Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands

Verschoof, W.B.; Lambers, K.; Kowalczyk, W.J.; Bourgeois, Q.P.J.

International Journal of Geo-Information

MDPI

Article

Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands

Wouter B. Verschoof-van der Vaart [1,2,*,†], Karsten Lambers [1], Wojtek Kowalczyk [3] and Quentin P.J. Bourgeois [1]

1 Faculty of Archaeology, Leiden University, P.O. Box 9514, 2300 RA Leiden, The Netherlands; k.lambers@arch.leidenuniv.nl (K.L.); q.p.j.bourgeois@arch.leidenuniv.nl (Q.P.J.B.)
2 Data Science Research Programme (Leiden Centre of Data Science), Leiden University, P.O. Box 9505, 2300 RA Leiden, The Netherlands
3 Leiden Institute of Advanced Computer Science, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands; w.j.kowalczyk@liacs.leidenuniv.nl
* Correspondence: w.b.verschoof@arch.leidenuniv.nl; Tel.: +31-71-527-5277
† Current address: Faculty of Archaeology, Leiden University, P.O. Box 9514, 2300 RA Leiden, The Netherlands.

Received: 30 March 2020; Accepted: 22 April 2020; Published: 1 May 2020

Abstract: This paper presents WODAN2.0, a workflow using Deep Learning for the automated detection of multiple archaeological object classes in LiDAR data from the Netherlands. WODAN2.0 is developed to rapidly and systematically map archaeology in large and complex datasets. To investigate its practical value, a large, random test dataset—next to a small, non-random dataset—was developed, which better represents the real-world situation of scarce archaeological objects in different types of complex terrain. To reduce the number of false positives caused by specific regions in the research area, a novel approach has been developed and implemented called Location-Based Ranking. Experiments show that WODAN2.0 has a performance of circa 70% for barrows and Celtic fields on the small, non-random testing dataset, while the performance on the large, random testing dataset is lower: circa 50% for barrows, circa 46% for Celtic fields, and circa 18% for charcoal kilns. The results show that the introduction of Location-Based Ranking and bagging leads to an improvement in performance varying between 17% and 35%. However, WODAN2.0 does not reach or exceed general human performance, when compared to the results of a citizen science project conducted in the same research area.

Keywords: citizen science; Deep Learning; LiDAR; The Netherlands; Faster R-CNN

1. Introduction

The manual analysis of remotely sensed data is a widespread practice in present-day archaeology and heritage management [1]. However, the amount of available high-quality remotely sensed data is continuously growing at a staggering rate, which creates new challenges to effectively and efficiently analyze these data manually [2,3]. Especially the advancement of Light Detection And Ranging (LiDAR) techniques has opened up extensive areas for survey, which were up to now difficult to investigate due to forest and other vegetation cover [4]. LiDAR uses laser pulses to measure distance, based on precise measurements of time, resulting in a collection of three-dimensional data points. Airborne LiDAR can be used to record the surface of the Earth, documenting the topography of the area and objects appearing on it, with a high degree of accuracy [5,6].

In the last decade, archaeologists started using computational approaches to (semi-)automatically detect archaeological objects in remotely sensed data [7]. Most of these approaches have been based on Template Matching or Geographic Object-Based Image Analysis (GeOBIA [8]), and to a lesser extent on Knowledge-based or Machine Learning techniques (see Figure 1 in [9]). These often handcrafted algorithms oversimplify the detection problem and are generally unable to come close to human performance for complicated object detection tasks in varying contexts. Specifically, the large number of incorrect detections (false positives) compared to correct detections (true positives) make most of these algorithms of little practical value in large-scale archaeological mapping over different types of terrain [10]. Recent years have seen an increase in the use of Deep Learning [11–13], a subfield of Machine Learning, in many domains including archaeology (see [14]). The main architecture used in Deep Learning is the Convolutional Neural Network (CNN), an image feature extractor and classifier loosely inspired by the animal visual cortex [15]. Comparable to other Machine Learning approaches, a CNN learns to generalize from given examples (i.e., a large set of labeled images) rather than relying on a human operator to set parameters or formulate rules. Especially the possibilities offered by transfer-learning [16], where a CNN is pre-trained on a large, generic dataset and subsequently is fine-tuned on a small, specific dataset has made CNNs feasible for many domains that, up to now, were restricted by the small size of available labeled datasets [9]. In archaeology, transfer-learning has been successfully implemented on very-high-resolution satellite images from the Alps [17,18], as well as on LiDAR data from England [19], Norway [20], Scotland [21], and the Netherlands [22]. These approaches are mainly single class detectors that classify small extracts or snippets of data.

However, in archaeological prospection [23], obtaining the position of multiple objects in the wider landscape (i.e., localizing), is as important as characterizing them (i.e., classifying, the typical task of a CNN). This combination of localizing and classifying—referred to as object detection in Deep Learning—is handled by a specialized type of neural networks, so-called Region-based CNNs (R-CNNs [24,25]). These are able to localize and classify multiple, adjacent or even overlapping objects within a single image—as opposed to general CNNs that give a single classification for the entire input image [12].

1.1. WODAN

To explore the potential of R-CNNs for archaeological object detection in remotely sensed data, a workflow called WODAN1.0 (**W**orkflow for **O**bject **D**etection of **A**rchaeology in the **N**etherlands [22]) has been developed as part of an ongoing PhD research in the Data Science Research Programme at the Faculty of Archaeology and the Leiden Centre of Data Science at Leiden University, the Netherlands. The workflow consists of three parts (Figure 1): (1) a preprocessing part that converts LiDAR data into input images; (2) an object detection part consisting of an adapted version of the Faster R-CNN model [25]; and (3) a post-processing part that converts the results of the prior step into geographical features, directly usable in a Geographic Information System (GIS).

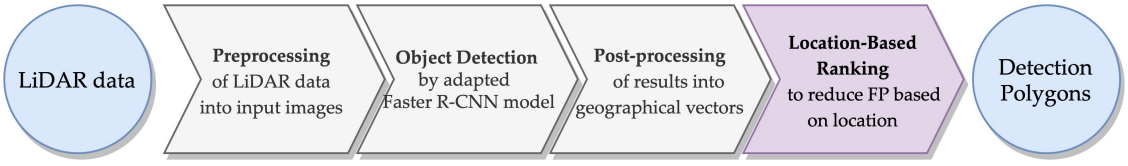


Figure 1. Simplified representation of WODAN1.0 and WODAN2.0 (with in purple the addition of Location-Based Ranking; FP stands for false positives); amended from Reference [9].

WODAN1.0 has been able to detect two different classes of archaeological objects, and thereby demonstrates that the Faster R-CNN model is usable as a multi-class archaeological object detector [9]. While the results of the experiments were promising, several points of improvement to the datasets and object detection model were identified: (1) the training dataset needed to be enlarged with more examples, to enable the detection of additional classes; (2) overlap needed to be introduced in all

datasets; and (3) the overall performance of the workflow needed to be improved by further adjusting the Faster R-CNN model [22].

Furthermore, in this study, we wanted to investigate the practical value of our object detection model for large-scale archaeological prospection over different types of terrain. Therefore, we needed to develop a large, random test dataset—that better represents the real-world situation of archaeological prospection—to replace the original small, non-random test dataset used in our prior research [22]. However, to make large-scale mapping feasible, an additional step needed to be added to the workflow to incorporate domain knowledge, to reduce false positives caused by specific regions in the research area, such as built-up areas, drift-sand areas, roads, and roundabouts [22]. This resulted in an updated version of the WODAN1.0 workflow, called WODAN2.0 (Figure 1), based on the aforementioned points of improvement. The development of WODAN2.0 is also part of the ongoing PhD research.

1.2. Outline of this Paper

In this paper, the WODAN2.0 workflow (Figure 1) is presented. The performance of this workflow for archaeological prospection is evaluated and compared to the results of WODAN1.0 [22] and a large-scale citizen science project [9] on the same test dataset. In Section 2, the research area and the datasets used are introduced. This is followed by an overview of the improvements made in Section 3, with a focus on the Location-Based Ranking step that has been added to the workflow (Section 4). In Sections 5 and 6, the results of the experimental evaluation are presented and discussed. The paper finishes with an overview of future developments planned (Section 7).

2. Research Area and Datasets

2.1. Research Area

The research area comprises the western part of the province of Gelderland in the Netherlands, known as the *Veluwe* (Figure 2). Nowadays, this area, approximately 2200 km^2 (circa 5% of the total area of the Netherlands), is predominantly covered by forest and heath, interspersed with agricultural fields and areas of habitation of various size (for a detailed overview of the research area, see [9,22]). The Veluwe holds one of the densest concentrations of known archaeological objects in the Low Countries, including prehistoric barrows [26] and Celtic fields [27], (post)medieval charcoal kilns [28], hollow roads [29,30], iron extraction pits, and *landweren* (border barriers), as well as more recent traces of conflict such as fortifications, military (support) structures, and bomb craters [31].

This research project is focused on detecting barrows, Celtic fields, and charcoal kilns (Figure 3). Barrows are round or oval-shaped earthen mounds that demarcate the burial place of a select group of people [32,33]. The majority of barrows on the Veluwe, individually or in small necropolises, were erected and used in the Neolithic and Bronze Age (between 2800 and 1400 cal BC [26,34]). Celtic fields are a characteristic checkerboard patterned parceling system from the late Bronze Age until the Roman Period (circa 1100 cal BC–200 AD), consisting of adjoining, roughly rectangular, embanked plots [35]. Charcoal kilns or charcoal burning platforms (generally known under the German term *Platzmeiler*) are the main remnants of pre-industrial charcoal production [36]. These consist of a shallow ditch or circle of pits surrounding a low, circular mound or platform, on which piles of wood, covered with sods, were carbonized under controlled conditions [28]. Above-ground charcoal kilns were mainly in use in the Low Countries from the Late Middle Ages until the second half of the 20th century (1250–1950 AD [37]).

