, the latter to facilitate multilabel classification [29].

An Nvidia Titan XP graphics processing unit (GPU) with 12 GB of RAM hosted at the Computer Vision Center (CVC) of the Autonomous University of Barcelona (UAB) was used to run the DL algorithms. The chosen work environment was the parallel computing platform CUDA 11.2, the ML library Tensorflow 2.1.0, the DL library cuDNN 8.1.1, the software development tool CMake 3.20.2 and the CV library OpenCV 4.5.2 as recommended for YOLOv3 [33].

As training and validation data, we used the current known burial mounds data obtained from the studies led by M. Carrero-Pazos and B. Vilas [16,34] in Galicia and J. Fonte in the area of Northern Portugal (Figure 1). The database comprised a total of 306 tumuli. From these, 200 were employed for training and 106 for validation. For the same map scale, the training, validation, and detection images size has been 1024 × 1024 pixels. The training and detection data size should not differ more than 40% to avoid FPs [33]. Since there is less representation of smaller diameter mounds in the training data, a DA process has been applied to add resized burial mound as new training data. DA seeks to generate more training data from our available data through a series of random transformations to the image [35]. As can be seen from Figure 2, after scaling the images to 75% and 50% of their size (DA1), practically the equivalent of the training data for small and large mounds was achieved, taking an average diameter of 18 m [15]. DA1 added 400 more mounds for training. Likewise, to label the training and validation images to create our custom data, we used LabelImg, a simple graphical image annotation tool [36] that allowed us to tag images directly in YOLO format. In this step, images without burial mounds were not included.

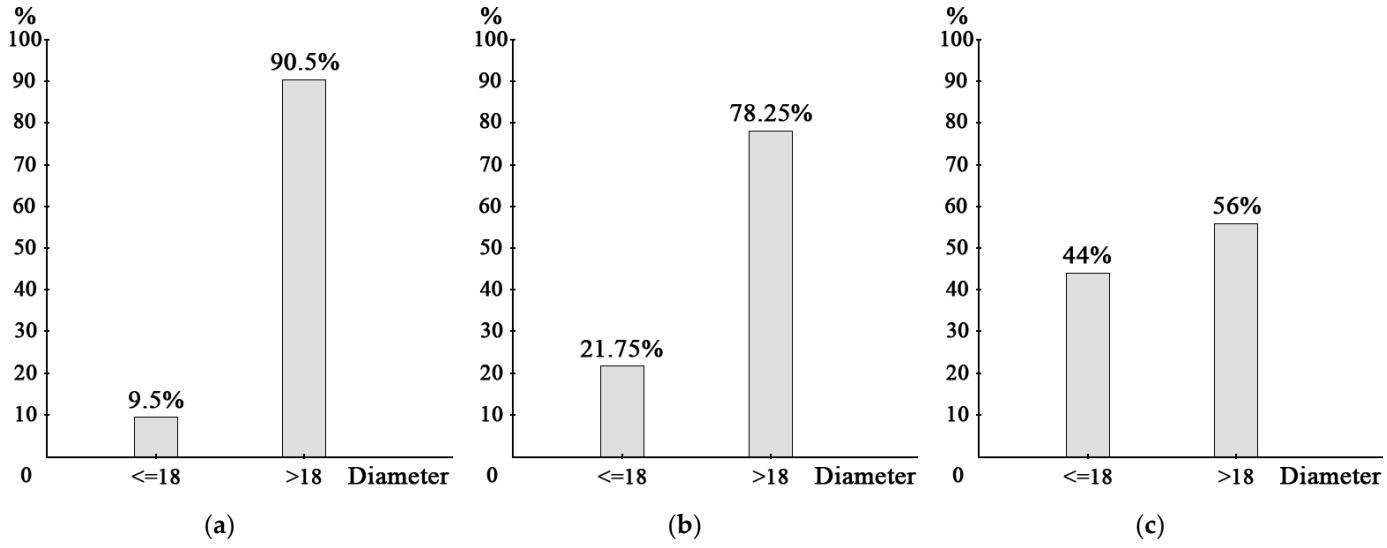


Figure 2. Number of training burial mounds with diameters of less than or equal to 18 m: (**a**) without DA; (**b**) adding the scaling to 75%; (**c**) adding the scaling to 75% and 50% (DA1).

The initially trained algorithm produced multiple false negatives (FNs) and FPs throughout Galicia (Figure 3). Additionally, some FNs were quite differently shaped from the training mounds. This issue led us to consider introducing model refinement procedures.

