scientific reports

Check for updates

OPEN

A human–AI collaboration workflow for archaeological sites detection

Luca Casini[1], Nicolò Marchetti[2], Andrea Montanucci[1], Valentina Orrù[2] & Marco Roccetti[1]✉

This paper illustrates the results obtained by using pre-trained semantic segmentation deep learning models for the detection of archaeological sites within the Mesopotamian floodplains environment. The models were fine-tuned using openly available satellite imagery and vector shapes coming from a large corpus of annotations (i.e., surveyed sites). A randomized test showed that the best model reaches a detection accuracy in the neighborhood of 80%. Integrating domain expertise was crucial to define how to build the dataset and how to evaluate the predictions, since defining if a proposed mask counts as a prediction is very subjective. Furthermore, even an inaccurate prediction can be useful when put into context and interpreted by a trained archaeologist. Coming from these considerations we close the paper with a vision for a Human–AI collaboration workflow. Starting with an annotated dataset that is refined by the human expert we obtain a model whose predictions can either be combined to create a heatmap, to be overlaid on satellite and/or aerial imagery, or alternatively can be vectorized to make further analysis in a GIS software easier and automatic. In turn, the archaeologists can analyze the predictions, organize their onsite surveys, and refine the dataset with new, corrected, annotations.

This paper documents the outcomes of a collaboration between data scientists and archaeologists with the goal of creating an artificial intelligence (AI) system capable of assisting in the task of detecting potential archaeological sites from aerial or, in our case, satellite imagery. Using semantic segmentation models allowed us to draw precise outlines and human-in-the-loop evaluation showed that detection accuracy is in the neighborhood of 80%.

This procedure falls into the domain of Remote Sensing (RS) which indicates the act of detecting and/or monitoring a point of interest from a distance. In the world of archaeology this operation has become invaluable with the availability of more and better imagery from satellites that can be combined with older sources of information (e.g., the CORONA satellite imagery) to spot a larger number of archaeological sites as well as tracking their successive degradation due to anthropic factors[1]. Depending on the area of investigation and the size of the archaeological features being surveyed, the effort necessary, especially in terms of time, can be huge for the researcher.

This collaboration aimed at solving exactly this issue by using deep learning models to streamline, but not completely automate, the process. Thus, starting from a dataset of vector shapes for all archaeologically recorded sites in the southern Mesopotamian floodplain (which represents a sufficiently coherent geo-morphological region), we trained a model to detect and segment sites in a given input image. As the project went on, a number of issues emerged that make this problem particularly hard to tackle and lead to an important reflection on the use of deep learning in general and its relationship to human experts. The dataset, while may be considered a very large one for near eastern archaeology with its almost 5000 sites, is hardly sufficient for training a model as large as the state-of-the-art ones we see in use today and, perhaps more significantly, contains many cases that are visible only on certain old imagery. The first issue is commonly solved through transfer learning[2]. This technique consists in starting from a model, pre-trained on a large and general dataset (e.g. imagenet[3]), and then in fine-tuning it on a smaller but more specific dataset, leveraging the skills it has previously learned to make the new task more manageable. The second one, however, puts both training and evaluation in jeopardy, as the model is pushed to make wrong classifications during training and even if it learned robust representations that ignore bad examples, we would then have a hard time detecting if is a mistake is by the model or in the labels.

We believe that the only way out of this conundrum is through a human-in-the-loop approach[1]. For this reason, throughout the paper we highlight the importance of integrating domain expertise during the training

[1]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy. [2]Department of History and Cultures, University of Bologna, Bologna, Italy. ✉email: marco.roccetti@unibo.it

and evaluation phase of our experiments, since that was crucial in improving the dataset used and, in turn, the model. The final outcome of this iterative process is a model capable of a detection accuracy of around 80%.

Based on these promising results, we envision a tool for human-AI collaboration to support the archaeologists in the remote sensing operations (rather than replace them) and propose a new kind of workflow, enhancing both their task and the model by providing improved data after every use[4,5]. All the results were achieved using open-source software and models, as well as openly available data (imagery, annotations) and computational resources (Google Colab), making this kind of work highly accessible and replicable even in resource-constrained research environments. All code, data and resources mentioned are available on GitHub (https://bit.ly/NSR_floodplains).

Research background

The Mesopotamian floodplain. The southern Mesopotamian floodplain is a crucial region for understanding the complex interplay between the spatial clustering of human communities and the development of irrigated farmland in an otherwise semi-arid environment[6]. Robert McCormick Adams' surveys in the area[7–9] were carried out according to standards that were unparalleled for the time: he used a set of aerial photographs from 1961 to locate potential sites and map canals whose traces were visible on the surface; he was systematic in recording sites ranging in time from the later 7th millennium BCE to the Ottoman period; above all, he was acutely aware of the historiographical potential of his survey work, which resulted in a powerful interpretation of settlement patterns and hydraulic activities[8].

After a long halt to fieldwork resulting from political instability, archaeological research resumed in southern Iraq in recent years, see[10] for an overview. In this area sites are usually referred with the Arabic word for mound, "Tell". The color and shape of these hills makes them especially visible from aerial and satellite imagery, which led to the use of remote sensing as a viable strategy to discover their location.

As Tony Wilkinson puts it "Tells comprise multiple layers of building levels and accumulated wastes built up through time, in part because the locus of occupation has remained stationary. Tell settlements frequently are defined by an outer wall that both contained and constrained the accumulated materials, thereby restricting their spread […]. The tell is by no means the sale locus of occupation […]. Outer or lower towns […] often appear as low humps or simply artifact scatters around tells, and they can extend the total occupied area of a site several fold"[11].

In Mesopotamia, Tells are often only slightly more elevated than the surrounding countryside, often being prone in such cases to artificial leveling in order to gain irrigable agricultural areas. Thus, the automatic detection of sites in such a dynamic environment is a highly complex operation, although contrasts are sufficiently marked to justify the attempt.

Remote sensing. By remote sensing one may refer to the use of any sensor (i.e., temperature, humidity, hyper-spectral, satellite images etc.) for detecting or monitoring a point of interest without the need for direct observation. This approach is relevant to a variety of fields, but solutions that work in one domain may not translate to others.