In addition to the archaeological objects on record in various national archaeological databases, a recent analysis of LiDAR data from the research area—within the framework of this research project—has shown an abundance of prospective archaeological objects that were previously unknown [9]. The majority of the known and previously unknown archaeological objects are currently situated under heath or forest cover. While their location has almost certainly contributed to their present-day preservation, this also hinders the physical investigation and management of these objects

and restricts the field survey of the surrounding landscape for potential new archaeological objects (see also [38]).

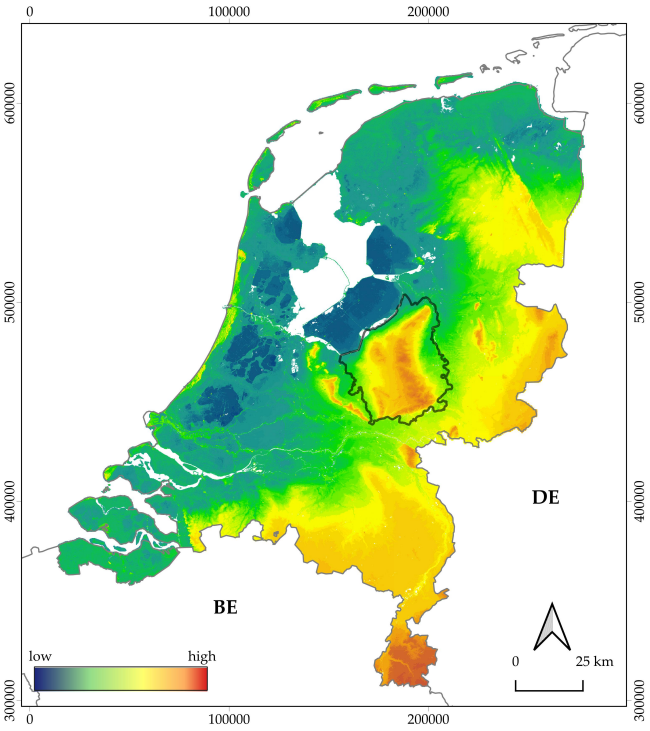


Figure 2. The research area on a height model of the Netherlands (source of the background image and height model: Reference [39]; coordinates in Amersfoort/RD New, EPSG: 28992; amended from Reference [9]).

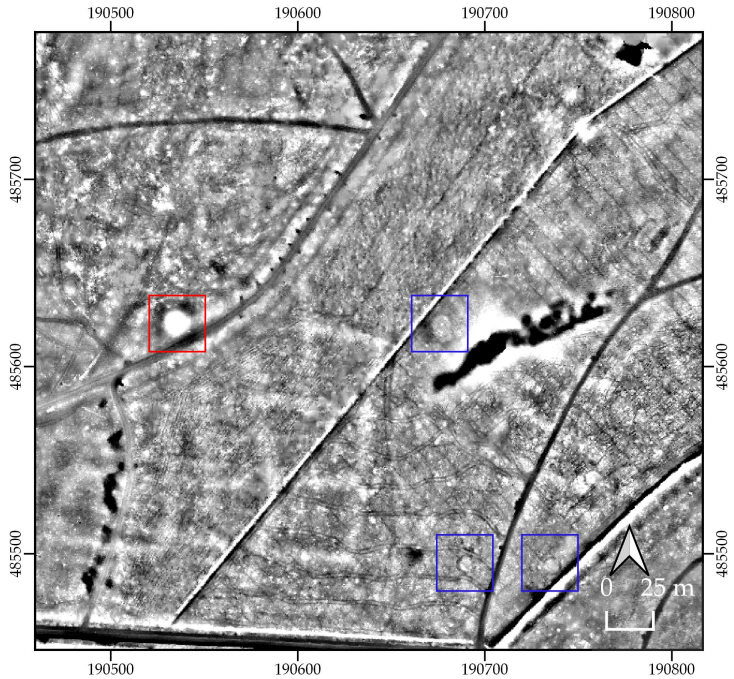


Figure 3. Excerpt of LiDAR data, visualized with Local Relief Model [40], from the research area (source height model: Reference [39]; coordinates in Amersfoort/RD New, EPSG: 28992), showing a barrow (red) and charcoal kilns (blue) in a Celtic field (white checkerboard pattern); from Reference [9]).

2.2. Datasets

For the research area, LiDAR data are freely available from the online repository PDOK [39] and the *Actueel Hoogtebestand Nederland* [41]. The LiDAR data were commissioned by the Dutch Directorate-General for Public Works and Water Management and collected by helicopter with a Riegl LMS-Q680i scanner in April 2010. The data were classified into non-ground and ground points and interpolated, resulting in a digital terrain model (ground points) with an average point density of 6–10 per m^2, a spatial resolution of 50 cm, and a vertical and planimetric accuracy of 5 cm [42]. The data are disseminated in GeoTIFF tiles measuring 10,000 by 12,500 pixels (5 km by 6.25 km). In the prior research, fourteen tiles (in total, 437.5 km^2) were dissected into 2940 subtiles measuring 1000 by 600 pixels. Then, 492 subtiles that contained archaeological objects were selected to train, validate, and test WODAN1.0 [22]. To enlarge the training dataset in the current research, two additional tiles (62.5 km^2), predominantly containing examples of charcoal kilns, were added (increasing the total area used to 500 km^2). Furthermore, in the datasets of WODAN1.0, only distinct examples of the archaeological objects (e.g., reconstructed barrows) were included. To enlarge the number of objects in the WODAN2.0 datasets, less conspicuous examples, in various state of preservation, were also added. To validate the Location-Based Ranking approach (see Section 4.2), 125 km^2 of LiDAR data from the northwestern Veluwe were used. These data do not coincide with the test datasets.

To construct the datasets, sixteen tiles of interpolated LiDAR data were downloaded. Thirteen tiles were designated as training data, one tile as validation data and two tiles as testing data. The tiles were loaded into *QGIS 3.4 Madeira* [43] and a *Fill_nodata* processing tool was used to reduce the number of no-data points. Subsequently, the tiles were visualized with the Local Relief Model visualization [40] from the *Relief Visualisation Toolbox 1.3* [44]. All tiles were sliced into subtiles of 600 by 600 pixels with 30 pixels overlap on all sides. The latter was done to eliminate potential edge effects resulting from the visualization of the LiDAR data (see [45]), and to avoid the dissecting of archaeological objects on the edges of subtiles in the datasets (see also [22]). Subtiles that contained archaeological objects were selected and labeled with *LabelImg*, a graphical image annotation tool to label object bounding boxes in images [46]. This resulted in a training dataset of 1024 subtiles and a validation dataset—used to monitor the model during training—of 88 subtiles (Table 1).

Table 1. The datasets used in this research (numbers in parentheses concern WODAN1.0, after Reference [22]). For convenience, the non-random test dataset is listed as well. The discrepancy in the number of Celtic fields in both test datasets is due to a change from counting individual plots to demarcated areas.

Dataset	Subtiles	Barrows	Celtic Fields	Charcoal Kilns	Objects
training	1024 (380)	1261 (805)	1504 (667)	575 (177)	3340 (1649)
validation	88 (39)	127 (49)	64 (199)	22 (24)	213 (272)
non-random test	73	78	235	23	336
random test	828	137	65/2.56 km^2	26	363
low confidence		65	1.48 km^2	14	
high confidence		72	1.08 km^2	12	

2.3. Test Datasets

To evaluate the results of WODAN2.0 a large, random test dataset or reference standard [47] was created to replace the original small, non-random test dataset. To create the reference standard, two expert researchers—the first and fourth authors, who both have ample experience in analyzing LiDAR data and considerable knowledge of the archaeology of the research area—independently classified archaeological objects in 828 subtiles from two separate areas on the Veluwe. Both areas have been extensively studied in the (recent) past [26,27], and contain multiple examples of the archaeological classes, in various states of preservation.

The LiDAR data were similarly pre-processed as the training and validation datasets (see above). The classifications were done in *LabelImg* [46] and brought together and compared in *QGIS 3.4 Madeira* [43]. Inter-analyst variability (also see [48]) was resolved by assigning different levels of confidence to individual classifications: objects that were marked by both researchers and/or extant archaeological objects on record in any of the national archaeological databases were given high confidence, while objects, marked by only one researcher, were given low confidence. The resulting random test dataset (see Table 1) consists of all 828 subtiles of which 164 contain in total 137 examples of barrows and 26 charcoal kilns. The total area covered by Celtic fields equals 2.56 km^2 spread over 65 demarcated areas. The discrepancy in the amount of Celtic fields between the non-random and the random test dataset derives from the fact that in the non-random dataset every individual plot within a Celtic field is counted as one example, while in the random dataset every demarcated area covered with Celtic fields, which can contain multiple individual plots, is counted as one example.

In comparison, the random test dataset includes 828 subtiles (of 600 by 600 pixels) of which only 164 (19.8%) contain a total of 363 objects, while the non-random test dataset consists of 73 subtiles (of 1000 by 600 pixels) of which 63 (86%) contain 336 objects in total (see Table 1). Therefore, the proportion of subtiles with or without archaeological objects on it (i.e., positive or negative subtiles) varies greatly between 6.7:1 (positive:negative) for the non-random and 1:4 (positive:negative) for the random test dataset (see also Section 3.3). Therefore, the random test dataset could better represent the real-world situation of the prospection of scarce archaeological objects and gives a better impression of the practical value of the object detection model.

2.4. Heritage Quest Dataset

The *Heritage Quest* project, from Leiden University and *Erfgoed Gelderland* [9,49], is the first large-scale citizen science project involving the archaeological interpretation of remotely sensed data in the Netherlands, and is conducted in the same research area as the current study (see Figure 2). Members of the public, generally called citizen researchers [50], are actively involved in two stages of archaeological prospection: (1) the classification of archaeological objects in LiDAR data; and (2) the validation of potential archaeological objects in the field. Professional archaeologists from the organizing institutions assist and direct both stages of the research. This approach, directly involving citizen researchers in the collection and/or interpretation of data, is uncommon in archaeology, although community engagement has been a long recurrent practice [51].