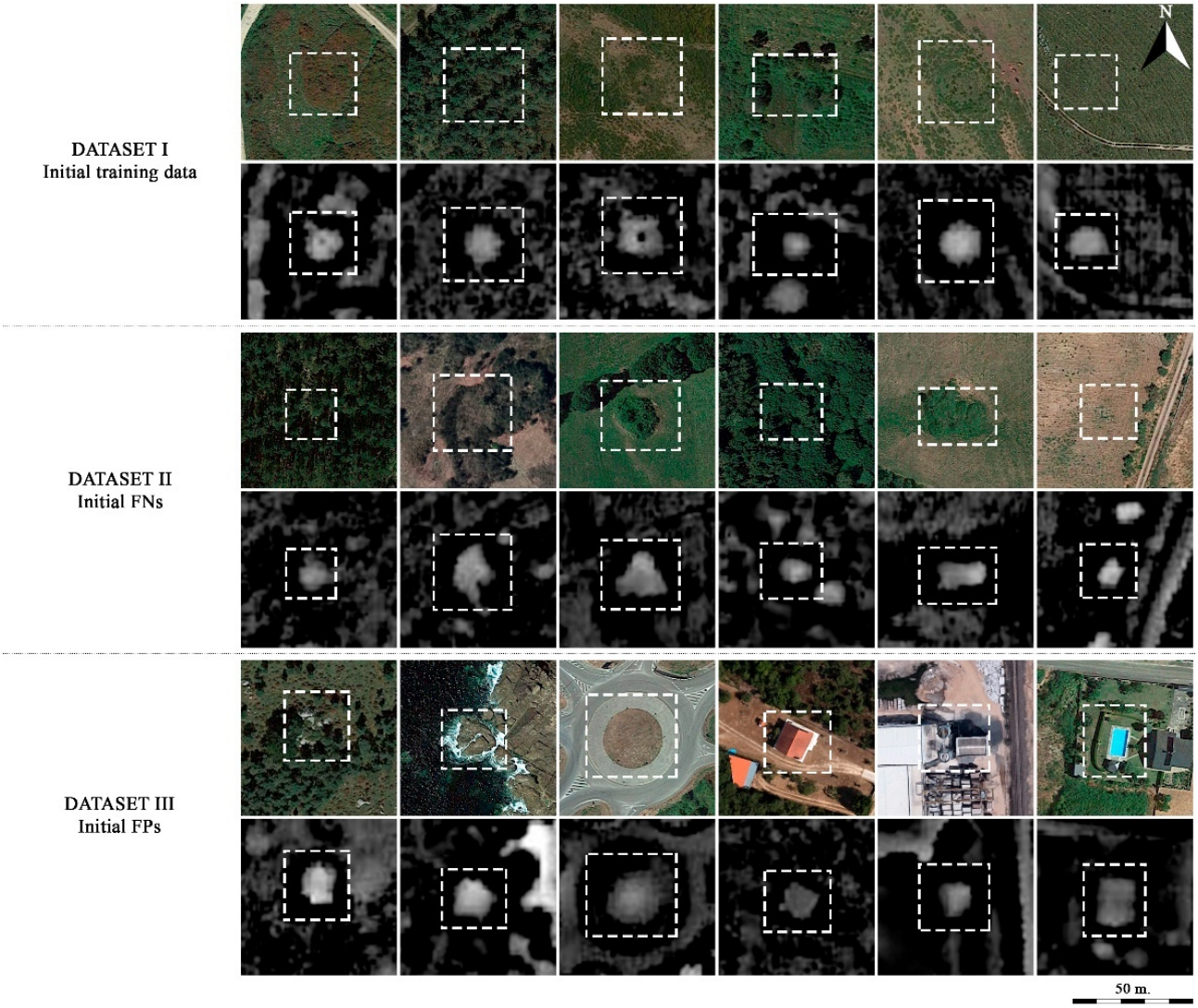


Figure 3. MSRM training data examples (Dataset I) used for the algorithm, and its FN (Dataset II) and FP examples (Dataset III) from our initial detection in Galicia. The corresponding top image for each pair is a visible satellite image, shown for the sake of visualisation, but not used in our process.

2.3. Model Refinement

In our initial model, the use of a CNN-based detection method and a filtered DSM detected hundreds of FPs corresponding to roundabouts, rock outcrops (also on the coast), house roofs, and swimming pools, but also multiple mounds in quarries, golf courses, shoot ranges and industrial sites between others. As many of these (particularly roundabouts) presented a tumular shape they could not have been filtered out to improve the training data without losing a large quantity of archaeological tumuli. This is a frequent problem in CNN-based mound detection. Through model refinement, we sought to reduce both FNs and FPs. In this retraining, 278 missing new burial mounds (FNs) and 88 FPs were collected from the previous training steps as new training data. From this step onwards, images without burial mounds were included as training, obtained from the aforementioned FPs. In addition to training with false positives, a second training was proposed adding the FPs as a new extra class to find out if the algorithm was able to filter them more efficiently. To increase the burial mounds correctly detected, we proceeded to apply DA, beyond that initially developed. For this, two new models were tested. The first was a random rotation at different angles of the training features (DA2) and the second was the use of a pre-trained initial weight created specifically for the detection of circular

shapes (DA3). However, these did not produce significant improvements, and were not incorporated in the final model (see discussion for details).

Despite model refinement, several FPs remained. To remove them, a filtering step using official urban, industrial and road layers was proposed. In a previous attempt to tackle this issue, Verschoof-van der Vaart et al. (2020) developed a three-level location-based ranking using the information provided by soil-type and land-use maps [8]. Instead of a ranking, such as that proposed by Verschoof-van der Vaart et al. (2020), we simply selected and eliminated the mounds detected in these areas after checking that all of them corresponded to FPs. Even though this approach eliminated most of the detected FPs in these areas, our results still included many FPs as land-use maps for the area do not classify as urban several areas in which isolated houses, swimming pools or roundabouts are present. Also, soil type maps included within the same category areas with potential archaeological mounds and FPs. For example, many archaeological mounds were located within granitic grasslands but at the same time, the specific nature and shape of granitic outcrops within these grasslands created many FPs that could not be filtered using this approach. In addition, some correct burial mounds close to the removed areas were also eliminated.

2.4. Random Forest Classification of Multitemporal Sentinel-2 Data

To overcome this problem, we decided to develop a binary soil classification map using GEE Code Editor, Repository and Cloud Computing Platform [28]. Our objective was to eliminate those pixels that could not correspond to archaeological mounds. To reach this objective we used cloud-filtered multitemporal Sentinel-2 multispectral imagery. Sentinel-2 incorporates 13 bands from which only the visible/near-infrared bands (VNIR B2–B8A) and the short-wave infrared bands (SWIR B11–B12) were employed. Bands B1, B9, and B10 (60 m/px each) correspond to aerosols, water vapor, and cirrus, respectively, and they were not employed in this study except for the use of the cirrus-derived cloud mask applied. Visible (B2–B4) and NIR (B8) bands provide a ground resolution of 10 m/px, while red-edge (B5-B7 and B8A) and SWIR (B11–B12) bands present a 20 m/px spatial resolution. Specifically, for this research Sentinel-2 Level 1C products representing top of atmosphere (TOA) reflectance were preferred due to the larger span of its mission (starting from June 2015). Sentinel-2 multispectral satellite images were a good compromise given their relatively high spatial and spectral resolutions and their open access policy. The use of cloud-filtered multitemporal satellite data has been successfully employed in previous research to provide long-term vegetation indices [37,38], but also for the development of machine learning classifications [3,5] as they provide images that are independent of specific environmental or land-use conditions that are particularly adequate for the development of classifications.

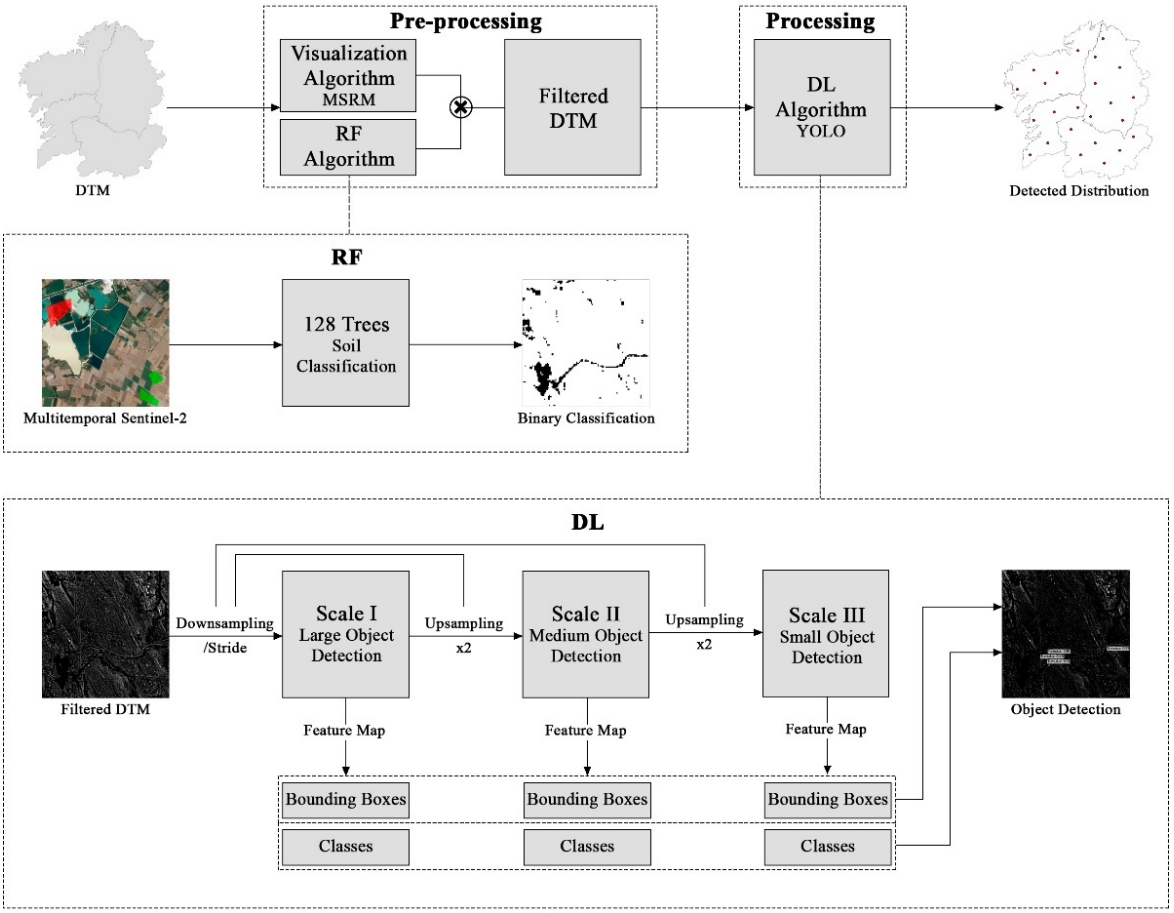
The use of GEE allowed us to access and join 1920 (at the moment of writing) Sentinel-2 images in a single 10-band composite, train the classification algorithm and execute the analysis, which would have been impossible using a desktop computer. It also provided an ideal environment to join the results of the classification with that resulting of the MSRM filter of the DTM also created using GEE (see previous section). Thirteen polygons defining training areas were drawn and tagged as class 0 (areas unsuited for the presence of tumuli), which included a variety of urban features (such as roofs, roads, swimming pools, etc.), water, rock and quarries and other industrial areas. Additionally, 19 class 1 polygons were drawn within grasslands, cultivation fields and forests. From these polygonal training areas, a total of 4398 sampling points corresponding to individual multispectral pixels (1832 for class 0 and 2566 for class 1) were extracted with values for all selected bands and a class identifier. These training data were employed to classify the composite raster using a RF algorithm with 128 trees, which resulted in a binary raster indicating areas where archaeological tumuli can (class 1) and cannot (class 0) be found.