Locating archaeological sites remotely was certainly possible even before the advent of modern computer technology by using aerial photographs and topographical maps of the area to be investigated, but today it is easier to combine multiple sources, using sensors of different nature or from different points in time, to get a more complete picture of the environment, especially since it can be changing due to natural or anthropic factors[12–14]. Depending on the characteristics of the sites, certain representations can be helpful like elevation models obtained from stereoscopic images or the use of parts of the electromagnetic spectrum other than visible light like infrared or radio waves[15,16]. Light Detection and Ranging (LiDAR) is also becoming popular as it gives satisfactory high-resolution images, but it can be difficult to employ as it often requires to be mounted on some kind of airborne craft like drones[17]. The problem with these types of sources is that they might not be available for every location or not have a high enough resolution for the task at hand. On the other hand, good quality, open-source RGB images of virtually any location on the planet are readily available, especially through the popularity of online services such as Google Maps or Bing Maps. Specifically, in this project, we use satellite imagery from the Bing Maps service, which, for the area under analysis, provides excellent visibility of the anthropogenic traces we focus on: Tells.

Deep learning for remote sensing and archaeology. Deep learning has found multiple uses in every field of application and archaeology is no exception. It can help in classifying objects and text, finding similarities, building 3D models and, as this paper illustrates too, the detection of sites[18–22]. A difficulty in dealing with such a model is that it requires domain experts in both archaeology and deep learning to come together, but it may also depend on the amount of data available. Neural networks are notoriously data hungry, and archaeology is a "slow data" field as Bickler put it[23]. Nonetheless, there are a few recent examples of deep learning being successfully applied to site detection in a variety of different scenarios[24–27]. Most applications either use neural network to perform a classification, detection or segmentation task. The first uses tiles sampled from maps that are marked as containing the site of interest or not; the second instead consists in predicting a bounding box around an object and classifying it if needed; in the third the individual pixels are classified, and the result is the prediction of a shape corresponding to the site. In this paper we use the second approach, as described below.

Semantic segmentation. Semantic segmentation is the task of dividing an image into parts that correspond to units with a specific meaning. These can correspond to a specific subject (e.g., the outline of persons, vehicles, etc.) or to a generic category that encompasses multiple entities (e.g., buildings, backgrounds, etc.). In the context of this paper, we only have two categories: one for mounded (tell) sites and another one for every-

thing else. Segmentation can be performed with various techniques that perform pixel-level classification. A very common approach uses pre-computed features, extracted by some algorithm, or manually engineered, which are then classified by a Random Forest algorithm[28]. The current state of the art is represented by end-to-end systems based on deep learning with convolutional neural networks. For this approach, the introduction of U-Net by Ronnenberger in the context of medical imaging represented a milestone[29]. This work leverages a more recent architecture, called MA-Net[30], which can be thought of as an upgrade of the U-Net architecture with the inclusion of a self-attention mechanism as proposed in the popular Transformer architectures[31]. This allows the model to weigh different latent features depending on contents, figuratively specifying where to "pay attention" in this latent space in order to learn better. While it was developed in the context of medical imaging, it has found use also in remote sensing tasks[32,33]. In the "Materials and methods" section below we provide more details.

Previous work and limitations. In a previous paper we tried to tackle this same problem using an image classification approach where the map was divided into tiles[34]. In that experiment, however, the dataset was an order of magnitude smaller, and we had to resort to aggressive data augmentation in order to boost performance. The best model obtained an AUC score of around 70% but when tested on an unseen portion of map it showed its limits in that it predicted many False Positives while also missing some sites. The biggest trade-off of this tile-based classification approach is between the size of the tiles and the granularity of the predictions with bigger squares that are more practical but result in a loss of detail. There is also the problem of dealing with sites that land on the edge of a tile. A solution we tried was creating a shingled dataset with in-between tiles to fill the gaps. This however greatly increased the amount of prediction to be created. Finally, most models for image classification are bound by the use of a fixed size of input which can be a huge limit when dealing with maps. In this new experiment, given the increased size of the dataset, we decided to leverage image segmentation models with fully convolutional layers which address both the limits in input size and the granularity trade-off.

Materials and methods

In this section we first describe the dataset used, which was built starting from openly available resources and then the open-source models we fine-tuned on that dataset.

Vector shapes for archaeological sites. We started with a dataset of geo-referenced vector shapes corresponding to contours of known mound sites in the survey area of the Floodplains Project that spans 66,000 km^2, as shown in Fig. 1. The dataset—developed at the University of Bologna by filing all published archaeological surveys in the area and geo-referencing anew the sites cataloged therein (https://floodplains.orientlab.net)—contains 4934 shapes, thus all referring to sites which had been confirmed by ground truthing and by the associated study of the surface scatter of artifacts.

Since the dataset was compiled as a comprehensive source of information for archaeologists rather than specifically to train a deep learning model, we needed to filter out some examples that provided no information and could actually impair the learning process. We started by removing the top 200 sites by area as these were considerably bigger than the rest of the dataset and visual inspection confirmed that they follow the shape of areas that are not just simply mounds. The number 200 emerges from noticing that these sites have an area bigger than the square region we use as an input and could thus result in a completely full segmentation mask which would not be very helpful. After a discussion between data scientists and archaeologists we convened that this was a good heuristic solution.

Additionally, we filtered out 684 sites that either presented an area too small to be a Tell or were earmarked by the archaeologists as having been destroyed. In particular, the size threshold was set at around 1000 m^2 which corresponds to a circle with a diameter of 30 m. These very small sites actually correspond to a generic annotation for known sites with unknown size or precise location.

Setting the input images. To generate a set of images to fine-tune our pre-trained model, we imported the abovementioned shapes into QGIS, an open-source GIS software[35] and using a Python script saved a square of length L centered on the centroid of the site that contains only satellite imagery from Bing Maps (displayed directly in the GIS environment via the QuickMapService plugin that allows access to images provided by various online services, including Bing Maps). We then saved the same image without a base map but with the contours of the site represented as a shape filled with a solid color, to serve as ground truth masks.

Thus, during training, our neural network learns to reproduce the shape of the site from the ground truthed one by only looking at the RGB satellite image; during inference, we can detect and outline new sites in a given input image if there are any.

In the first experiments we set L to be 1000 m, but we imagined that increasing the size of the prediction area could be beneficial due to the inclusion of a larger context. Consequently, we also tried using L = 2000 m and obtained improved performance overall.