In the first stage of Heritage Quest—the classification of archaeological objects in LiDAR data—the web-based citizen science platform Zooniverse [52] was used. Participants were shown LiDAR snippets of 300 m by 300 m (600 by 600 pixels) from the research area and asked to mark the location of every potential barrow, Celtic field, and charcoal kiln. The participants were presented with two different LiDAR visualizations (shaded relief and Local Relief Model; see [44]) to assist them in their classification. This stage of Heritage Quest produced circa 120,000 detections, spread over the entire research area. Every individual LiDAR snippet was classified by fifteen different users before it was retired, therefore providing possibilities to aggregate the classifications and to explore inter-analyst agreement [53]. This type of "consensus" [54] improves accuracy of the classifications and is an established method to produce reliable data by guaranteeing minimal inter-analyst variability [55].

The task performed in the online Heritage Quest project and the object detection task performed by WODAN2.0 are very similar in design and execution, and are implemented on the same dataset. Therefore, the performance of both can be compared and the results of Heritage Quest offer us a benchmark for human performance on the task of detecting barrows, Celtic fields, and charcoal kilns in LiDAR data from the Veluwe. Although citizen science arouses skepticism among some scientists [56,57], datasets produced by and performance of citizen researchers can be of reliable high quality, on par with those from professionals, if appropriate strategies are employed in the design, execution, and validation of the project [54,55,57]. We are aware that the performance of a group of citizen researchers, with predominantly little experience in both archaeology and remote sensing, does not necessarily

equal the performance of experts or even novel experts (e.g., students). However, studies have shown that the difficulty of the task is a more important predictor of performance, rather than background, experience, or locality [58,59]. To determine the quality and reliability of the Heritage Quest data, the performance was tested on the large, random test dataset (see Section 5.2).

3. Methodology

In the main part of the WODAN2.0 workflow (Figure 1), an adapted version of the Faster R-CNN model, written in *Python 3* [60] and *Keras* [61] (see also [62]), is employed to detect barrows, Celtic fields, and charcoal kilns in LiDAR images. Faster R-CNN is one of the latest instalments of R-CNN [24]. The concept of the original R-CNN architecture is: (1) produce object proposals with Selective Search [63]; (2) extract features for every object proposal with a CNN; (3) classify whether a proposal contains an object of interest with a Support Vector Machine (SVM); and (4) use a linear regressor to tighten the bounding box to fit the true sizes of the object [24]. Fast R-CNN, the successor of R-CNN, improved on its predecessor by speeding up the feature extraction and classification step, and by joining the CNN, SVM, and linear regressor into one CNN model [64]. Further improvements were made to speed up the object proposal step, resulting in the Faster R-CNN model [25] that is used in this research. Faster R-CNN utilizes a fully connected convolutional Region Proposal Network (RPN) to generate object proposals (instead of Selective Search). The feature extraction and classification of the candidate regions is done with the Fast R-CNN model. Both the RPN and Fast R-CNN are trained simultaneously during the training of Faster R-CNN [65].

Faster R-CNN was selected for this research because this model has achieved great success in detecting (small) objects in natural scene images [12], and it generally outperforms the traditional sliding window based methods [65] and single shot object detectors [66]. As the backbone network of the Faster R-CNN model, VGG16 [67] was used. This CNN performs better than most shallower networks and needs significantly less memory than some deeper networks, while yielding comparable results [68]. To improve the performance of the Faster R-CNN model, specific measures were adopted that are discussed in detail below.

3.1. Anchor Box Sizes

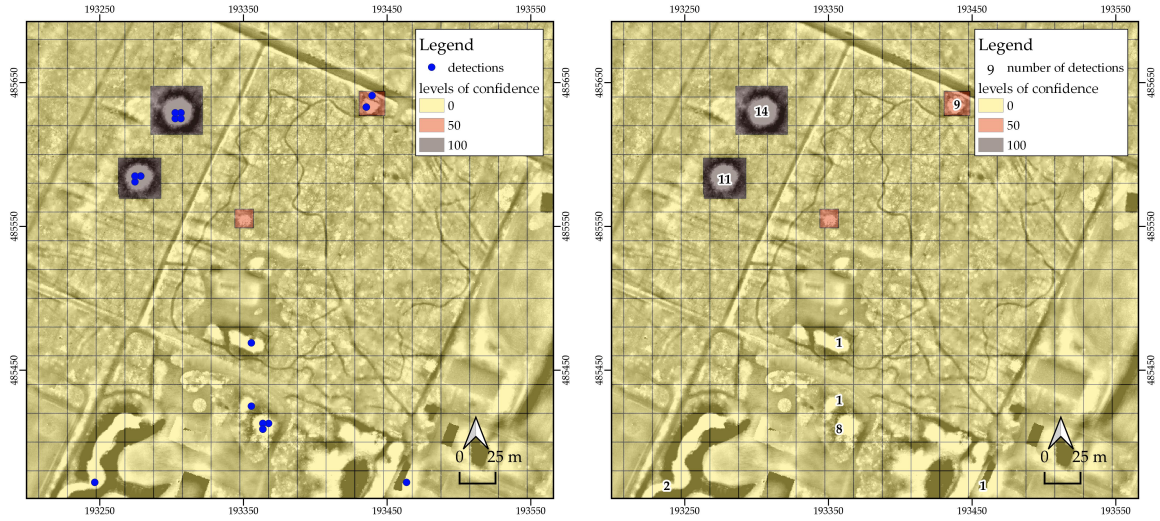
In WODAN1.0, it was already noticed that the RPN anchor boxes were too large for most objects in the datasets [22]. Several researchers have noted that the performance of Faster R-CNN on small objects can be improved by lowering the sizes of the anchor boxes [65,69,70]. Based on the approximate size of the archaeological objects, the size of the square shaped anchor boxes was lowered to 16^2, 64^2, and 512^2 pixels. For the aspect ratios of the anchor boxes, the values of the original paper (1:1, 1:2, and 2:1) were maintained [25].

3.2. Bootstrap Aggregating

Bootstrap aggregating (or bagging) is a form of ensemble learning used to improve the stability and performance of Machine Learning and Deep Learning classification and regression algorithms [71]. Bagging also reduces variance and helps to avoid overfitting. The concept of bagging is threefold: (1) bootstrapping of the training dataset; (2) training of multiple models; and (3) aggregating of the predictions of these models. Bootstrapping involves the repeated resampling with replacements of a dataset into a number of new datasets. If the size of the new datasets equals the size of the original dataset, the former are expected to have circa 63% of the unique examples of the original training dataset, the rest being duplicates [72]. Bootstrapping of the datasets in this research was done in *Python 3* [60] by randomly selecting and copying an image from the original training dataset into an new, resampled dataset and repeating this action a number of times equal to the total number of images in the original dataset. After bootstrapping, a number of models, equal to the number of resampled datasets, are trained with the same (hyper)parameters. After training, the models are tested

on the same test dataset and the outputs are aggregated into a single result, for instance through majority voting.

However, in this research, the outputs of the testing have a spatial element and need to be combined based on their position in relation to areas containing archaeological objects, as opposed to a specific location. Therefore, a GIS-based spatial aggregation method was developed (see Figure 4) to facilitate the combination of the results of the bagging, to ease the comparison of the results with other (archaeological) geospatial data, and to provide opportunities to visualize the results (see [73]). In the post-processing step of WODAN2.0, the output of the object detection step is converted from bounding boxes with pixel coordinates into geospatial features with real-world coordinates (see [22]). Barrow and charcoal kiln detections are converted into points by taking the central coordinate of the bounding box. These points are compared to a map of the test area that has been divided into cells of 20 by 20 m^2 for barrows and 15 by 15 m^2 for charcoal kilns, based on the average size of these archaeological objects (Figure 4). The points are aggregated by counting the number of them within each cell through the *Join_by_location* processing tool in *QGIS 3.4 Madeira* [43]. The bounding boxes depicting Celtic fields are turned into polygon features. Subsequently, the features are combined into larger polygons, turned from multipart input features to individual singlepart features, and overlapping polygons are joined with, respectively, the *Union*, *Multipart_to_singlepart* and *Spatial_Join* processing tools in *ArcMap 10.6.1* [74]. These polygons are subsequently compared to a spatial layer containing polygon features for all the confirmed Celtic fields in the test area.



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

Figure 4. Excerpts of LiDAR data, visualized with Local Relief Model [40], showing: (**a**) the results of the post-processing step with points for individual barrow detections (blue) on the cells of the reference standard (with different levels of confidence); and (**b**) the results of the spatial aggregation showing the number of detections per cell (with different levels of confidence); (source of the height model: Reference [39]; coordinates in Amersfoort/RD New, EPSG: 28992).

3.3. Negative Examples

One of the drawbacks of Faster R-CNN is that the Region Proposal Network cannot adequately distinguish small objects from complex backgrounds [65]. Especially in large remotely sensed imagery, the backgrounds are generally more intricate than in natural scene images. According to Tang et al. [65], this problem is due to the lack of negative examples, only containing background, in the training dataset. By adding these, the model learns their specific texture features. Thus, when these areas are subsequently detected during testing, the model is trained to recognize them as (non-archaeological) background areas [75]. To investigate whether adding negative examples improves performance, an additional training and validation dataset was created containing, besides subtiles with

archaeological objects (i.e., positive subtiles), also subtiles without archaeology (i.e., negative subtiles; see Table 2). As mentioned in Section 2.3, both test datasets already contained subtiles without archaeological objects. As shown by Gao et al. [75], the proportion of positive and negative subtiles is of influence to the performance of the model and a proportion between 1:1 and 1:2 (positive:negative subtiles) for training data yields the best results. In this research, a proportion of circa 1:1.6 was used. The proportion of the validation and random test dataset is 1:3 and 1:4, respectively, to better simulate the real-world scarcity of archaeological objects in the landscape. Separate experiments were performed with WODAN2.0 trained on the training and validation datasets with and without negative examples (see Section 5).

Table 2. The number of positive and negative subtiles in the datasets. For convenience, the non-random test dataset is listed as well (after Reference [22]).