2.5. Hybrid Machine Learning Approach

The combination of DL for shape detection and traditional ML for binary soil classification is described in Scheme 1. The use of GEE for the generation of both MSRM and the binary classification map made it possible to integrate both processes in a single script, which, as a last step, multiplied both outputs to produce a MSRM in which all areas not conductive to the presence of mounds had been removed.

A similar approach combining DL and traditional ML was recently published by Davis et al. (2021) [1]. While we used the RF classification to eliminate areas of source of FPs for the application of the DL detector, they used the multisource multitemporal RF approach developed by Orengo et al. (2020) [3] to evaluate the detection results from a Mask R-CNN detector. Although this approach was useful to confirm many of the detected features, it was not integrated into the detection workflow and did not contribute to reduce the large number of FPs reported.

In our case, the DL algorithm was retrained using the new raster produced by the multiplication of the MSRM and the classified binary raster. The RF removed 11 real archaeological tumuli from our initial training data and 13 from the refinement step, leaving 560 burial mounds to work with. Of those 560 mounds, 456 were employed for training and 104 for validation.



Scheme 1. The implemented workflow for object detection with the detail of the structure and behaviour of the RF and DL algorithms.

3. Results

3.1. Digital Terrain Model Pre-Processing

MSRM was the most effective DTM pre-processing method for the detection of barrows (Table 1), with an AP of 63.03% and higher recall and precision values. Despite showing a better result, the initial detection using MSRM presents a recall value of 0.58, which highlights the presence of a large proportion of FNs, and a precision of 0.95 indicating that some FPs were detected.

Table 1. Evaluation of the YOLOv3 models using MSRM, Slope gradient and SLRM as input data.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision
MSRM	63.03%	62	3	44	0.58	0.95
SLOPE	53.58%	49	5	57	0.46	0.91
SLRM	52.89%	44	8	62	0.42	0.85

3.2. Model Refinement and Data Augmentation

As said before, two different models were tested applying model refinement: a two-classes model with the FPs as the new class and one class model with the FPs as background. As shown in Table 2, model refinement works similarly in both cases because the background of the images is considered in the training. Although the recall and precision values have not improved significantly compared to the previous case, the key is that this result now includes the mentioned FPs and the FNs. Even though the number of FPs was reduced, several are nonetheless included.

Table 2. Evaluation of the YOLOv3 models using model refinement for one class and two classes.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision
1 class	66.77%	63	3	43	0.59	0.95
2 classes	70.30%	66	3	40	0.62	0.96

The use of DA methods provided mixed results. Although all DA methods improved the results provided by the training without DA, the resizing of the training data (DA1) proved the most effective (Table 3). Even if it increased the presence of FPs it also increased the number of true positives (TPs) while reducing the presence of FNs. Therefore, DA1 was implemented in the final model.

Table 3. Results of the YOLOv3 models using different types of DA.

DA	AP@0.5	TPs	FPs	FNs	Recall	Precision
None	68.31%	63	2	43	0.59	0.97
DA1	70.30%	66	3	40	0.62	0.96
DA1 + DA2	67.62%	65	2	41	0.61	0.97
DA1 + DA3	66.77%	66	6	40	0.62	0.92

3.3. Integration of Random Forest Classification

The use of the RF classification of satellite data aimed at reducing the number of FPs, by eliminating those areas with soils not conducive to the presence of burial mounds. The results of the validation (Table 4) show that the RF classification and filtering of the DTM improved the model in all respects. It increased the number of TPs while reducing the presence of FPs and FNs. The model trained with the classification-filtered MSRM was also able to detect 1538 tumuli more than that without the filter with a lower presence of FPs and FNs. Although a percentage of false positives are still present after using the classification to filter the MSRM (see the evaluation section for details) it was successful in

eliminating all urban areas and road related infrastructure (all roundabouts were also eliminated), even those not considered as such in the official land-use maps.

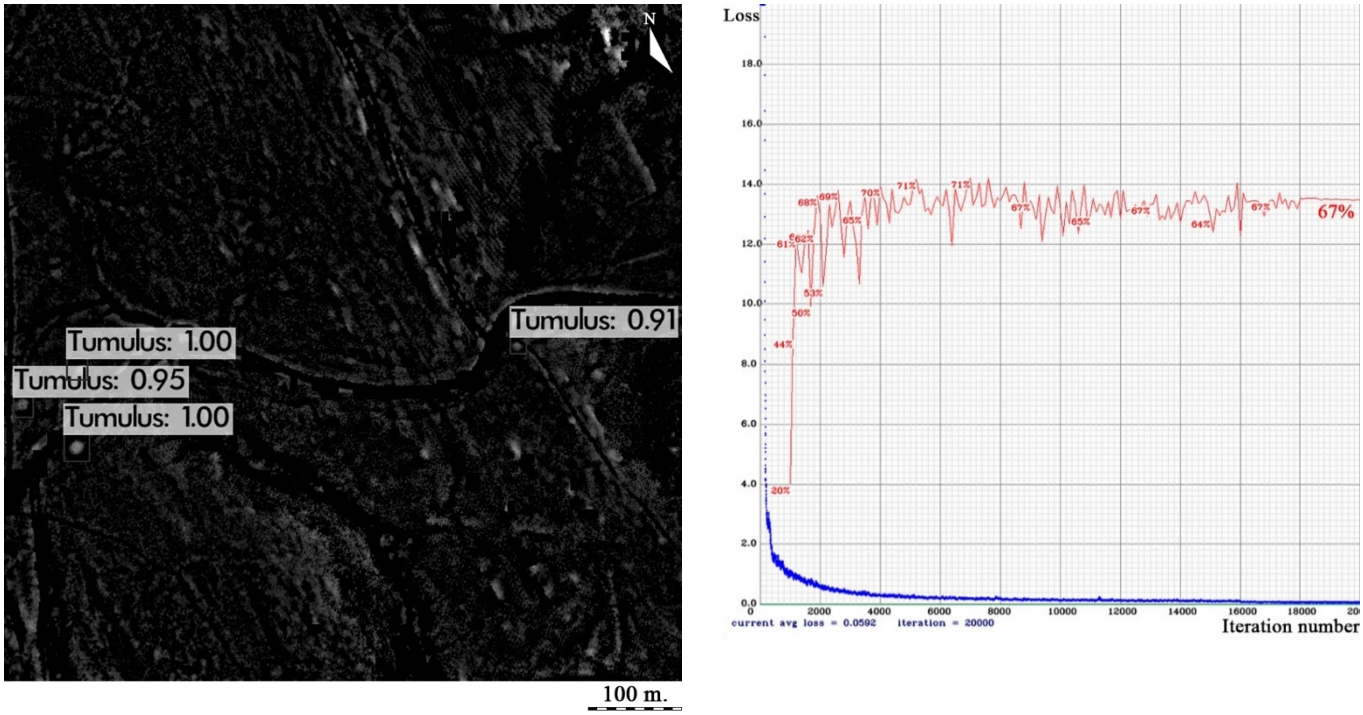
Table 4. Evaluation of the YOLOv3 models using RF filtering and not using it.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision	Mounds
Not RF	71.65%	66	3	38	0.63	0.96	8989
RF	66.75%	67	2 [1]	37	0.64	0.97	10,527