From the starting square image, we randomly crop a square of length L/2 to be used as the input. This ensures that the model does not learn a biased representation for which sites always appear at the center of the input and additionally serves as data augmentation. Beside this crop, we also augment the dataset by applying a random rotation and mirroring, as well as a slight shift in brightness and contrast, all these operations being applied in a different manner at each training iteration. When extracting from QGIS, we saved images with a resolution of around 1 pixel per meter (1024 pixels for 1000 m, double that for the model with increased input size) but the inputs were then scaled down to half of that to ease computational requirements while having low impact on the overall performance[36].

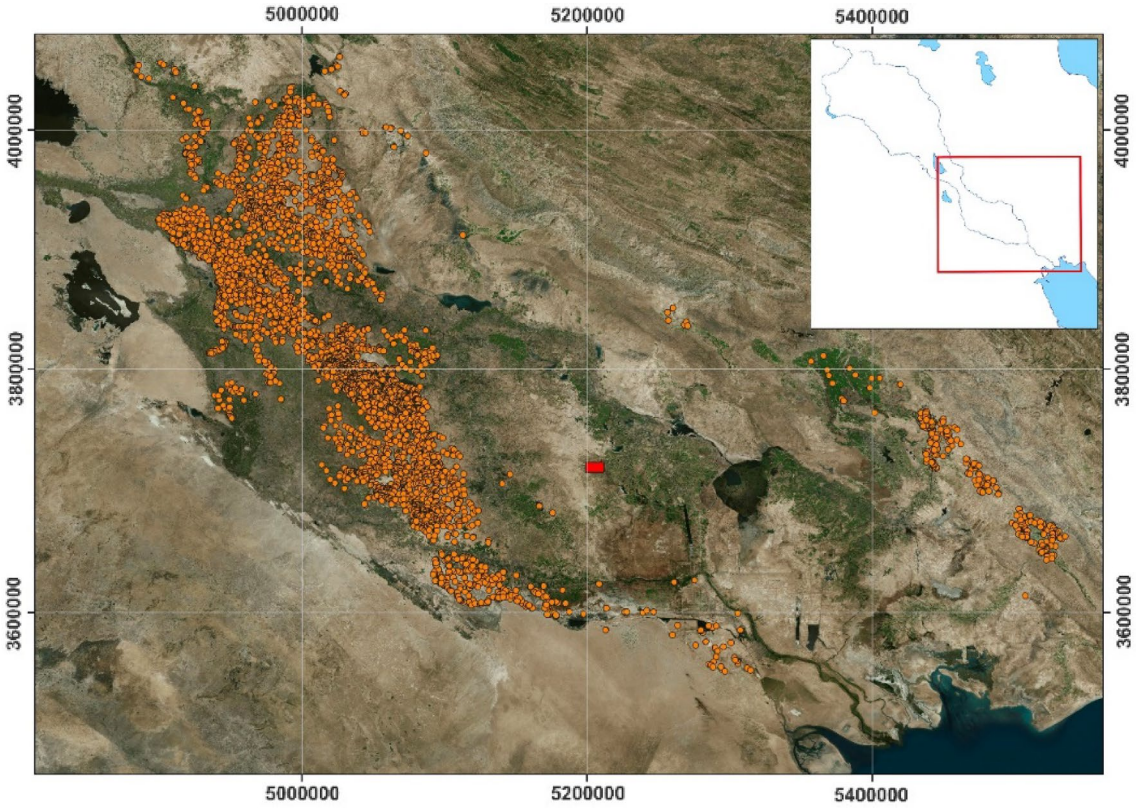

Figure 1. Investigation area. Orange dots represent surveyed sites in the Mesopotamian floodplain. The solid red rectangle is a selected test area in Maysan. All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://bit.ly/NSR_floodplains).

Finally, we introduced 1155 images with empty masks (no sites to predict) sampled from locations suggested by the archaeologists. These include highly urbanized areas, intensive agricultural areas, locations subject to flooding (i.e., artificial lakes and basins) and rocky hills and mountains.

The number was chosen arbitrarily, taking into consideration the size of each suggested area and of the tiles. The final number of images is thus 5025. We split the dataset into a 90% training set and a 10% holdout test set, stratifying the "empty" images we added. 10% of the training set was also randomly selected to be used as a validation set.

We tried integrating CORONA imagery as an additional input[37], as in the usual archaeological workflow that historical imagery is very useful (since it refers to a situation so much less affected by development) and often combines with the satellite basemaps and the topographical maps (but since CORONA were used here as a complement, we did not pursue automatic detection on them alone and thus sites destroyed after the 1970s have been excluded from the analysis). After importing the imagery into QGIS, we followed the same procedure to create the inputs, ensuring the crop operation was equal for both Bing and CORONA imagery.

Semantic segmentation models. This project started as an experiment to investigate the viability of pretrained semantic segmentation models as tools for detecting sites. For this reason, we decided to compare pretrained open-source models made available as part of a library written in PyTorch. The library allows one to choose an encoder convolutional neural network for feature extraction and a segmentation architecture independently, as well as providing a number of different loss functions[38].

In a previous preliminary paper, we experimented with different choices of architecture, encoders and loss functions[36]. We compared U-Net versus MA-net, Resnet18 versus Efficientnet-B3 and Dice Loss versus Focal Loss. The performance differences were small, within a few percentage points at best, which could be very well explained by fluctuations due to the random data augmentation.

Nonetheless, we took the best model which uses MA-net, Efficientnet-B3 and Focal Loss, trained for 20 epochs. We further tested for the effects of our filtering procedure (slightly improved from the previous work), and additionally experimented with the introduction of CORONA imagery and increased the input size.

Tepa sites in Uzbekistan. We also performed an additional test on another large dataset (https://www.orientlab.net/samark-land/) elaborated by the Uzbek-Italian Archaeological Project at Samarkand[39]. Given the similarity between the Mesopotamian Tell and the Uzbek Tepa, we wanted to see if the model was able to detect those sites without the need of additional retraining.

The dataset features 2318 point-like annotations categorized in different ways which also come with attributes related to their preservation states. We selected only sites classified as either *Tepa* or *Low Mound,* with the *Well-preserved* label. The final number of sites ends being 215: 148 Tepa and 67 Mounds. The actual test set images were created following the same procedure described above.

Results

Mesopotamia. First, we present the results in terms of average Intersection-over-Union (IoU) score on the test dataset. We define the metrics as follows: $IoU = \frac{P \cap G}{P \cup G}$ with P indicating the predicted shape and G the ground truth shape. IoU represents the degree of correspondence between the predicted shape and annotation in the dataset. While it gives us an idea of how the model behaves and helps us to select the best one, we must recognize that it does not indicate how many sites are identified or not, which is our primary goal.