Dataset	Positive Subtiles	Negative Subtiles	Proportion
training	1024	1634	1:1.6
validation	88	259	1:3
non-random test	63	10	6.7:1
random test	164	664	1:4

4. Introducing Domain Knowledge: Location-Based Ranking

To make the WODAN2.0 workflow usable in large-scale archaeological mapping over different types of complex terrain, domain information needed to be introduced to the classification to reduce the number of false positives caused by specific regions in the research area, such as built-up areas, drift-sand areas, and roundabouts [22]. This resulted in the Location-Based Ranking (LBR) step being implemented in WODAN2.0 (Figure 1). The basic assumption of LBR is comparable to archaeological predictive modeling [76,77] in that it is assumed that the location of archaeological objects in the present landscape is not random, but is, among others, the result of certain characteristics of the past and present environment. These landscape characteristics, such as subsoil and (current) land-use, either influence the preservation of archaeological objects (e.g., erosion and deposition) or restrict the ground visibility conditions (e.g., vegetation and agricultural practices) [78]. For instance, gravel quarries will have destroyed archaeological objects within their confines, while certain agricultural practices effectively act as 'blankets' for remote sensing techniques, greatly reducing visibility [79]. A comparison of the distribution of a sample of known archaeological objects with the driving characteristics allows for an exploratory analysis of relationships between them. These trends can subsequently be extrapolated to a larger area, i.e., the research area [77]. Although archaeological predictive modeling attempts to incorporate social and cognitive factors of past human behavior, LBR focuses on post-depositional processes rather than choice of location and could be considered more related to environmentally-based predictive modeling (see [76]).

Location-Based Ranking consists of determining, ranking, and mapping of the principal (present-day) landscape characteristics, such as subsoil and land-use, that have had an impact on the preservation and/or visibility of archaeology. The influential characteristics within a research area can be determined based on prior research in the formation of the archaeological landscape and/or by a broad-brush landscape characterization (see [80]). The subsequently assigned ranks correspond to the potential for the occurrence of specific types of archaeological objects within that zone. For instance, the formation of large scale drift-sand areas on the Veluwe, starting in the (Late) Middle Ages [81], has had a negative impact on the preservation and visibility of barrows and Celtic fields. Drift-sand areas can therefore be considered to have a low potential for the occurrence of these objects. After determining and categorizing, the effectiveness of the chosen characteristics and the ranking system can be evaluated by a validity test (see Section 4.2).

The result of Location-Based Ranking is a ranked map of the research area (see, for example, Figure 5) on which the location of detections, in our case from the object detection step, can be

compared and assigned to different ranks. Detections in high ranking zones are more likely to be archaeological objects, while detections in low ranking zones have a much higher likelihood of being false positives. Therefore, Location-Based Ranking can be used to reduce the amount of false positives by ignoring detections in low ranking areas.

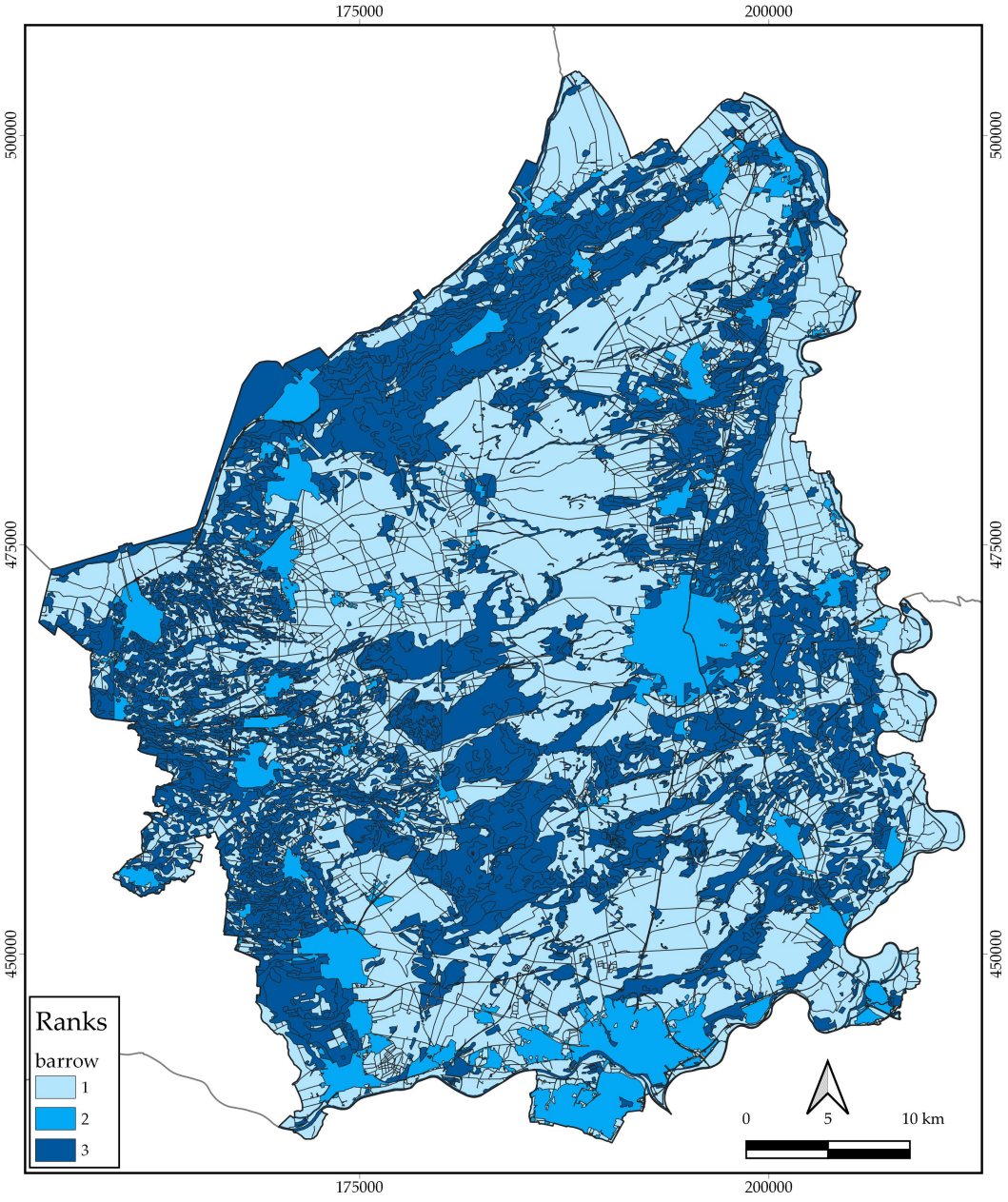


Figure 5. Location-Based Ranking map for the Veluwe research area (coordinates in Amersfoort/RD New, EPSG: 28992) showing the ranks of the areas for barrows in shades of blue (see legend).

4.1. Location-Based Ranking in Practice: The Veluwe

In the current study, Location-Based Ranking was implemented on our research area, the Veluwe (Figure 2). To determine the principal landscape characteristics involved, recent research on the distribution of barrows and the formation of barrow landscapes on the Veluwe was taken as a starting point [26]. This evaluation identified two processes as being the most detrimental to the preservation and visibility of barrows (and other archaeological objects; see, e.g., [27]) on the Veluwe: erosion and sedimentation by wind (i.e., medieval drift-sand) and (post)medieval agricultural activities and

urbanization [26]. The former has eroded barrows and other prehistoric objects and/or covered them with sand dunes. The latter have either destroyed (urbanization) or covered (agricultural activities) barrows and other archaeological objects. (Post)medieval agriculture on the Veluwe is evidenced by the presence of *enken* or *plaggen* soils [82]. *Enken* are arable complexes with an anthropogenic topsoil that has formed through the centuries-long spreading of heather or grass sods (*plaggen*) mixed with animal manure over the fields [83]. Generally, prior to being turned into arable land, above-ground (prehistoric) objects were razed to produce level fields [79]. Subsequently, the fields were gradually raised with layers of soil on top of the old (prehistoric) surface. Several other present-day landscape features (called 'badlands' in this research) that either entail ground disturbances (such as dikes, golf courses, and quarries) or wet areas (such as marshes, surface water, and crevasse splays) have also had a (major) negative impact on the preservation and visibility of archaeological objects. Urbanization and concurrent infrastructural developments (i.e., roads) have been of less impact than the above features. The best chances for survival can be found in heathland and forested areas (see also Table 4.1 in [26]).

Based on the above, a three-tiered ranking map (Table 3 and Figure 5) for the research area was developed, using open-source geo(morph)ological and topographical data from the online spatial data repository PDOK [39]. The digital geological map (*Bodemkaart van Nederland*, scale 1:50.000) and geomorphological map (*Geomorfologische Kaart van Nederland*, scale 1:50.000) were used to determine the location of drift-sand areas, *plaggen* soils, and 'badlands' (e.g., dikes, quarries, etc.). The digital topographical map of the Netherlands (*Basisregistratie Grootschalige Topografie*) was used to demarcate built-up areas. Roads were extracted from the national road dataset (*Nationaal Wegen Bestand*). The following ranks were determined:

- The lowest rank (3) is given to barrow and Celtic field detections in drift-sand areas. Charcoal kiln detections in drift-sand areas are given the highest rank (1). Any detections, regardless of class, in (post)medieval agricultural areas (*plaggen* soils) or in 'badlands' (e.g., dikes, quarries, etc.) are also given this lowest rank.
- The middle rank (2) is given to detections located in urbanized or built-up areas and in the direct vicinity of roads. While many Celtic fields are intersected by roads, this has had a limited negative impact on the preservation of the overall objects. Therefore, roads are considered Rank 1 in the case of Celtic fields.
- Any detections not located in one of the aforementioned zones are given the highest rank (1). These are generally located in heathland or forested areas, and in the case of charcoal kilns also in drift-sand areas.

Table 3. Different landscape features and their rank in the Location-Based Ranking map for the Veluwe.