[1] Three FPs were eliminated because they were in a wooded area and it was not possible to determine if they were actually TPs or FPs.

3.4. Results and Test Dataset-Based Validation

The YOLOv3 algorithm has validated the known burial mounds with an AP@0.50 of 66.75% and a loss value of 0.0592 (Figure 4). Furthermore, 10,527 burial mounds were detected all over Galicia with a minimum similarity of 25%, a minimum size of 7 m, a maximum size of 74 m, a mean size of 29 m and a mode of 25 m. Likewise, the locations of these detected tumuli were indicated in order to facilitate their identification in the field. The implemented parameters were *classes* = 1, *channels* = 1, *max_batches* = 20000, *width* = 832 px and *height* = 832 px for training configuration, and *width* = 1024 px and *height* = 1024 px for the detection one. DA1 was the DA dataset implemented.

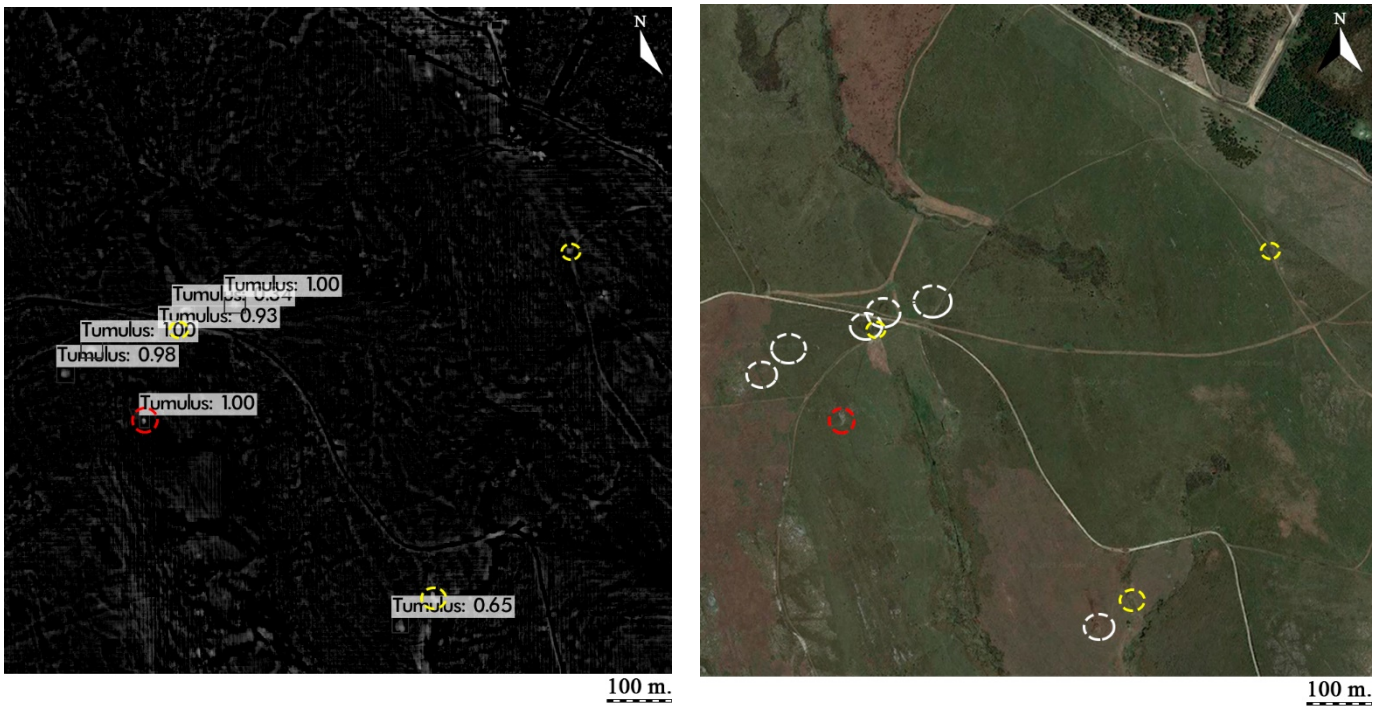


(**a**) (**b**)

Figure 4. YOLOv3 model: (**a**) tumulus detection example; (**b**) loss (blue) and AP@0.50 (red) vs. iteration number function.

This model proved to have similar robustness to the previous one despite having a slightly lower AP (Table 4). As the AP is the used area under the precision/recall curve for each recall value, it is possible that even if the precision and recall values improve, the AP may be lower. However, the AP value, calculated for an IoU threshold of 0.5, were not completely successful. On the one hand, a 0.97 precision value on the test dataset shows that the algorithm distinguishes burial mounds with high precision, but also that there were two FPs, 1.92% of the total (Figure 5). Both of these corresponded to small and isolated rock outcrops. On the other hand, the 0.64 recall value reveals that most of the burial

mounds have been correctly validated by the algorithm, but also that there were 37 FNs, 35.58% of the total (Figure 5).



(**a**) (**b**)

Figure 5. Tumulus detection using YOLOv3 where there were six TPs (white circles), three FNs (yellow circles) and a single FP (red circle): (**a**) output data; (**b**) satellite view.

Finally, there were 67 correctly detected burial mounds (TPs), 64.42% of the total. This indicates that numerous burial mounds were detected in Galicia despite the aforementioned FNs (Figure 6), showing their large-scale distribution (Figure 7).