Table 1 summarizes the results for all models on the holdout dataset, as described in the Methods section. Note that, for each model, we report a mean score and the associated standard deviation. This is due to the fact that we are performing a random crop on the images, even on the test set, and thus we run ten tests with different crops to average out this effect.

The first thing that can be noted is the marked improvement given by the increase in the input size. We imagine that the larger area provides more context to the predictions and makes the model more accurate. As important is the filtering procedure described above, that tries to remove small and undetectable sites, resulting in a bump in performance regardless of the input size.

Finally, the use of CORONA imagery is a bit controversial. For the smaller input size, it seems to provide no benefits (the lower error score is within the margin of error) and we can hypothesize this is due to the low resolution of this imagery. With larger areas they instead seem to provide an increase in performance, maybe again due to the larger context. Inspecting the prediction, however, revealed the absence of a marked difference, perhaps meaning the IoU is increasing just as the result of slightly more precise contours.

Detection accuracy. To further assess the results, we moved on to detection accuracy. First, we transformed the raster predictions from the model into vector shapes using the well-known library GDAL[40] and then we looked for the intersection between the site annotations and the predictions. To obtain smoother shapes, before the conversion we first applied a Gaussian blur to the prediction rasters and then clipped values above a certain threshold (0.5, but the number can be changed for a more or less sensitive model) to 1.0, while everything else would be set to 0.0.

This automatic evaluation gives good but not too exciting results, with an accuracy score of 62.57% for Model 5 and 60.08% for Model 6. A model able to find two out of three sites would already provide a good starting point for human analysis. However, archaeologists must provide a verification of the predictions and differentiate the cases in which the model commits proper mistakes from those in which it makes justifiable errors that a human would do too[41–43].

First of all, there are a considerable number of sites that are no longer visible from present day satellite imagery and were not filtered from the dataset. This was expected as only half of the annotations had additional information and even less contained indication of their visibility. Any input image containing only sites that are no longer visible should be considered as True Negative rather than False Negatives if the model produces no contour.

When it comes to predictions marked as False Positive, sometimes the model predicts another site close by, instead of not the one being tested. This can be considered a mistake or not depending on the nature of the "missed" site. In the case the missed site is one of those no longer visible, but we detect a near visible one, the prediction is actually a True Positive. On the other hand, the missed site can be one that is still visible but maybe less so than another one in the picture. In this situation we could either consider both a False Negative and a true positive, or just as a true positive given that, in a real-world scenario, the closeness to other sites would result in a useful suggestion as the human expert, who would then be able to retrieve them all. Alternatively, we could avoid considering non-visible sites altogether, but the difference would be minimal (accuracy 78.37% and recall 82.01%).

Model	Input	Filter	IoU (%)	St.dev
Model 1	Bing 1k		74.17	0.38
Model 2	Bing 1k	✓	78.10	0.54
Model 3	Bing + CORONA 1k		74.06	0.39
Model 4	Bing 2k		79.77	0.34
Model 5	Bing 2k	✓	81.54	0.35
Model 6	Bing + CORONA 2k	✓	83.45	0.18

Table 1. IoU scores for the different experimental setup we tested. The standard deviation comes from the repeated testing used to average out random cropping.

Lastly, some predictions were actually present in the outputs but too faint for the cutoff threshold we imposed. We did not adjust for those errors, but they indicate a possible approach for interaction: using predictions as overlays and manually looking at the map. Alternatively setting a lower threshold could solve the problem.

The adjustment raises accuracy and recall to around 80, giving a more objective idea of the actual model performance.

Table 2 summarizes the results for the automatic evaluation and the adjusted values after the human evaluation highlighted non-visible sites. The following equations define the metrics used in terms of True/False Positive/Negatives. We chose Accuracy, Precision, Recall and the Matthews Correlation Coefficient.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

$$Recall = \frac{TP}{TP + FN}; Precision = \frac{TP}{TP + FP},$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

It is interesting to see how Model 6, which got a higher IoU score, seems to actually be performing worse now. Looking at the images, it appears that this model is a little bit more restrained and cautious, resulting in less positive predictions and thus less False Positives. In turn, this can result in a higher IoU because it reduces the Union term, and, if areas are a little bit more precise, it even raises the Intersection term. However, for detection's sake, we need the presence of an intersection rather than a perfect match and in this situation the lower number of positives is punishing. Overall, the difference in accuracy is not excessive, so both models are useful and could be used in parallel, but we must also consider the additional complexity and cost of using two sets of input images which make Model 6 a bit cumbersome. For this reason, we moved on using just Model 5.

We concluded this subsection with Fig. 2, which contains a few examples from the test dataset to display the quality of the model's outputs. Note how the colors correspond to probability values, and that faint areas would be cut off by the 0.5 threshold we use in creating the vector shapes. The model is very accurate at tracing the site outlines and in some cases (i.e., the first column in Fig. 2) these are even more accurate than the ground truth with respect to current satellite imagery.

A test in the Maysan province. After assessing detection performance, we wanted to try the model on a rectangular area within the unsurveyed Maysan province for which we carried out remote sensing. This test had the goal of evaluating how many False Positives the model would predict and to give an example of the mistakes the model makes in an operational scenario.

The area we selected contains 20 alleged sites and spans 104 km^2. Figure 3 shows the area with the annotation from the archaeologist and the prediction from the model. As it can be seen the model is able to recover 17 of the 20 sites while also suggesting around 20 more shapes (or less, depending on what is considered a single instance). Most of those suggestions are not useful but are also easily and quickly sifted out by an expert eye, especially in context, given their size or their location.

Figure 4 instead shows an overlay produced by stitching together the various predictions and using the probabilities values as a sort of heatmap. "Hotter" colors correspond to higher probabilities while black indicates the absence of a site. Note that the palette is the same as the one seen in Fig. 2, with dark purple colors indicating a relatively low probability (less than 0.5). The transparency is obtained through the use of the Overlay filter in QGIS.

Uzbekistan. Unfortunately, human evaluation of the outputs showed that the model is able to correctly identify only around 25% to 30% of the sites in this region, depending on how thresholds are chosen. The remaining part contains either sites that are missed completely or sites that are somehow hinted either too faintly or inside a huge area that appears meaningless.