Landscape Features			Rank		
Type	**Area (km^2)**	**Ratio of Research Area (%)**	**Barrow**	**Celtic Fields**	**Charcoal Kilns**
drift-sand	338.4	15.2%	3	3	1
plaggen soils	460.8	20.7%	3	3	3
badlands	73.4	3.3%	3	3	3
build-up	218.3	9.8%	2	2	2
roads	42.2	1.9%	2	1	2
other	1090.6	49.1%	1	1	1
total	**2223.7**	**100%**			

4.2. Validity Test

To test the validity of the proposed LBR map for the research area, the locations of all known and extant barrows, Celtic fields, and charcoal kilns in a 125 km^2 area on the northwestern Veluwe (see Section 2.2) were ranked. This area contains several minor and major villages, an extensive road network (including a motorway and smaller roads), different areas of drift-sand, and multiple quarries.

Table 4 shows the results of the ranking. In the case of all three archaeological classes, more than 93% of the objects or area can be assigned to the highest rank (1). This shows the effectiveness of the ranking system and the landscape characteristics chosen. Furthermore, it demonstrates that, by only considering Rank 1 detections—ignoring detections in Ranks 2 and 3—the number of missed archaeological objects will be low, while the number of false positives, caused by these zones, will be reduced.

Table 4. Validity test of the Location-Based Ranking map for the research area.

Rank	Archaeological Objects					
	Barrows		Celtic Fields		Charcoal Kilns	
	Number	Ratio	m^2	Ratio	Number	Ratio
1	341	93.4%	414.7	100%	174	99.4%
2	17	4.7%	0	0%	0	0%
3	7	1.9%	0	0%	1	0.6%
total	365	100%	414.7	100%	175	100%

5. Experimental Evaluation

5.1. Implementation Details

In our experiments, we used the adapted version of the Faster R-CNN model (as detailed above) with VGG16 [67] as the backbone network. The Faster R-CNN model [62] was written in *Python 3* [60] and *Keras* [61]. VGG16 was pre-trained on the ImageNet image dataset [84] and fine-tuned on our own training dataset (see Section 2.2). The training dataset was resampled fifteen times in *Python 3* [60], and fifteen adjusted Faster R-CNN model were fine-tuned for fifteen epochs with a learning rate of 1×10^{-5}. We used stochastic gradient descent with the Adam optimizer [11], implemented in *Keras* [61]. To cope with the fact that the input images are grayscale, these were turned into RGB by copying the value from the first channel to the other two color channels, as is done by default in Keras' *ImageDataGenerator* [61]. In the training process, the sizes of the anchor boxes were lowered following Section 3.1, and the input images were flipped horizontally and vertically, as well as rotated. Every two epochs the model was validated on the validation dataset (Table 1).

To investigate whether the addition of negative subtiles improved performance (see Section 3.3), the above training regime was repeated with the training and validation datasets containing negative subtiles (indicated with NEG in Table 5). All experiments were performed on an NVIDIA Tesla K80 GPU.

Both WODAN1.0 and WODAN2.0 were tested on the non-random and random dataset (see Section 2.3), indicated with, respectively, (NR) and (R) in Table 5. Heritage Quest was also tested on the random dataset. To evaluate the workflow(s), the number of true positives (TP), false positives (FP), and false negatives (FN) were determined and the commonly used metrics for measuring the performance of object detection models were calculated [85]: recall (R; Equation (1)), precision (P; Equation (2)), and the F1-score (F1; Equation (3)). Recall gives a measure of how many relevant objects are selected. Precision measures how many of the selected items are relevant. The F1-score is the harmonic average of the precision and recall and a measure of the model's performance per class [85]. These measurements are normally restricted between 0 and 1, with higher values indicating a better performance. For readability, the values for all metrics are presented in percentages (see Table 5).

$$R = \frac{TP}{(TP + FN)} \tag{1}$$

$$P = \frac{TP}{(TP + FP)} \tag{2}$$

$$F1 = 2 \times \frac{R \times P}{(R + P)} \tag{3}$$

Every detection, generated during the object detection step, consists of a rectangular bounding box with a category label and a softmax or confidence score (range 0–100) [25]. The confidence threshold is typically set to 80: if the confidence score equals or exceeds 80, the detection is outputted by the object detection model, otherwise the detection is discarded [25]. However, by changing the threshold, redetermining the number of TP, FP, and FN, and recalculating the performance metrics, an optimal trade-off between recall and precision can be found, resulting in the highest F1-score [86]. The same can be done for the number of detections within a grid cell (see Figure 4), resulting from the aggregation of the bagging (see Section 3.2) or the Heritage Quest results (see Section 2.4). For instance, the number of detections per grid cell can have a threshold set to three, meaning that only grid cells with three or more detections in it will be taken into account. In this research, the confidence threshold was varied between 80 and 91, with intervals of 1, and the threshold for the number of detections per grid cell was varied between 1 and ≥10, with intervals of 1 (see Table 6). By finding the optimal trade-off between confidence and number of detections per grid cell, the highest F1-score is obtained. However, a drawback of using thresholds is that the maximum achievable precision and recall is not shown. Therefore, in Table 7, the results of both WODAN2.0 and Heritage Quest without the use of thresholds are shown.

Table 5. The performance of WODAN1.0, WODAN2.0, and Heritage Quest per archaeological class on the non-random (NR) and random (R) test dataset. NEG indicates the training and validation datasets with negative subtiles were used.

Method	Barrows			Celtic fields			Charcoal kilns		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
WODAN1.0 (NR)	62.3	55.2	58.5	82.3	57.6	67.8	–	–	–
WODAN1.0 (R)	53.3	9.0	15.3	43.0	20.5	27.7	–	–	–
WODAN2.0 (NR)	67.1	73.3	70.1	74.6	66.0	70.0	–	–	–
WODAN2.0 (R)	44.5	56.5	49.8	40.4	52.1	45.5	34.6	12.2	18.0
WODAN2.0+NEG (R)	47.4	46.4	46.9	38.5	45.4	41.7	19.2	10.2	13.3
Heritage Quest (R)	45.3	80.5	57.9	75.7	85.0	80.1	38.5	55.6	45.5

Table 6. Heatmap showing the performance (F1) of WODAN2.0 on barrows as a result of using a confidence threshold (80-91) and a threshold of the number of detections per grid cell (between 1 and ≥10); also see Section 3.2.

	80	81	82	83	84	85	86	87	88	89	90	91
1	24.0	25.4	26.9	27.9	29.2	31.1	32.1	33.0	33.7	35.3	36.8	39.1
2	31.6	34.0	34.6	36.1	38.0	40.0	41.3	41.1	41.8	43.2	42.5	42.8
3	35.9	39.6	40.7	41.5	43.8	44.3	43.9	44.2	45.2	45.3	45.3	47.2
4	41.6	42.1	42.6	43.0	44.4	43.3	44.8	45.5	48.1	49.8	45.7	44.3
5	42.5	44.4	44.3	44.3	44.4	44.7	45.0	45.3	46.6	46.2	42.3	40.0
6	43.2	44.8	45.1	44.4	44.8	42.7	45.1	46.0	45.1	42.6	40.6	37.8
7	44.3	44.6	43.8	41.7	42.2	43.9	43.9	42.7	40.4	38.5	36.7	32.7
8	46.0	45.1	43.1	40.9	41.1	42.5	42.6	42.5	38.7	36.9	28.6	27.9
9	43.1	40.3	39.6	40.8	41.6	41.6	40.4	38.5	34.5	32.7	25.8	22.6
10	40.7	39.6	41.0	39.6	40.2	38.7	37.8	34.1	30.6	27.8	24.7	16.9

Table 7. The performance of WODAN2.0 and Heritage Quest on the test datasets (non-random, NR; random, R) without using thresholds.

Method	Metric	Barrows	Celtic Fields	Charcoal Kilns
WODAN2.0 (NR)	recall	80.5	92.8	–
	precision	23.3	40.8	–
	F1-score	36.2	56.7	–
WODAN2.0 (R)	recall	79.6	82.9	38.5
	precision	14.1	13.3	5.1
	F1-score	24.0	22.9	8.9
Heritage Quest (R)	recall	82.5	89.6	76.9
	precision	8.1	43.4	2.6
	F1-Score	14.8	58.8	5.0

5.2. Results

Table 5 shows the performance of WODAN2.0 on both the non-random (NR) and the random (R) test datasets. WODAN2.0 has a performance (F1) of circa 70% for barrows and Celtic fields using thresholds on the non-random test dataset. The performance (F1) on the random test dataset is lower: circa 50% for barrows, circa 46% for Celtic fields, and circa 18% for charcoal kilns using thresholds. As shown in Table 5, the use of the training and validation datasets with negative subtiles (indicated with NEG in Table 5; see Section 3.3) did not improve the performance of WODAN2.0. The exact reason of this is unknown and will be further investigated in future research.

A cursory analysis of the false positives produced by WODAN2.0 shows that these include a wide range of anthropogenic and natural landscape objects that generally have a comparable geometric shape as the archaeological objects (Figure 6). No significant pattern in the nature or location of these 'objects of confusion' can be observed, because the more common patterns (e.g., drift-sand dunes, roundabouts, etc.) are already excluded from the results with LBR.

The performance of WODAN1.0 is displayed in Table 5. Comparing the performance (F1) of WODAN1.0 and WODAN2.0 shows that WODAN2.0 outperforms its predecessor WODAN1.0 on the detection of barrows and Celtic fields in both the non-random and random test datasets. The comparison also shows the considerable impact of bootstrap aggregating and LBR—the main differences between the WODAN1.0 and WODAN2.0 workflow, which led to an improvement in performance (F1) varying between 17% and 35%. A large increase in precision and a small decrease in recall can be observed, mostly due to the discarding of false positives and true positives in low ranking areas of the Location-Based Ranking map (see Figure 5).