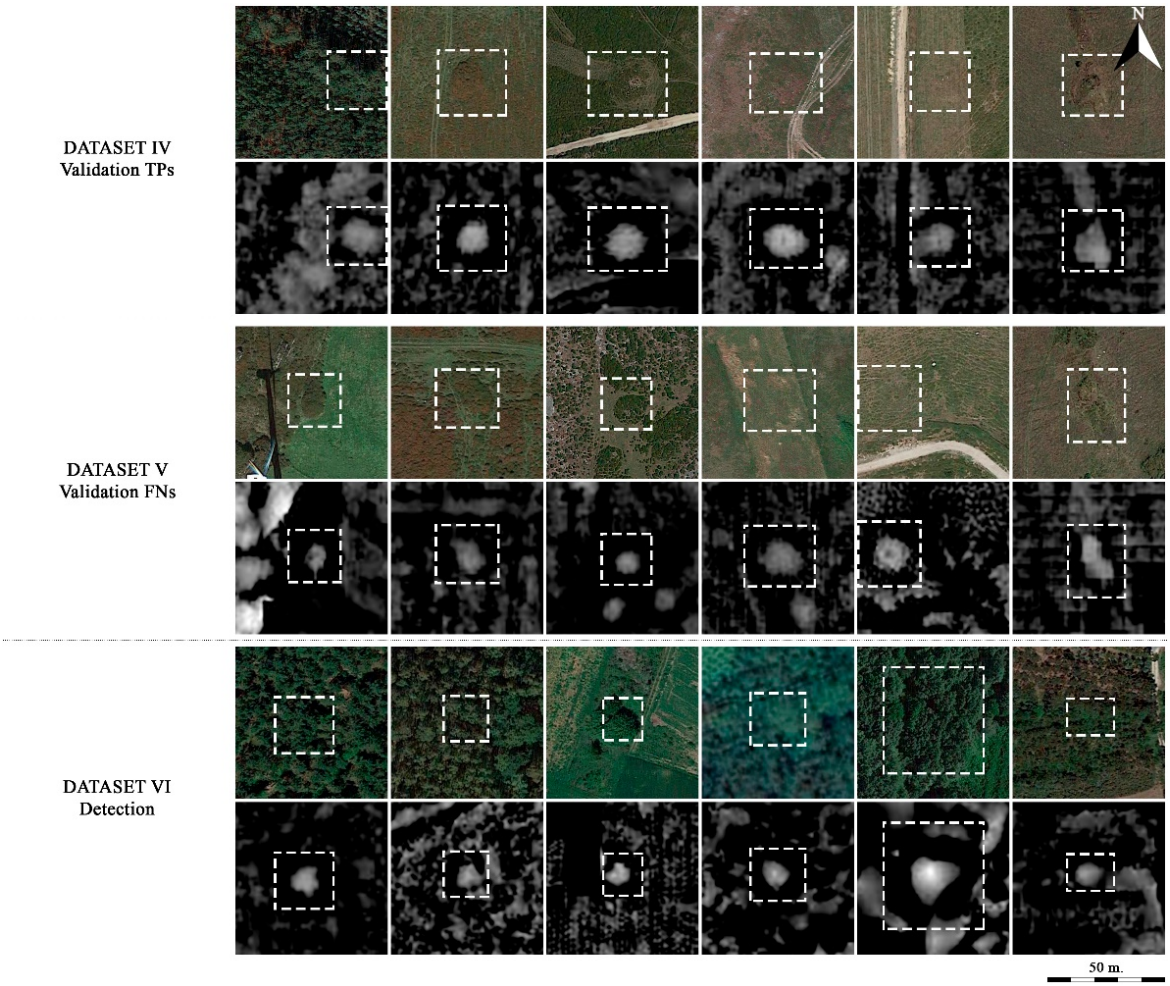
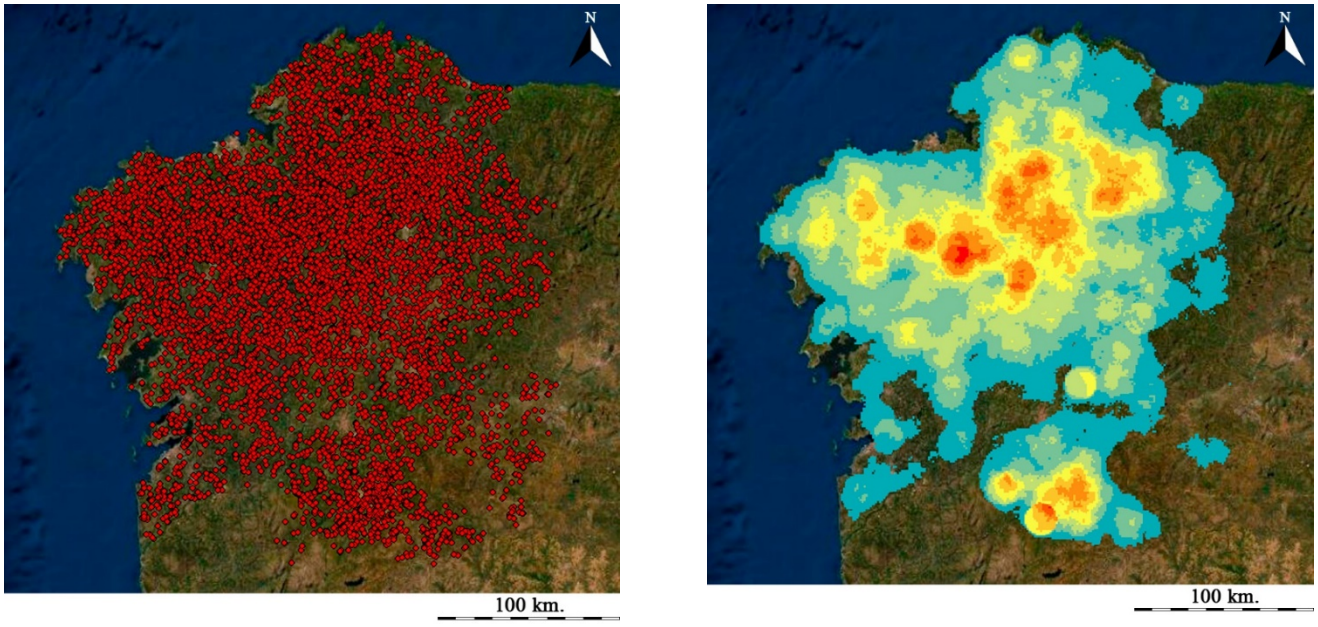


Figure 6. Validation TPs (Dataset V), FN (Dataset VI) data examples, and detections (Dataset VII). The latter were detected with a similarity of 100%, 90%, 80%, 60%, 40% and 25% (from left to right). The corresponding top image for each pair is a visible satellite image, shown for the sake of better visualization, but not used in our process.



(**a**) (**b**)

Figure 7. Detected tumuli in Galicia (Spain): (**a**) point distribution; (**b**) heat map.

3.5. Manual Model Validation

A last validation step consisted of manually evaluating the results. Although we extracted statistically significant performance metrics from the test dataset (see above), this dataset was extracted from a single area that did not have the variety of soil and land-use types present in the whole of the study area. As this can greatly influence the presence of FPs (e.g., areas with isolated houses could present false positives in the form of houses' roofs and eroded highland areas in the form of rock outcrops), a manual validation was considered necessary. This is a basic measure in archaeological detection studies, in particular with respect to mound detection work, as FPs tend to constitute a very high proportion of the detected features (see for example, [1,8]). For the manual visual inspection of the detected features, we used three different series of high-resolution imagery provided by Google, Bing, and ESRI, accessed as XYZ Tiles, and aerial imagery at 0.25 m/px provided by SIGPAC as a WMS through QGIS. These sources cover several years before and after the acquisition of the LiDAR dataset and helped to evaluate the possible presence of barrows independently of specific circumstances. From the 10,527 tumuli detected, we evaluated a total of 3086 individual tumuli in non-forested areas where the aerials allowed good visibility of ground conditions. We found that, of these, 324 corresponded to FPs, as follows: 225 were identified as rock outcrops, 33 as isolated houses' roofs, 9 as swimming pools and 57 to other mound-shaped features, most of them of anthropogenic nature. We should also note that, among this last type of FPs, some were only identified as FPs because of their context (such as mounds in golf courses) and were otherwise indistinguishable from archaeological tumuli. Mound identifications in forested areas were not considered to be FPs or TPs, as the only inspection method available for them was the LiDAR dataset, and this would have made it impossible for us to identify quite common occurrences such as rock outcrops.