The reason for this severe drop in performance is most probably due to the different nature of the landscape in the region which in some locations appear to be way more urbanized and in general features more vegetation: thus, not all floodplain environments are similar enough for a direct cross-comparison. Furthermore,

Model	Evaluation	TP	TN	FP	FN	Accuracy (%)	Precision (%)	Recall (%)	MCC (%)
Model 5	Automatic	228	98	70	125	62.57	76.51	64.59	21.65
	Adjusted	258	185	40	68	80.40	86.58	79.14	60.53
Model 6	Automatic	209	104	57	151	60.08	78.57	58.06	20.94
	Adjusted	239	197	27	88	79.13	89.85	73.09	59.99

Table 2. Site detection performance for the best models. Automatic evaluation considers the labels as they come, adjusted evaluation compensates for incorrect labels with a human in the loop.

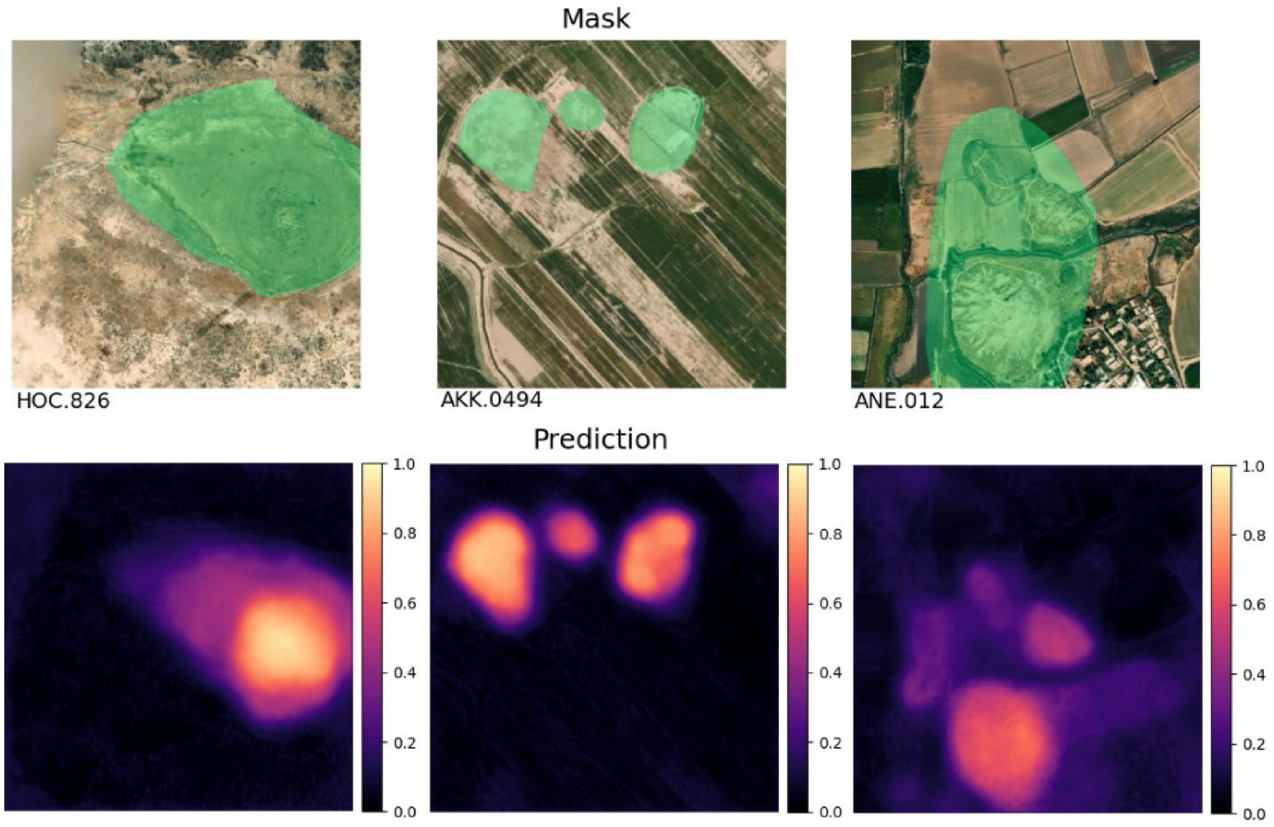

Figure 2. A few sample predictions from the test set. On the left is the target mask overlaid on the input image. On the right the model output. The color bar corresponds to classification probability. Note how the model is capable of matching accurately the site outline. All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://bit.ly/NSR_floodplains).

the conventions which lie behind the annotations in the Uzbek dataset might not be perfectly aligned with the Mesopotamian one further complicating the situation.

This partial failure must be set into a context, since we do believe that our method can be applied to a wide array of similar environments in Asia and beyond having multi-period settlement histories: the only way of dealing with this problem here is that of creating a small dataset of selected Tepa sites and perform an additional round of transfer learning so that the model may grasp the new context and characteristics of the given region.

Discussion

The results obtained can be considered satisfactory even if the IoU metric, when compared to other semantic segmentation applications, is not extremely high. When testing for detection performance, however, we found that the model is still able to detect most sites in the dataset, leaving us with good expectations for its use in other parts of the survey area. As the Uzbek test shows however, when it comes to new areas with similar sites but in a different context, performance may drop severely. This issue of transferability, as it is referred to in archaeology, is an active research topic. A retraining phase, even with a smaller dataset, could hopefully fix the issue and future work may explore this research direction.

It is important to note how evaluation metrics in this task seem to hit a wall when confronted with the fact that they are computed against annotations that oftentimes are not homogeneous and contain various spurious labels[44]. In our case we coped with the fact that there are many sites that are only visible on some historical photographs or maps that are part of the dataset even if they do not provide useful examples. Fortunately, the model seems to be robust enough to learn useful concepts and ignore these confounding data points. Still a smaller, cleaner dataset could drastically improve performance while also reducing computational load. Obviously, such cleaning operations would be a massive investment in terms of time and archaeologists would rather spend it actively searching for sites themselves, instead.

Our model, however, opens up the possibility of going through already surveyed areas automatically and then producing a list of predictions that contrast the annotations to be manually reviewed. Subsequently a new, cleaner dataset could be assembled by the archaeologists and a new improved model could be trained. See Lambers et al.