The results of testing Heritage Quest (see Table 5) on the random test dataset show a performance (F1) of circa 58% for barrows, 80% for Celtic fields, and circa 46% for charcoal kilns using thresholds. When comparing the performance of WODAN2.0 and Heritage Quest, it can be observed that the citizen researchers outperform WODAN2.0 by a margin of circa 8% for barrows, while the margin for Celtic fields and charcoal kilns is higher (34% and 27%, respectively). However, if the performance of WODAN2.0 and Heritage Quest without using thresholds is compared (Table 7), the recall differs little for barrows (circa 3%) and Celtic fields (circa 7%), but varies greatly for charcoal kilns (circa 37%).

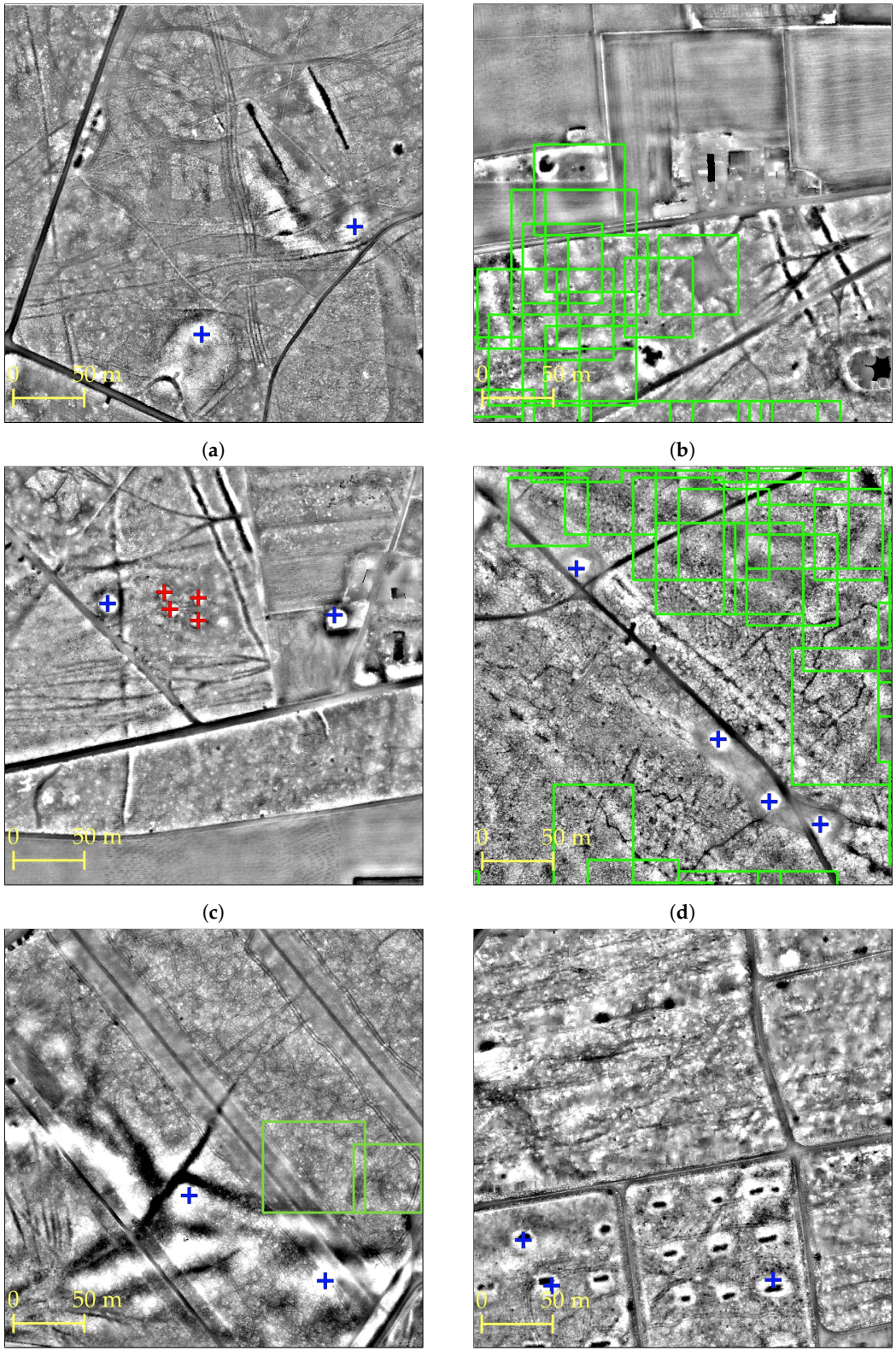


Figure 6. Excerpts of LiDAR data, visualized with Local Relief Model [40], showing correct (**a**–**d**) and incorrect (**e**,**f**) detections of barrows (blue), Celtic fields (green), and charcoal kilns (red) by WODAN2.0 (source of the height model: Reference [39]).

6. Discussion

The results of the experimental evaluation (Table 5) show that WODAN2.0 is able to detect multiple archaeological object classes in LiDAR data (also see Figure 6). The strategies incorporated in WODAN2.0 improve the performance of the object detection workflow compared to is predecessor WODAN1.0. However, the performance of WODAN2.0 still has room for improvement. The performance of WODAN2.0 on charcoal kilns can be considered low and is probably related to the diversity in the shape of charcoal kilns in the research area and an insufficient number of examples (see Table 1) in the training dataset (see [22]). The performance of WODAN2.0 on barrows and Celtic fields can be considered high on the non-random test dataset, with performance (F1) of circa 70%, but is considerably lower on the random test dataset. The decrease in performance can in large part be attributed to the introduction of this random test dataset, as discussed below.

6.1. Non-Random versus Random Test Dataset

The influence of the different test datasets on the performance of WODAN1.0 and WODAN2.0 is clearly illustrated by comparing the performance of both on the non-random and random test datasets (Table 5). The main differences between the test datasets are the number of negative subtiles, the 'density' of archaeological objects, and the variety in the state of preservation of the archaeological objects. As stated above (see Section 2.3), the proportion of positive and negative subtiles (i.e., subtiles with or without archaeological objects) varies greatly between the non-random (6.7:1, positive:negative) and the random (1:4, positive:negative) test dataset. This increased amount of negative subtiles in the latter results in more false positives. Furthermore, the precision of object detection models strongly correlates to the total labeled objects in the research area (i.e., the density), as a result of the higher proportion of false positives:true positives detected in low-density areas as compared to high density areas [10]. In addition, identifying objects in low-density images is a challenging task, even for domain experts [10]. Our random test dataset—following the definition of density in Reference [10]—has a low density, while the non-random test dataset has a high density. Therefore, the random test dataset will have had a negative influence on the precision, due to the proportion of negative subtiles and change in density. On the other hand, the decrease in recall is probably caused by the increased variety in the state of preservation of the barrows and Celtic fields in the random test dataset as compared to the non-random test dataset. The former contains many more examples of archaeological objects in a bad state of preservation, which in general are harder to detect (also see [87]).

Although the introduction of the random test dataset leads to reduced performance of the object detection model, we discern that this test dataset better represents the real-world situation of archaeological prospection over different types of complex terrain, and therefore gives a better impression of the practical value of the object detection model.

6.2. Computer and Human Performance

Comparing the performance of WODAN2.0 and Heritage Quest shows that the former has not reached general human performance on the object detection task in the research area. Table 5 shows that the citizen researchers of Heritage Quest outperform WODAN2.0 on all archaeological classes. The main difference in performance is related to the precision (see Table 5). This might be due to the fact that the citizen researchers can more easily determine possible detections as being objects of confusion by consulting the two different LiDAR visualizations and by looking at the direct vicinity of the detection. The variation in performance on barrows is low, when comparing the performance using thresholds (Table 5) and without using thresholds (Table 7). The performance (with and without using thresholds) on Celtic fields and charcoal kilns varies more. The large difference in performance on the former might be related to the fact that the citizen researchers are looking for the telltale checkerboard pattern of Celtic fields, which has few parallels in the natural landscape (see [87]). Contrarily, the object detection model looks for the individual plots within a Celtic field, a shape much more abundant in

the landscape. The low performance of WODAN2.0 on charcoal kilns is most likely related to the problems mentioned above.

6.3. Object Detection Models Users

That WODAN2.0 and Heritage Quest have the potential to detect the majority of the archaeological objects in the random test datasets is shown by the recall in Table 7. However, without using thresholds, the number of false positives is high. This is a recurrent problem in object detection models [56]. Whether the precision values without using thresholds are acceptable depends on the perspective of the envisioned user of the object detection model. A field archaeologist can, due to financial and time constraints, only investigate a limited number of detected, potential archaeological objects, and would need an object detection model in which high precision is essential. Hence, every false positive investigated reduces the amount of archaeological information gained during the field campaign. On the other hand, for cultural heritage managers, high recall is more important as localizing as many of the archaeological objects as possible is paramount for appropriate conservation. Failing to localize an archaeological object can lead to inadequate protection, potential damage, and ultimately the destruction of the archaeological object.

Either way, the results in Table 7 show that the focus of further research should lie on reducing the number of false positives in order to improve precision. The implementation of Location-Based Ranking is a first attempt to specifically combat this problem. Furthermore, this method also offers opportunities to make informed decisions regarding the allocation of (limited) resources for (field) validation, for instance by targeting archaeological objects with the highest potential or by drawing a relevant sample from all different ranks for (field) validation. Depending on the characteristics chosen, Location-Based Ranking can also be used to redirect resources and prioritize the validation of objects threatened by human activity, such as urbanization and agricultural practices.

7. Conclusions

This paper presents the results of the implementation of a Region-based Convolutional Neural Network (Faster R-CNN [25]) in a workflow, called WODAN2.0, for the automated detection of archaeological objects in LiDAR data. WODAN2.0 is the updated version of WODAN1.0 [22] and incorporates several strategies to improve performance, including reduced anchor box sizes and bootstrap aggregating. To reduce the number of false positives caused by specific regions, a novel approach called Location-Based Ranking has been developed and implemented into the workflow.

To investigate the practical value of WODAN2.0 for large-scale archaeological prospection over different types of complex terrain, a large, random test dataset was developed, replacing the original small, non-random dataset. To evaluate the performance, as compared to humans, the results of the citizen science project Heritage Quest [9] were used as a benchmark for general human performance on the task of archaeological object detection in the research area.