Therefore, the manual validation indicated that 10.5% of the detected features were FPs, resulting in a detection rate of 89.5%. This suggests that, of the 10,527 tumuli detected approximately 9422 correspond to TPs. This number could be slightly higher, as approximately 23% of the tumuli are located in forested areas where, from all types of FPs, only rock outcrops (69% of the FPs) could be found.

Of course, this does not mean that all 9422 are archaeological tumuli, but their criteria did correspond to those used to identify them. Only a proper field survey and/or test pit excavations can truly document the archaeological nature of these remains, as there are many natural and human activities that could produce indistinguishable shapes in the same types of contexts.

In conjunction with the information provided by the presence of FNs (35.58% of the test data), our results suggest that the approximate number of tumular features that could correspond to archaeological tumuli in Galicia approximates 14,626 (9422 estimated TPs plus the estimation of those not detected according to the percentage of FNs).

4. Discussion

The automated detection of archaeological tumuli is a complex task given their common morphology. The study case presented here is particularly complex, considering the very large study area, the largest ever for this type of research. It includes multiple environmental conditions, land uses comprising urban, industrial, recreational and natural areas, and many other complex topographic settings such as granitic ranges and coasts which typically produce shapes similar to those of barrows.

Despite the complexity and scale of this study, the results are well beyond previous attempts to detect mounds using LiDAR data. The assessment of the test data provides a recall value of 0.64 (which means that the algorithm has detected a 64% of the known tumuli in the test area) and a precision of 0.97 (so 97% of the detections correspond to TPs). Further to that, the visual validation on randomly selected tumuli throughout the study area indicates that 89.5% of the detected features correspond to potential mounds,

a total of approximately 9422 tumuli. The most recent approaches to the detection of archaeological mounds using LiDAR-derived data are usually able to detect a high percentage of the test dataset's true mounds, but they also include a large proportion of FPs. For example, Davis et al. (2021) detected 17 of 18 mounds present in their study area but they also detected 3237 more mounds [1]. After visual validation, they confirmed that from the 3254 detected mounds, only 287 corresponded to possible burial structures (equivalent to an 8.8% success rate), pending field validation. Verschoof-van der Vaart et al. (2020) obtained a recall value of 0.796, but the precision value was 0.141 (86% of detected tumuli were FPs) [8]. Trier et al. (2021) detected 38% of known tumuli in their study area, but 89% of the detected features were identified as FPs [9].

4.1. Digital Terrain Model Pre-Processing

The use of MSRM instead of the most commonly used relief visualisation tools, such as LRM [25] and slope gradient [23,24], improved the detection rate of the algorithm. In contrast to other kernel-based methods, where the size of the feature can strongly affect its resulting shape, the multiscale nature of the MSRM produced more consistent shapes independently of the size of the tumuli. This is consistent with the results obtained by Guyot et al. (2021) [39], which, after comparing 13 microrelief visualisation methods, concluded that multiscale approaches consistently showed better performances in CNN-based detections.

4.2. Model Refinement

Despite the high detection and low FP rate of our algorithm, the recall value indicates that more training data could have improved the detection rate. An increment in the burial mound's training data would increase the variability in the shape of the tumulus. In the study area, there were only 584 known tumuli to work with (306 coming from previous works and 278 added in the refinement step), of which 478 were employed for training and 106 for validation. Theoretically, this number could be increased using DA, but given the circular nature of the mounds, the improvement with augmentation techniques such as DA2 or DA3 is very small. The obtained AP value without any DA was 68.31%, slightly lower than when implementing DA procedures (Table 3). This is because the validation mounds with a diameter of less than 18 m represent only 7.55% of the total. An increase of the DTM resolution to 0.5 m per pixel would allow a better detection of the smallest burial mounds, and an improvement in shape definition that would have allowed to distinguish FPs such as rock outcrops and houses' roofs.

4.3. Hybrid Model

Concerning the FPs, both the RF filtering and the refinement eliminated most of the possible FPs detected but we could still find some small rock outcrops, particularly those isolated and surrounded by soil types conductive to the detection of mounds. This is due to the 10-m-per-pixel ground resolution provided by Sentinel-2. In many cases, rock outcrops or other FP were located at the intersection of several 10 m pixels and the value taken was that of the positive pixel (i.e., that corresponding to a valid soil/land-use type) in some other cases the feature originating the FP was too small and the classification's pixel value was an average of the Sentinel-2 pixel footprint, resulting in the misclassification of the specific pixel. In relation to the previous point, the manual identification of FPs using high resolution imagery allowed us to identify elements that would not have been visible in lower-resolution images such as those acquired by Sentinel-2 (see the first column of Figure 8, where a rock outcrop is partially visible in between the tree cover). This was particularly the case for isolated small rock outcrops in grasslands, which constitute most of the FPs detected. With an improvement in the spatial resolution of the multispectral imagery employed, all those small diameter FPs in the RF classification stage would

have been removed. In this study, we preferred to employ Sentinel-2 given its public nature and the fact that it is directly accessible from GEE. However, the use of imagery from SPOT, RapidEye or any of the very high-resolution satellites commercially available could importantly reduce the presence of FPs. Another problem encountered with the use of multitemporal multispectral data for the classification of soil type/land use is precisely the multitemporal nature of the data. Although it helped to achieve better classification results, in some cases, particularly recent constructions, it included data from before the construction, producing median pixels classified as positive, which resulted in several houses' roofs being included as FPs.

It would have also been possible to include the multispectral bands as channels to the DL model (*channels* = 11) whose spectral data would have made it possible to discern the rocky areas from burial mounds and roofs from mounds in cultivation fields (which present similar values in RGB images). However, this would have resulted in a significant increase in the use of computational resources and training time.