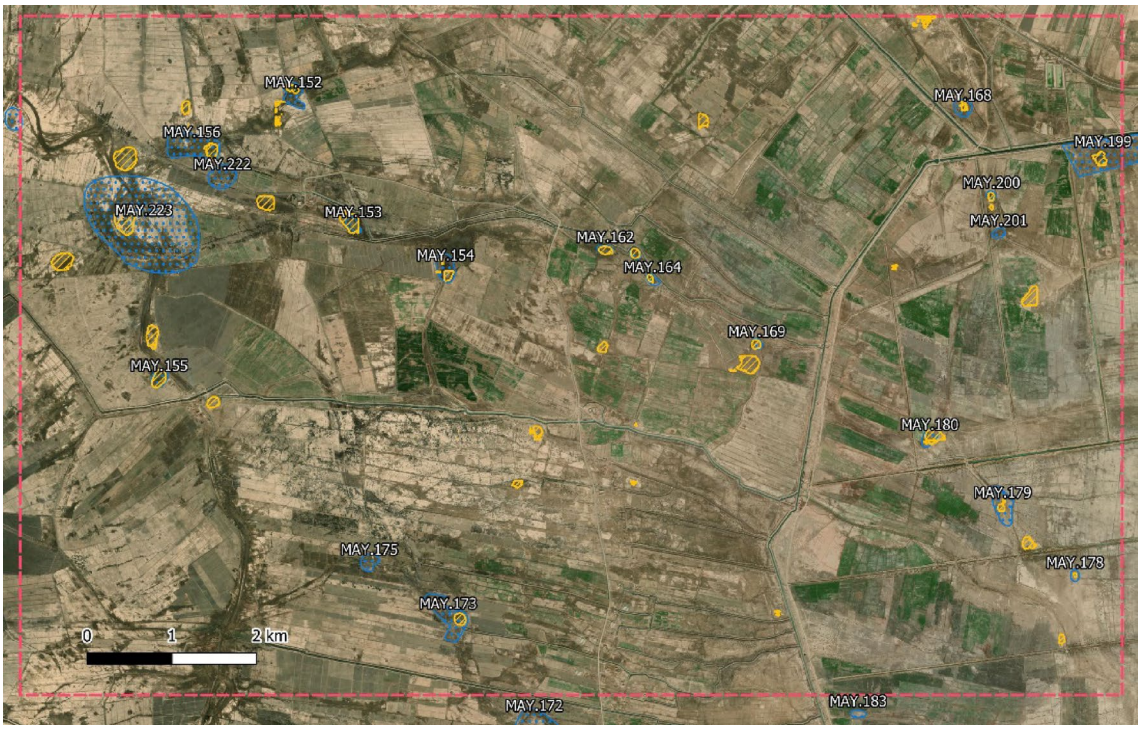

Figure 3. Maysan province test area (pink, dashed line) with sites remotely identified by archaeologists (blue, dot-filled) and model predictions (yellow, line-filled). The sites identified by the trained eye and the model are equivalent and, most importantly, the model is able to ignore areas without significant features. All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://bit.ly/NSR_floodplains).

for an example using citizen science[45]. This same procedure also works in applications to new areas, where novel predictions can be manually checked and added to a new dataset overtime.

In addition to the automatic procedure, the model could also be used to produce an overlay to guide the eye of the archaeologist inside a GIS software. This graphical approach allows the users to also compare the overlay with other maps they might be using and use their expertise to infer the existence of a site based on all contextual information they have[46,47]. We only tried this approach on a small area as shown in Fig. 4, but the computation could be easily scaled up to cover huge areas, as it takes less than a second to produce an output and there is no need to complete the operation in one go anyway. The only shortcoming of this method is the evident mismatch at the border between different input images, which give the overlay its mosaic-like appearance. In theory, semantic segmentation could work with inputs of arbitrary size, but doing so requires a huge amount of memory which might not be available. A solution might be the creation of overlapping prediction maps that would then be averaged, trading off computational time for increased precision.

Figure 5 summarizes the use we envision for the model we described, in the vein of similar solutions[48,49]. Starting from the dataset the model produces prediction masks that we can manipulate through post-processing to obtain either a vector shapefile that can be used for automatic evaluation and detection of sites. At this stage the user has the possibility of choosing a threshold to cut prediction off and the use of techniques to smooth the output shapes, like blurring or buffering the vectors. Similarly, the map overlay can be adjusted by selecting different graphical representations directly into the GIS software. The goal in this case is that of spotting sites that might go undetected by the automatic comparison because their probability is lower than the threshold, while still being distinguishable for a human. Each time the model is used, in either way, after reviewing the outputs the users would be able to obtain either a new set of annotations or a list of sites to be removed or relabeled. If such a workflow is used by more than one team it could also greatly speed up the search efforts: the use of open technologies in this case makes the results easier to share between research groups, which could greatly help archaeology as a field[50].

The experiments with CORONA imagery also hint at the possibility of combining more models, perhaps trained with different basemaps or a combination of them, and compare the prediction given by all of these. Especially if historical images are present, we could end up with a dataset that also contains temporal information about when a site is visible and when it becomes undetectable. This latter aspect is quite novel and represents a potential breakthrough in automated remote sensing. Use of stereoscopic images for the creation of elevation models could also benefit the task, if the resolution is sufficient to highlight the low mounds we are looking for.

Figure 4. The Maysan test area prediction probabilities layer visualized as the top one within QGIS. This visualization allows the user to decide where to look instead of relying on a predefined threshold value. All the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under the Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://bit.ly/NSR_floodplains).

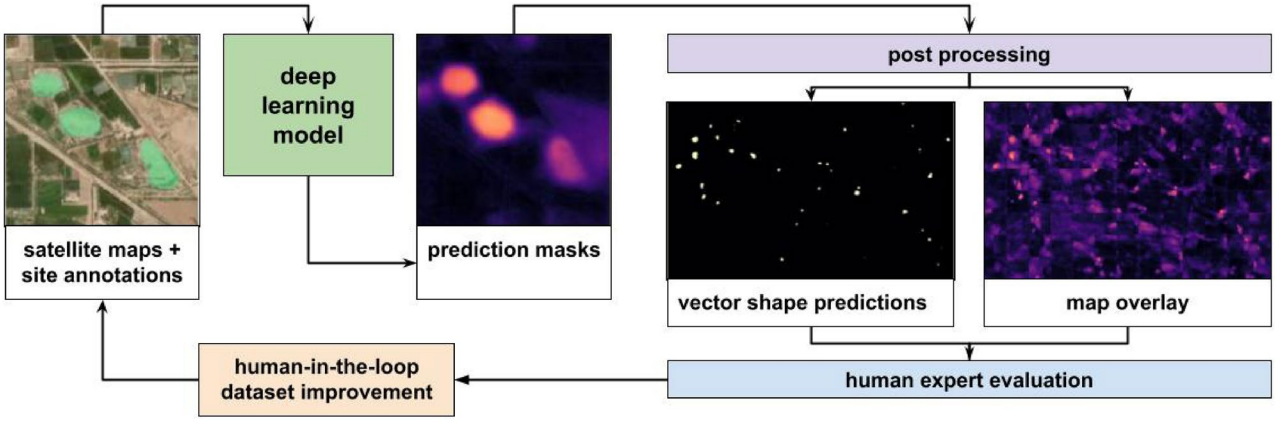

Figure 5. A human-in-the-loop workflow based on our model. A model is trained from annotated images and provides predictions masks. The masks can be used as an overlay or vectorized. Human evaluation is conducted on the outputs and in turn a refined dataset can be created to improve the model.