The results of the experimental evaluation (Table 5) on the non-random and random test dataset show that WODAN2.0 outperforms its predecessor WODAN1.0. However, the performance of WODAN2.0 does not reach or exceed general human performance. While the recall (see Table 7) without using thresholds is high, the object detection model has low precision. To make WODAN2.0 feasible for large-scale archaeological prospection, future research will therefore focus on improving the precision of the workflow. A possible improvement lies in the use of CNNs pre-trained on remotely sensed data (see [19]), for instance using the BigEarthNet archive [88]. Furthermore, the recent initiatives to combine Deep Learning methods and citizen science in biology [89], environmental sciences [90], and even archaeology [9] are promising as well, and future research will also focus on different means to combine these two methods, for instance by incorporating the results of Heritage Quest in the training dataset [9], or by developing a task allocation strategy [90].

In the end, the goal of this research is not to develop a method to either outperform or replace archaeological experts or 'automate archaeology' [91]. Rather, object detection models are meant

to become another instrument in the archaeologists' toolkit that assists in the rapid and systematic mapping of objects of interest over extensive areas, in large and complex datasets [10]. The subsequent archaeological interpretation remains the domain of the human expert. The utilization of object detection models and post-processing steps such as Location-Based Ranking are in essence about reducing the time invested into mapping archaeological objects. By tending to the task of localizing, the specialist's time can be reallocated to analysis, (field) validation, and interpretation of the results.

Author Contributions: Conceptualization, Wouter B. Verschoof-van der Vaart, Karsten Lambers, and Quentin P.J. Bourgeois; methodology, Wouter B. Verschoof-van der Vaart; software, Wouter B. Verschoof-van der Vaart and Wojtek Kowalczyk; validation, Wouter B. Verschoof-van der Vaart and Quentin P.J. Bourgeois; formal analysis, Wouter B. Verschoof-van der Vaart; investigation, Wouter B. Verschoof-van der Vaart; resources, Wouter B. Verschoof-van der Vaart, Karsten Lambers and Quentin P.J. Bourgeois; data curation, Wouter B. Verschoof-van der Vaart; writing—original draft preparation, Wouter B. Verschoof-van der Vaart; writing—review and editing, Wouter B. Verschoof-van der Vaart, Karsten Lambers, Wojtek Kowalczyk, and Quentin P.J. Bourgeois; visualization, Wouter B. Verschoof-van der Vaart; supervision, Karsten Lambers and Wojtek Kowalczyk; project administration, Karsten Lambers; and funding acquisition, Karsten Lambers and Quentin P.J. Bourgeois. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Data Science Research Programme of Leiden University, the Netherlands. The Heritage Quest project was funded by *Erfgoed Gelderland* and the Province of Gelderland.

Acknowledgments: We are grateful to Eva Kaptijn and Roel Kramer at *Erfgoed Gelderland*, and Sigrid van Roode at *Provincie Gelderland* for their collaboration in developing and implementing Heritage Quest. We would like to thank the reviewers for their valuable comments and insights.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cowley, D.C. In with the new, out with the old? Auto-extraction for remote sensing archaeology. *Proc. SPIE* **2012**, *8532*, 853206. [CrossRef]
2. Bennett, R.; Cowley, D.; De Laet, V. The data explosion: Tackling the taboo of automatic feature recognition in airborne survey data. *Antiquity* **2014**, *88*, 896–905. [CrossRef]
3. Bevan, A. The data deluge. *Antiquity* **2015**, *89*, 1473–1484. [CrossRef]
4. Devereux, B.J.; Amable, G.S.; Crow, P.; Cliff, A.D. The potential of airborne lidar for detection of archaeological features under woodland canopies. *Antiquity* **2005**, *79*, 648–660. [CrossRef]
5. Crutchley, S.; Crow, P. *Using Airborne Lidar in Archaeological Survey: The Light Fantastic*, 2nd ed.; Historic England: Swindon, England, 2018.
6. Opitz, R.S. An overview of airborne and terrestial laser scanning in archaeology. In *Interpreting Archaeological Topography: Airborne Laser Scanning, 3D Data and Ground Observation*; Opitz, R.S., Cowley, D.C., Eds.; Oxbow Books: Oxford, UK; Oakville, ON, Canada, 2013; Chapter 2, pp. 13–31.
7. Opitz, R.; Herrmann, J. Recent trends and long-standing problems in archaeological remote sensing. *J. Comput. Appl. Archaeol.* **2018**, *1*, 19–41. [CrossRef]
8. Hay, G.J.; Castilla, G. Geographic object based image analysis (GEOBIA): A new name for a new discipline. In *Object Based Image Analysis*; Blaschke, T., Lang, S., Hay, G., Eds.; Springer: Heidelberg, Germany, 2008; pp. 93–112.
9. Lambers, K.; Verschoof-van der Vaart, W.B.; Bourgeois, Q.P. Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection. *Remote. Sens.* **2019**, *11*, 794. [CrossRef]
10. Soroush, M.; Mehrtash, A.; Khazraee, E.; Ur, J.A. Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq. *Remote Sens.* **2020**, *12*, 500. [CrossRef]
11. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
12. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
14. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine Learning for Cultural Heritage: A Survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. [CrossRef]
15. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *J. Appl. Remote. Sens.* **2017**, *11*, 042609. [CrossRef]

16. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshop, Columbus, OH, USA, 23–28 June 2014; pp. 806–813. [CrossRef]
17. Zingman, I. Semi-Automated Detection of Fragmented Rectangular Structures in High Resolution Remote Sensing Images with Application in Archaeology. Ph.D. Thesis, University of Konstanz, Konstanz, Germany, 2016.
18. Zingman, I.; Saupe, D.; Penatti, O.A.B.; Lambers, K. Detection of fragmented rectangular enclosures in very-high-resolution remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* **2016**, *54*, 4580–4593. [CrossRef]
19. Gallwey, J.; Eyre, M.; Tonkins, M.; Coggan, J. Bringing Lunar LiDAR Back Down to Earth: Mapping Our Industrial Heritage through Deep Transfer Learning. *Remote. Sens.* **2019**, *11*, 1994. [CrossRef]
20. Trier, Ø.D.; Salberg, A.B.; Pilø, L.H. Semi automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In *CAA 2016: Oceans of Data, Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology*; Matsumoto, M., Uleberg, E., Eds.; Archaeopress: Oxford, UK, 2018; pp. 219–231.
21. 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. [CrossRef]
22. Verschoof-van der Vaart, W.B.; Lambers, K. Learning to look at LiDAR: The use of R-CNN in the automated detection of archaeological objects in LiDAR data from the Netherlands. *J. Comput. Appl. Archaeol.* **2019**, *2*, 31–40. [CrossRef]
23. David, A. The Role and Practice of Archaeological Prospection. In *Handbook of Archaeological Sciences*; Brothwell, D., Pollard, A., Eds.; John Wiley & Sons, LTD: Chichester, UK, 2005; pp. 521–527.
24. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE CVPR, Columbus, OH, USA, 24–27 June 2014; pp. 580–587. [CrossRef]
25. 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.* **2017**, *39*, 1137–1149. [CrossRef]
26. Bourgeois, Q.P.J. *Monuments on the Horizon. The Formation of the Barrow Landscape throughout the 3rd and 2nd Millennium BC*; Sidestone Press: Leiden, The Netherlands, 2013.
27. Arnoldussen, S. The fields that outlived the Celts: The use-histories of later prehistoric field systems (Celtic Fields or Raatakkers) in The Netherlands. *Proc. Prehist. Soc.* **2018**, *84*, 303–327. [CrossRef]
28. Groenewoudt, B. Charcoal Burning and Landscape Dynamics in the Early Medieval Netherlands. *Ruralia* **2007**, *6*, 327–337.
29. Van Lanen, R.J.; Groenewoudt, B.J.; Spek, M.; Jansma, E. Route persistence. Modelling and quantifying historical route-network stability from the Roman period to early-modern times (AD 100–1600): A case study from the Netherlands. *Archaeol. Anthropol. Sci.* **2018**, *10*, 1037–1052. [CrossRef]
30. Vletter, W.F.; Van Lanen, R.J. Finding vanished routes: Applying a multi-modelling approach on lost route and path networks in the Veluwe Region, The Netherlands. *Rural. Landscapes Soc. Environ. Hist.* **2018**, *5*, 1–19. [CrossRef]
31. Van der Schriek, M.; Beex, W. The application of LiDAR-based DEMs on WWII conflict sites in the Netherlands. *J. Confl. Archaeol.* **2017**, *12*, 94–114. [CrossRef]
32. Lohof, E. Tradition and change: Burial practices in the Late Neolithic and Bronze Age in the north-eastern Netherlands. *Archaeol. Dialogues* **1994**, *1*, 98–118. [CrossRef]
33. Louwen, A.; Fontijn, D.R. *Death Revisited. The Excavation of Three Bronze Age Barrows and Surrounding Landscape at Apeldoorn-Wieselseweg*; Sidestone Press: Leiden, The Netherlands, 2019.
34. Bourgeois, Q.P.J.; Fontijn, D.R. The Tempo of Bronze Age Barrow Use: Modeling the Ebb and Flow in Monumental Funerary Landscapes. *Radiocarbon* **2015**, *57*, 47–64. [CrossRef]
35. Kooistra, M.; Maas, G. The widespread occurrence of Celtic field systems in the central part of the Netherlands. *J. Archaeol. Sci.* **2008**, *35*, 2318–2328. [CrossRef]