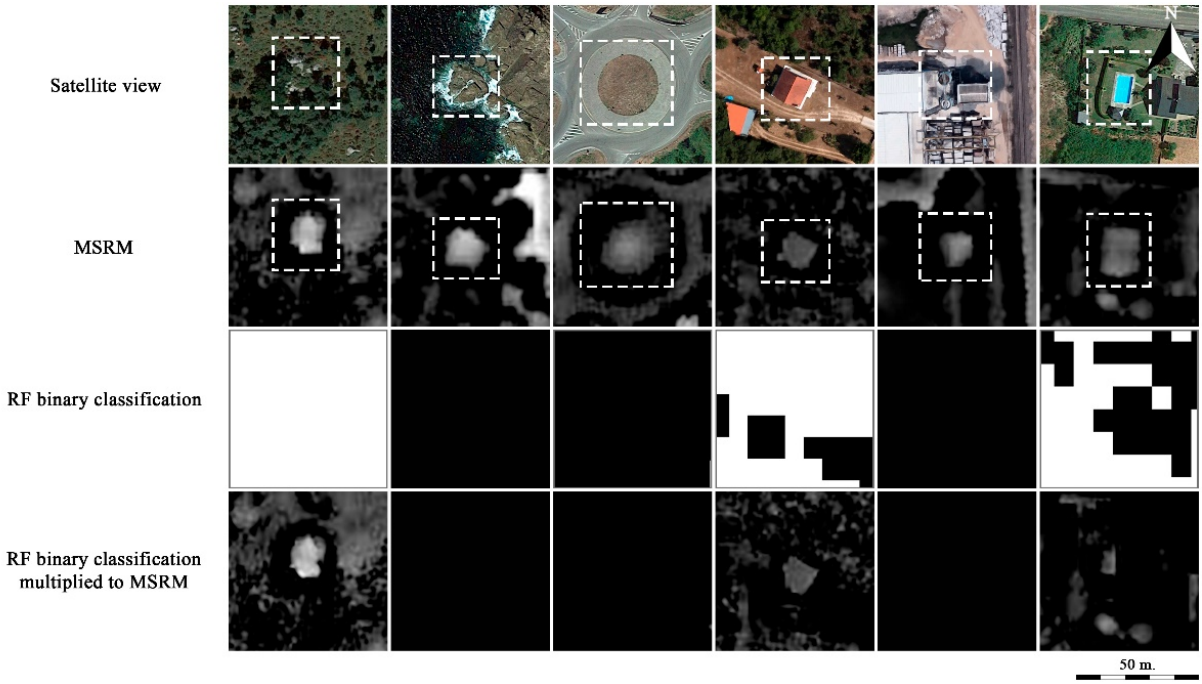


Figure 8. FP examples (Dataset III) obtained when we carried out the detection in Galicia and the same examples after applying the RF filtering. The corresponding top images are a high-resolution RGB satellite image, used for the manual validation, but not in the detection process.

4.4. Algorithm Accessibility and Reproducibility

The algorithm was designed to be accessible and reusable. All data employed are publicly available, even if the use of private imagery providers could have significantly improved the results of this study. The code is provided as Supplementary Material and is also available in GitHub (where future updates will be made available). Beyond evaluation and reproducibility concerns, the code has been designed to avoid personal limitations in computing power as a personal computer with an Internet browser and an Internet connection are the only requirements to apply the whole model. GEE is used to process the MSRM and RF classification (both very costly using desktop computers), and Google Colaboratory can be used to import the resulting raster and apply the YOLO algorithm seamless using a single cloud project. The design of the algorithm with a single channel source (the RF classification-filtered MSRM) instead of a costly multichannel DL approach significantly reduces computing costs allowing the detector to be applied over large areas using GEE and Colaboratory cloud computing resources.

The results and the detected FPs are also available as Shapefiles with associated metadata. In this way, these data can be used to better understand, manage and protect the cultural heritage of Galicia.

5. Conclusions

The algorithm presented in this paper constitutes an important improvement over previous and current approaches to the detection of archaeological tumuli and presents, for the first time, a valid alternative to the manual detection of this very common type of archaeological structure. The comparison of the results with regional heritage databases will make it possible to validate and improve both datasets. The large number of burial mounds detected in Galicia will allow the development of future investigations on their cultural distribution, achieving a better knowledge of the Galician megalithic complex.

Future research will implement newer versions of YOLO (v4 and v5, published during the development of this study), which improve the AP and the frame rate of YOLOv3. However, given the performance of the training algorithm presented here for the detection of burial mounds, our method already constitutes a practical tool that can be applied to any other areas where tumuli are present with few modifications, thus making it a general tool for archaeological research and cultural heritage management in many areas of the world. This is also prompted by making open-access the code presented in this work.

The process could also be greatly simplified by the use of Google Cloud Projects, where GEE and Colaboratory can be combined. GEE allows the ingestion of the user's preferred source for both LiDAR and satellite multispectral data (allowing to boost the results of this research with higher resolution sources without the need to modify the algorithm's code) and the training of the RF classification algorithm can be easily achieved within GEE using its simple vector drawing tools. Colaboratory's Jupyter notebook environment requires no configuration, runs entirely in the cloud, and allows the use of Keras, TensorFlow and PyTorch. It provides free accelerators like GPU or specialized hardware like tensor processing units, 12 GB of RAM, 68 GB of disk and a maximum of 12 h of continuous running.

Supplementary Materials: The following Supplementary Materials are available online at www.mdpi.com/article/10.3390/rs13204181/s1. Document explaining the use of the code and the scripts necessary to run it: script1.txt, script2.ipynb, JPEGtoPNG.atn, result.txt, script3.txt, resultsGIS.xlsx. Scripts can also be found in GitHub: https://github.com/horengo/Berganzo_et_al_2021_DTM-preprocessing (Accessed on 1 October 2021) and https://github.com/iberganzo/darknet (accessed on 1 October 2021).

Author Contributions: I.B.-B. and H.A.O. wrote the paper with the collaboration of all other authors. I.B.-B. created all illustrations. M.C.-P., J.F. and B.V.-E. provided training data and input during the evaluation of the results. I.B.-B., H.A.O. and F.L. designed the algorithm. H.A.O. designed the project and obtained funding for its development. All authors have read and agreed to the published version of the manuscript.

Funding: I.B.-B.'s PhD is funded with an Ayuda a Equipos de Investigación Científica of the Fundación BBVA for the Project DIASur. H.A.O. is a Ramón y Cajal Fellow (RYC-2016-19637) of the Spanish Ministry of Science, Innovation and Universities. F.L. work is supported in part by the Spanish Ministry of Science and Innovation project BOSSS TIN2017-89723-P. M.C.-P. is funded by the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie Grant Agreement No. 886793). J.F. is funded by the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie Grant Agreement No. 794048). Some of the GPUs used in these experiments are a donation of Nvidia Hardware Grant Programme.

Data Availability Statement: All relevant material has been made available as Supplementary Materials.