Conclusions

We presented a deep learning model for detection of mounded archaeological sites in the Mesopotamian floodplain. The model was implemented using pretrained models for semantic segmentation, fine-tuned on satellite imagery and masks of the site shapes coming from a dataset containing almost 5000 examples.

The result of our experiments is a model which obtains an IoU score of 0.8154 on the test dataset and detects sites with 80% of accuracy. This statistic accuracy however is adjusted for the considerable number of sites that appear mislabeled as they are no longer visible on modern satellite imagery. While we cleaned up the dataset to the best of our ability, many undetectable sites still remained. The model seems to be quite robust, however.

Following this result, we propose a workflow for the archaeologists to adopt, in which their already established remote sensing practices are supported and enhanced by the use of a model like our own. The outputs can be used both for very fast automatic detection, being aware of the mistakes this could introduce, or combined to generate

a graphical overlay to direct the user's attention towards certain areas. In turn, the use of the model will result in new shape files and annotations that can be used for retraining and improving the model, as well as enabling further analyses. The potential applications of this method are far reaching and do not only concern its speed: it should rather be seen as a necessary complement to traditional expert-based photointerpretation, adding to the latter in many cases site features which may go overlooked but are likely to be significant.

Data availability

In addition to the specific information provided within the paper, all code, data and various resources are available on GitHub (https://bit.ly/NSR_floodplains). As to geographical data, all the displayed data fall under the condition of fair use utilization of geographical data for academic purposes. The list of all relevant data/software provider(s) is as follows: (i) original maps creation under Section 5 of Microsoft Bing Maps Platform APIs' terms of use (https://www.microsoft.com/en-us/maps/product/print-rights); (ii) maps display achieved with an open source software, under the GNU licenses of QGIS (https://qgis.org/en/site/) and QuickMapsServices (https://github.com/nextgis/quickmapservices); (iii) final maps elaboration achieved with a software developed by the authors and available at (https://bit.ly/NSR_floodplains).

Received: 9 March 2023; Accepted: 27 May 2023
Published online: 29 May 2023

References

1. Verschoof-van der Vaart, W. B. & Landauer, J. Using CarcassonNet to automatically detect and trace hollow roads in LiDAR data from the Netherlands. *J. Cult. Herit.* **47**, 143–154. https://doi.org/10.1016/j.culher.2020.10.009 (2021).
2. Torrey, L. & Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (eds Torrey, L. & Shavlik, J.) 242–264 (IGI Global, 2010).
3. Deng, J. *et al*. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
4. Traviglia, A., Cowley, D. & Lambers, K. Finding common ground: Human and computer vision in archaeological prospection. *AARGnews Newslett. Aerial Archaeol. Res. Group* **53**, 11–24 (2016).
5. Palmer, R. Editorial. *AARGnews* (2021).
6. Wilkinson, T. J., Gibson, M. & Widell, M. *Models of Mesopotamian Landscapes: How Small-Scale Processes Contributed to the Growth of Early Civilizations* (Archaeopress, 2013).
7. Adams, R. M. *Land Behind Baghdad: A History of Settlement on the Diyala Plains* (University of Chicago Press, 1965).
8. Adams, R. M. *Heartland of Cities: Surveys of Ancient Settlement and Land Use on the Central Floodplain of the Euphrates* (University of Chicago Press, 1981).
9. Adams, R. M. & Nissen, H. J. *The Uruk Countryside: The Natural Setting of Urban Societies* (University of Chicago Press, 1972).
10. Marchetti, N. *et al.* The rise of urbanized landscapes in Mesopotamia: The QADIS integrated survey results and the interpretation of multi-layered historical landscapes. *Z. Assyriol. Vorderasiat. Archäol.* **109**, 214–237. https://doi.org/10.1515/za-2019-0016 (2019).
11. Wilkinson, T. J. *Archaeological Landscapes of the Near East* (University of Arizona Press, 2003).
12. Lyons, T. R. & Hitchcock, R. K. *Aerial Remote Sensing Techniques in Archeology* (Chaco Center, 1977).
13. Kucukkaya, A. G. Photogrammetry and remote sensing in archeology. *J. Quant. Spectrosc. Radiat. Transf.* **88**, 83–88 (2004).
14. Karamitrou, A., Sturt, F., Bogiatzis, P. & Beresford-Jones, D. Towards the use of artificial intelligence deep learning networks for detection of archaeological sites. *Surf. Topogr. Metrol. Prop.* **10**, 044001 (2022).
15. Hendrickx, M. *et al.* The use of stereoscopic images taken from a microdrone for the documentation of heritage—An example from the Tuekta burial mounds in the Russian Altay. *J. Archaeol. Sci.* **38**, 2968–2978 (2011).
16. Küçükdemirci, M. & Sarris, A. GPR data processing and interpretation based on artificial intelligence approaches: Future perspectives for archaeological prospection. *Remote Sens.* **14**, 3377 (2022).
17. Balsi, M. *et al.* Preliminary archeological site survey by UAV-borne lidar: A case study. *Remote Sens.* **13**, 332 (2021).
18. Assael, Y. *et al.* Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022).
19. Verschoof-van der Vaart, W. B., Lambers, K., Kowalczyk, W. & Bourgeois, Q. P. Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from the Netherlands. *ISPRS Int. J. Geo Inf.* **9**, 293 (2020).
20. 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.* **26**, 165–175 (2019).
21. Anichini, F. *et al.* The automatic recognition of ceramics from only one photo: The ArchAIDE app. *J. Archaeol. Sci. Rep.* **36**, 102788 (2021).
22. Mantovan, L. & Nanni, L. The computerization of archaeology: Survey on artificial intelligence techniques. *SN Comput. Sci.* **1**, 1–32 (2020).
23. Bickler, S. H. Machine learning arrives in archaeology. *Adv. Archaeol. Pract.* **9**, 186–191 (2021).
24. Guyot, A., Lennon, M., Lorho, T. & Hubert-Moy, L. Combined detection and segmentation of archeological structures from LiDAR data using a deep learning approach. *J. Comput. Appl. Archaeol.* **4**, 1 (2021).
25. Trier, Ø. D., Salberg, A.-B. & Pilø, L. H. Semi-automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In *CAA2016: Oceans of Data. Proc. 44th Conference on Computer Applications and Quantitative Methods in Archaeology* 219–231 (Archaeopress, 2018).
26. Bickler, S. H. & Jones, B. Scaling up deep learning to identify earthwork sites in Te Tai Tokerau, Northland, New Zealand. *Archaeology* **16**, 1 (2021).
27. Caspari, G. & Crespo, P. Convolutional neural networks for archaeological site detection—Finding "princely" tombs. *J. Archaeol. Sci.* **110**, 104998 (2019).
28. Orengo, H. A. *et al.* Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *Proc. Natl. Acad. Sci.* **117**, 18240–18250. https://doi.org/10.1073/pnas.2005583117 (2020).
29. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
30. Fan, T., Wang, G., Li, Y. & Wang, H. MA-Net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access* **8**, 179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372 (2020).
31. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems* (eds Vaswani, A. *et al.*) 5998–6008 (MIT Press, 2017).
32. da Costa, L. B. *et al.* Deep semantic segmentation for detecting eucalyptus planted forests in the Brazilian territory using sentinel-2 imagery. *Geocarto Int.* **37**, 6538–6550 (2022).