36. Raab, A.; Takla, M.; Raab, T.; Nicolay, A.; Schneider, A.; Rösler, H.; Heußner, K.U.; Bönisch, E. Pre-industrial charcoal production in Lower Lusatia (Brandenburg, Germany): Detection and evaluation of a large charcoal-burning field by combining archaeological studies, GIS-based analyses of shaded-relief maps and dendrochronological age determination. *Quat. Int.* **2015**, *367*, 111–122. [CrossRef]
37. Deforce, K.; Boeren, I.; Adriaenssens, S.; Bastiaens, J.; De Keersmaeker, L.; Haneca, K.; Tys, D.; Vandekerkhove, K. Selective woodland exploitation for charcoal production. A detailed analysis of charcoal kiln remains (ca. 1300–1900 AD) from Zoersel (northern Belgium). *J. Archaeol. Sci.* **2013**, *40*, 681–689. [CrossRef]
38. Kenzler, H.; Lambers, K. Challenges and perspectives of woodland archaeology across Europe. In *Concepts, Methods and Tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*; Giligny, F., Djindjian, F., Costa, L., Moscati, P., Robert, S., Eds.; Archaeopress: Oxford, UK, 2015; pp. 73–80.
39. Publieke Dienstverlening Op de Kaart (PDOK). Available online: https://www.pdok.nl/ (accessed on 18 March 2020).
40. Hesse, R. LiDAR-derived Local Relief Models—A new tool for archaeological prospection. *Archaeol. Prospect.* **2010**, *17*, 67–72. [CrossRef]
41. Actueel Hoogtebestand Nederland (AHN). Available online: https://ahn.arcgisonline.nl/ahnviewer/ (accessed on 18 March 2020).
42. Van Der Zon, N. *Kwaliteitsdocument AHN2*; Technical Report; Rijkswaterstaat: Amersfoort, The Netherlands, 2013.
43. QGIS Development Team. QGIS Geographic Information System. Available online: https://www.qgis.org/ (accessed on 27 April 2020).
44. Kokalj, Ž.; Hesse, R. *Airborne Laser Scanning Raster Data Visualisation: A Guide to Good Practice*; Založba ZRC: Ljubljana, Slovenian, 2017.
45. Nyffeler, J. *Kulturlandschaft in neuem Licht: Eine Einführung zu LiDAR in der Archäologie*; University of Bamberg Press: Bamberg, Germany, 2018.
46. LabelImg. Available online: https://github.com/tzutalin/labelImg/ (accessed on 8 April 2020).
47. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2009.
48. Sadr, K. The impact of coder reliability on reconstructing archaeological settlement patterns from satellite imagery: A case study from South Africa. *Archaeol. Prospect.* **2016**, *23*, 45–54. [CrossRef]
49. Heritage Quest. Available online: https://www.zooniverse.org/projects/evakap/heritage-quest (accessed on 18 March 2020).
50. Eitzel, M.V.; Cappadonna, J.L.; Santos-Lang, C.; Duerr, R.E.; Virapongse, A.; West, S.E.; Kyba, C.C.M.; Bowser, A.; Cooper, C.B.; Sforzi, A.; et al. Citizen science terminology matters: Exploring key terms. *Citiz. Sci. Theory Pract.* **2017**, *2*, 1–20. [CrossRef]
51. Van den Dries, M.H. Community Archaeology in the Netherlands. *J. Community Archaeol. Heritage* **2014**, *1*, 68–88. [CrossRef]
52. The Zooniverse. Available online: https://www.zooniverse.org (accessed on 25 March 2020).
53. Lyman, R.L.; VanPool, T.L. Metric data in archaeology: A study of intra-analyst and inter-analyst variation. *Am. Antiq.* **2009**, *74*, 485–504. [CrossRef]
54. Kosmala, M.; Wiggins, A.; Swanson, A.; Simmons, B. Assessing data quality in citizen science. *Front. Ecol. Environ.* **2016**, *14*, 551–560. [CrossRef]
55. Swanson, A.; Kosmala, M.; Lintott, C.; Packer, C. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conserv. Biol.* **2016**, *30*, 520–531. [CrossRef] [PubMed]
56. Casana, J. Global-Scale Archaeological Prospection using CORONA Satellite Imagery: Automated, Crowd-Sourced, and Expert-led Approaches. *J. Field Archaeol.* **2020**, *45*, S89–S100. [CrossRef]
57. Freitag, A.; Meyer, R.; Whiteman, L. Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data. *Citiz. Sci. Theory Practice.* **2016**, *1*, 2. [CrossRef]
58. Herfort, B.; Höfle, B.; Klonner, C. 3D micro-mapping: Towards assessing the quality of crowdsourcing to support 3D point cloud analysis. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *137*, 73–83. [CrossRef]
59. Salk, C.; Sturn, T.; See, L.; Fritz, S. Local knowledge and professional background have a minimal impact on volunteer citizen science performance in a land-cover classification task. *Remote. Sens.* **2016**, *8*, 774. [CrossRef]

60. Python 3.6. Available online: https://www.python.org/ (accessed on 8 April 2020).
61. Keras. Available online: https://keras.io/ (accessed on 8 April 2020).
62. Keras Implementation of Faster R-CNN. Available online: https://github.com/moyiliyi/keras-faster-rcnn (accessed on 8 April 2020).
63. Uijlings, J.R.R.; Van De Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
64. Girshick, R. Fast R-CNN. In Proceedings of the IEEE ICCV, Santiago, Chile, 7–13 December 2015. [CrossRef]
65. Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining. *Sensors* **2017**, *17*, 336. [CrossRef]
66. Mohamed, E.; Sirlantzis, K.; Howells, G. Application of transfer learning for object detection on manually collected data. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 1037, pp. 919–931. [CrossRef]
67. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
68. Ding, P.; Zhang, Y.; Deng, W.J.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *141*, 208–218. [CrossRef]
69. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for Small Object Detection. In *Computer Vision—ACCV 2016. Lecture Notes in Computer Science, vol 10115*; Lai, S., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 214–230. [CrossRef]
70. Ren, Y.; Zhu, C.; Xiao, S. Small object detection in optical remote sensing images via modified Faster R-CNN. *Appl. Sci.* **2018**, *8*, 813. [CrossRef]
71. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
72. Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, NY, USA, 1993.
73. Gupta, N.; Devillers, R. Geographic Visualization in Archaeology. *J. Archaeol. Method Theory* **2017**, *24*, 852–885. [CrossRef]
74. Esri. ArcMap. Available online: https://desktop.arcgis.com/en/arcmap/ (accessed on 27 April 2020).
75. Gao, L.; He, Y.; Sun, X.; Jia, X.; Zhang, B. Incorporating negative sample training for ship detection based on deep learning. *Sensors* **2019**, *19*, 684. [CrossRef]
76. Verhagen, P.; Whitley, T.G. Integrating Archaeological Theory and Predictive Modeling: A Live Report from the Scene. *J. Archaeol. Method Theory* **2012**, *19*, 49–100. [CrossRef]
77. Verhagen, P.; Whitley, T.G. Predictive spatial modelling. In *Archaeological Spatial Analysis: A Methodological Guide*; Gillings, M., Hacigüzeller, P., Lock, G., Eds.; Routledge: Oxon, UK; New York, NY, USA, 2020; Chapter 13, pp. 231–246.
78. Casarotto, A.; Stek, T.D.; Pelgrom, J.; Otterloo, R.H.v.; Sevink, J. Assessing visibility and geomorphological biases in regional field surveys: The case of Roman Aesernia. *Geoarchaeology* **2018**, *33*, 177–192. [CrossRef]
79. Gerritsen, F. *Local Identities: Landscape and Community in the Late Prehistoric Meuse-Demer-Scheldt Region*; Amsterdam University Press: Amsterdam, The Netherlands, 2003. [CrossRef]
80. Cowley, D. Remote sensing for archaeology and heritage management-site discovery, interpretation and registration. In *Remote Sensing for Archaeological Heritage Management, Proceedings of the 11th EAC Heritage Management Symposium, Reykjavîk, Iceland, 25–27 March 2010*; Cowley, D.C., Ed.; Europae Archaeologia Consilium: Brussel, Belgium, 2011; Chapter 4, pp. 43–55.
81. Koster, E.A. The "European Aeolian Sand Belt": Geoconservation of Drift Sand Landscapes. *Geoheritage* **2009**, *1*, 93–110. [CrossRef]
82. Blume, H.P.; Leinweber, P. Plaggen soils: Landscape history, properties, and classification. *J. Plant Nutr. Soil Sci.* **2004**, *167*, 319–327. [CrossRef]
83. Spek, T. *Het Drentse Esdorpenlandschap. Een Historisch-Geografische Studie*; Matrijs: Utrecht, The Netherlands, 2004.
84. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
85. Sammut, C.; Webb, G.I. *Encyclopaedia of Machine Learning*; Springer: Boston, MA, USA, 2010. [CrossRef]

86. Zou, Q.; Xie, S.; Lin, Z.; Wu, M.; Ju, Y. Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Res.* **2016**, *5*, 2–8. [CrossRef]
87. Risbøl, O.; Bollandsås, O.M.; Nesbakken, A.; Ørka, H.O.; Naesset, E.; Gobakken, T. Interpreting cultural remains in airborne laser scanning generated digital terrain models: Effects of size and shape on detection success rates. *J. Archaeol. Sci.* **2013**, *40*, 4688–4700. [CrossRef]
88. Sumbul, G.; Kang, J.; Kreuziger, T.; Marcelino, F.; Costa, H.; Benevides, P.; Caetano, M.; Demir, B. *BigEarthNet Deep Learning Models with A New Class-Nomenclature for Remote Sensing Image Understanding*; Technical Report; Technische Universitat Berlin: Berlin, Germany, 2020.
89. Torney, C.J.; Lloyd-Jones, D.J.; Chevallier, M.; Moyer, D.C.; Maliti, H.T.; Mwita, M.; Kohi, E.M.; Hopcraft, G.C. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods Ecol. Evol.* **2019**, *10*, 779–787. [CrossRef]
90. Herfort, B.; Li, H.; Fendrich, S.; Lautenbach, S.; Zipf, A. Mapping Human Settlements with Higher Accuracy and Less Volunteer Efforts by Combining Crowdsourcing and Deep Learning. *Remote. Sens.* **2019**, *11*, 1799. [CrossRef]
91. Traviglia, A.; Cowley, D.; Lambers, K. Finding Common Ground: Human and Computer Vision in Archaeological Prospection. *AARGnews—Newsl. Aerial Archaeol. Res. Group* **2016**, *53*, 11–24.