Acknowledgments: We would like to thank Daniel Ponsa (Computer Vision Center, Autonomous University of Barcelona) for his help in setting up the docker images and server access we employed for the development of this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Davis, D.S.; Gaspari, G.; Lipo, C.P.; Sanger, M.C. Deep learning reveals extent of Archaic Native American shell-ring building practices. *J. Archaeol. Sci.* **2021**, *132*, 105433.
2. Menze, B.H.; Ur, J.A. Mapping patterns of long-term settlement in Northern Mesopotamia at a large scale. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E778–E787.
3. Orengo, H.A.; Conesa, F.C.; Garcia-Molsosa, A.; Lobo, A.; Green, A.S.; Madella, M.; Petrie, C.A. Automated detection of archaeological mounds using machine-learning classification of multisensory and multitemporal satellite data. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 18240–18250.
4. Liss, B.; Howland, M.D.; Levy, T.E. Testing Google Earth Engine for the automatic identification and vectorization of archaeological features: A case study from Faynan, Jordan. *J. Archaeol. Sci.* **2017**, *15*, 299–304.
5. Orengo, H.A.; Garcia-Molsosa, A. A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery. *J. Archaeol. Sci.* **2019**, *112*, 105013.
6. Garcia-Molsosa, A.; Orengo, H.A.; Lawrence, D.; Philip, G.; Hopper, K.; Petrie, C.A. Potential of deep learning segmentation for the extraction of archaeological features from historical map series. *Archaeol. Prospect.* **2021**, *28*, 187–199.
7. Orengo, H.A.; Garcia-Molsosa, A.; Berganzo-Besga, I.; Landauer, J.; Aliende, P.; Tres-Martínez, S. New developments in drone-based automated surface survey: Towards a functional and effective survey system. *Archaeol. Prospect.* **2021**, 1–8. https://doi.org/10.1002/arp.1822
8. Verschoof-van der Vaart, W.B.; Lambers, K.; Kowalczyk, W.; Bourgeois, Q.P.J. Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 293.
9. Trier, Ø.D.; Reksten, J.H.; Løseth, K. Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *95*, 102241.
10. Guyot, A.; Hubert-Moy, L.; Lorho, T. Detecting Neolithic Burial Mounds from LiDAR-Derived Elevation Data Using a Multi-Scale Approach and Machine Learning Techniques. *Remote Sens.* **2018**, *10*, 225. https://doi.org/10.3390/rs10020225.
11. Trier, Ø.D.; Cowley, D.C.; Waldeland, A.U. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeol. Prospect.* **2019**, *26*, 165–175. https://doi.org/10.1002/arp.1731.
12. Soroush, M.; Mehrtash, A.; Khazraee, E.; Ur, J.A. Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in Kurdistan Region of Iraq. *Remote Sens.* **2020**, *12*, 500.
13. Orengo, H.A.; Petrie, C.A. Multi-scale relief model (MSRM): A new algorithm for the visualization of the subtle topographic change of variable size in digital elevation models. *Earth Surf. Process. Landf.* **2018**, *43*, 1361–1369.
14. Rodríguez-Casal, A.A. El fenómeno tumular y megalítico en Galicia: Caracterización general, problemas y perspectivas. In *Proceedings of the International Congress on the Study of Megaliths and Other Contemporary Burials in a Social, Economic and Cultural Context*, 1st ed.; Fernández-Eraso, J., Mujika-Alustiza, J.A., Eds.; Sociedad de Ciencias Aranzadi: Donostia-San Sebastián, Spain, 2007; pp. 58–93.
15. Carrero-Pazos, M.; Vilas-Estévez, B. The possibilities of the aerial LiDAR for the detection of Galician megalithic mounds (NW of the Iberian Peninsula). The case of Monte de Santa Mariña (Lugo). In *CAA2015. Keep the Revolution Going, Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, 1st ed.; Campana, S., Scopigno, R., Carpentiero, G., Eds.; Archaeopress: Oxford, UK, 2016; pp. 901–908.
16. Carrero-Pazos, M. El Fenómeno Tumular y Megalítico en Galicia. Aportaciones Desde los Sistemas de Información Geográfica y la Estadística Especial para el Estudio de los Patrones de Localización. Ph.D. Thesis, University of Santiago de Compostela, Santiago de Compostela, Spain, 2017.
17. Bourgeois, Q.P.J. Monuments on the Horizon: The Formation of the Barrow Landscape throughout the 3rd and the 2nd Millennium BCE. Ph.D. Thesis, University of Leiden, Leiden, The Netherlands, 2013.
18. Rodríguez-Del Cueto, F.; Carrero-Pazos, M. Límites y posibilidades de los análisis Lidar aplicados al megalitismo asturiano. Revisión de cuatro conjuntos tumulares prehistóricos en el concejo de Salas (España). *Veleia* **2021**, *38*, 9–31.
19. Carrero-Pazos, M.; Vilas-Estévez, B.; Romaní-Fariña, E.; Rodríguez-Casal, A.A. La necropolis del Monte de Santa Mariña revisitada: Aportaciones del LIDAR aéreo para la cartografía megalítica de Galicia. *Gallaecia* **2015**, 33, 39–57.
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1st ed.; IEEE: New York, NY, USA, 2014; pp. 580–587.
21. Berganzo-Besga, I.; Orengo, H.A.; Belarte, M.C.; Canela, J. Multitemporal lidar applied to the detection of architectural features in dense perennial Mediterranean forest. *Remote Sens.* **2021**, in preparation.
22. Información Xeográfica de Galicia. Available online: http://mapas.xunta.gal/portada (accessed on 1 September 2021).
23. Challis, K.; Forlin, P.; Kincey, M. A Generic Toolkit for the Visualization of Archaeological Features on Airborne LiDAR Elevation Data. *Archaeol. Prospect.* **2011**, *18*, 279–289.

24. Doneus, M.; Briese, C. Full-waveform airborne laser scanning as a tool for archaeological reconnaissance. In *From Space to Place. Proceedings of the 2nd International Conference on Remote Sensing in Archaeology. BAR International Series 1568*, 1st ed.; Campana, S., Forte, M., Eds.; Archaeopress: Oxford, UK, 2015; pp. 99–106.
25. Hesse, R. LiDAR-derived Local Relief Models—A new tool for archaeological prospection. *Archaeol. Prospect.* **2010**, *17*, 67–72.
26. Kokalj, Z.̌.; Somrak, M. Why Not a Single Image? Combining Visualizations to Facilitate Fieldwork and On-Screen Mapping. *Remote Sens.* **2019**, *11*, 747.
27. Zakšek, K.; Oštir, K.; Kokalj, Ž. Sky-View Factor as a Relief Visualization Technique. *Remote Sens.* **2011**, *3*, 398–415.
28. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27.
29. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767
30. Liu, G.; Xing, J.; Xiong, J. Spatial Pyramid Block for Oracle Bone Inscription Detection. In *ICSCA 2020, Proceedings of the 2020 9th International Conference on Software and Computer Applications*, 1st ed.; Association for Computing Machinery: New York, NY, USA, 2020; pp. 133–140.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149.
32. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1st ed.; IEEE: New York, NY, USA, 2017; pp. 936–944.
33. Bochkovskiy, A. YOLOv3. GitHub Repository. Available online: https://github.com/AlexeyAB/darknet (accessed on 25 January 2021).
34. Vilas-Estévez, B. Estudio de las Orientaciones y Emplazamientos de los Túmulos de la Necrópolis de la Serra do Laboreiro en base a la Arqueología del Paisaje y la Arqueoastronomía. Master's Thesis, University of Santiago de Compostela, Santiago de Compostela, Spain, 2015.
35. Torres, J. *Python Deep Learning. Introducción Práctica con Keras y TensorFlow2*, 1st ed.; Marcombo: Barcelona, Spain, 2020; pp. 231–253.
36. Lin, T. LabelImg. GitHub Repository. Available online: https://github.com/tzutalin/labelImg (accessed on 25 January 2021).
37. Orengo, H.A.; Petrie, C.A. Large-Scale, Multi-Temporal Remote Sensing of Palaeo-River Networks: A Case Study from Northwest India and its Implications for the Indus Civilisation. *Remote Sens.* **2017**, *9*, 735.
38. Garcia-Molsosa, A.; Orengo, H.A.; Conesa, F.C.; Green, A.S.; Petrie, C.A. Remote Sensing and Historical Morphodynamics of Alluvial Plains. The 1909 Indus Flood and the City of Dera Ghazi Khan (Province of Punjab, Pakistan). *Geosciences* **2019**, *9*, 21.
39. Guyot, A.; Lennon, M.; Hubert-Moy, L. Objective comparison of relief visualization techniques with deep CNN for archaeology. *J. Archaeol. Sci. Rep.* **2021**, *38*, 103027.