33. Li, R. *et al.* Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–13 (2021).
34. Roccetti, M. *et al.* Potential and limitations of designing a deep learning model for discovering new archaeological sites: A case with the Mesopotamian floodplain. In *Proc. 6th EAI International Conference on Smart Objects and Technologies for Social Good* 216–221 (Association for Computing Machinery, 2020).
35. QGIS Development Team. *QGIS Geographic Information System* (QGIS Association, 2022).
36. Casini, L., Orrù, V., Roccetti, M. & Marchetti, N. When machines find sites for the archaeologists: A preliminary study with semantic segmentation applied on satellite imagery of the Mesopotamian floodplain. In *Proc. 2022 ACM Conference on Information Technology for Social Good* 378–383 (2022).
37. Casana, J. & Cothren, J. The CORONA atlas project: Orthorectification of CORONA satellite imagery and regional-scale archaeological exploration in the Near East. In *Mapping Archaeological Landscapes from Space* (eds Comer, D. C. & Harrower, M. J.) 33–43 (Springer, 2013).
38. Iakubovskii, P. Segmentation models pytorch. *GitHub Repository* (2019).
39. Mantellini, S. & Berdimuradov, A. E. Evaluating the human impact on the archaeological landscape of Samarkand (Uzbekistan): A diachronic assessment of the Taylak district by remote sensing, field survey, and local knowledge. *Archaeol. Res. Asia* **20**, 100143. https://doi.org/10.1016/j.ara.2019.100143 (2019).
40. GDAL/OGR Contributors. *GDAL/OGR Geospatial Data Abstraction Software Library* (Open Source Geospatial Foundation, 2022).
41. Baeza-Yates, R. & Estévez-Almenzar, M. The relevance of non-human errors in machine learning. In *EBeM'22: Workshop on AI Evaluation Beyond Metrics* (2022).
42. Cowley, D. C. In with the new, out with the old? Auto-extraction for remote sensing archaeology. In *Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2012* 37–45 (SPIE, 2012).
43. 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.* **11**, 1994. https://doi.org/10.3390/rs11171994 (2019).
44. Fiorucci, M. *et al.* Deep learning for archaeological object detection on LiDAR: New evaluation measures and insights. *Remote Sens.* **14**, 1694. https://doi.org/10.3390/rs14071694 (2022).
45. Lambers, K., Verschoof-van der Vaart, W. B. & Bourgeois, Q. P. J. Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection. *Remote Sens.* **11**, 794. https://doi.org/10.3390/rs11070794 (2019).
46. Verschoof-van der Vaart, W. B. *Learning to Look at LiDAR: Combining CNN-Based Object Detection and GIS for Archaeological Prospection in Remotely-Sensed Data* (Leiden University, 2022).
47. Verschoof-van der Vaart, W. B. & Lambers, K. Applying automated object detection in archaeological practice: A case study from the southern Netherlands. *Archaeol. Prospect.* **29**, 15–31. https://doi.org/10.1002/arp.1833 (2022).
48. Herfort, B. *et al.* Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning. *Remote Sens.* **11**, 1799. https://doi.org/10.3390/rs11151799 (2019).
49. Ponti, M. & Seredko, A. Human-machine-learning integration and task allocation in citizen science. *Humanit. Soc. Sci. Commun.* **9**, 1–15. https://doi.org/10.1057/s41599-022-01049-z (2022).
50. Marchetti, N. *et al.* NEARCHOS. Networked archaeological open science: Advances in archaeology through field analytics and scientific community sharing. *J. Archaeol. Res.* **26**, 447–469 (2018).

Author contributions

L.C. wrote the manuscript, designed, and run the experiments with A.M.; V.O. provided the dataset and human-evaluation; N.M. and M.R. supervised the study and revised the manuscript.

Funding

The funding was provided by European Commission (CSOLA/2016/382-631), the Volkswagen Foundation (Kalam Project) and the University of Toronto (CRANE 2.0 project).

Competing interests

Nicolò Marchetti has been funded by the following projects: (i) the "EDUU—Education and Cultural Heritage Enhancement for Social Cohesion in Iraq" project, funded by EuropeAid (CSOLA/2016/382–631), www.eduu.unibo.it, in the framework of which the FloodPlains project was developed, https://floodplains.orientlab.net/; (ii) the "KALAM. Analysis, protection and development of archaeological landscapes in Iraq and Uzbekistan through ICTs and community-based approaches" project, funded by the Volkswagen Foundation, www.kalam.unibo.it; (iii) the CRANE 2.0 project of the University of Toronto, which provided the geospatial servers on which FloodPlains is running. All the other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.