**Archeologia e Calcolatori**
**37.1, 2026, 229-254**
**doi 10.19282/ac.37.1.2026.12**

# AIOH: ARTIFICIAL INTELLIGENCE OPERATOR FOR HERITAGE. THE CASE STUDY OF NURAGHE DETECTION

## 1. Introduction: AI, remote sensing and archaeology

The study of aerial images has long been an essential tool in archaeology for surveying the territory and identifying sites and monuments. Starting from the earliest cases at the end of the nineteenth and throughout the twentieth century, this proved to be a highly effective method for research activities (Bourgeois, Meganck 2005; Barber 2011). Once satellite cameras were available, such images became an extraordinary resource, leading to the emergence of a new branch of the discipline focused on photo-interpretation (Parcak 2009). Given the vast amount of data potentially available for analysis, over time there has been a growing trend toward developing automatic and semi-automatic systems for identifying elements of archaeological interest (Lambers, Traviglia 2016).

Within this context, the advent of Artificial Intelligence in relation to satellite photo-interpretation studies takes place. Alongside conventional image-processing techniques across different areas of the electromagnetic spectrum, the possibility of using AI has been introduced. Different domains have used the potential of this technology, enabling novel analytical approaches and effective interpretative models. This has also sparked discussions about the types of algorithms used and their flexibility and transparency for research purposes in the field of Cultural Heritage (Fiorucci *et al.* 2020; Jamil *et al.* 2022). Nevertheless, it is argued that the use of Deep Learning (DL) for archaeological data detection is still insufficient. In other words, although Machine Learning (ML) and DL techniques are widely adopted for analyzing and categorizing finds, there is still a significant gap in their use for detecting new archaeological data (Jamil *et al.* 2022).

### 1.1 *DL and ViT models*

Among the various approaches that are grouped under the broad label of 'Artificial Intelligence', so-called "deep learning" (DL) is certainly one of the most fascinating and promising. To explain how it works in the simplest possible way, we can say that in this approach the machine is able to analyse a dataset (images, in our case) by considering a very large number of mathematical parameters in the relationships between its elementary components (pixels). With appropriate training on cases where the elements of interest are either present or absent, the system can detect both features visible to the human eye (for example, a change in the colour of a stone structure compared to its surroundings) and features that are invisible, emerging only through numerical computation (such as the recurrence of a particular shade within vegetation). This can therefore lead to the discovery of previously unknown recurring patterns, which scholars must then interpret. In this sense, it is not uncommon to encounter instances of 'serendipity' in these models: the possibility of discovering aspects of the sample that were not part of the original research hypotheses.

DL processes have found extensive applications in various domains in the past decade to aid classification tasks and they operate through a category of algorithms, or models, known as Neural Networks. Historically, several approaches have leveraged models grounded in Convolutional Neural Networks (CNN), like VGG (Simonyan, Zisserman 2015) and ResNet (He *et al.* 2015). These models are fundamentally based on hierarchical, multilayered structures designed to automatically and adaptively learn spatial hierarchies of features from input images. They are renowned for their depth and complexity, enabling the learning of intricate patterns and representations from the input data, which is crucial for tasks like image classification and/or object detection. Some models, such as SqueezeNet (Iandola *et al.* 2016) and MobileNet (Howard *et al.* 2019), are particularly noteworthy for their efficiency and ability to balance accuracy and computational speed.

Recently, the advent of Vision Transformers (ViTs) (Dosovitskiy *et al.* 2021) has represented a paradigm shift in image classification and computer vision tasks. Unlike CNNs, which analyze small, individual sections of an image sequentially, ViTs, through the attention mechanism (Vaswani *et al.* 2017), can assess multiple sections of an image simultaneously, determining the importance and focus of different parts, thus gaining a more comprehensive understanding of the image. Studies show this architectural difference makes ViTs more robust to texture bias (ViTs rely less on local textures and more on structural shapes) (Bai *et al.* 2021).

It is clear that algorithms of this kind, if proved effective for archaeological research as well, would offer enormous potential for rapid application across very large areas. Our challenge is therefore to test the use of a DL system for the large-scale detection of archaeological monuments, with the aim of developing tools that could be widely adopted, using easily manageable and freely available satellite images as to cover entire regional areas. We tried to create a prototype of this type using for the images, probably the most widespread and publicly accessible satellite photo dataset in the world: Google Maps; and, as a case study, a kind of archaeological structure sufficiently standardized, characteristic and widespread over a large territory: the Sardinian nuraghes.

In this paper, we propose AIOH (Artificial Intelligence Operator for Heritage), a system to automatically classify images on the basis of a score

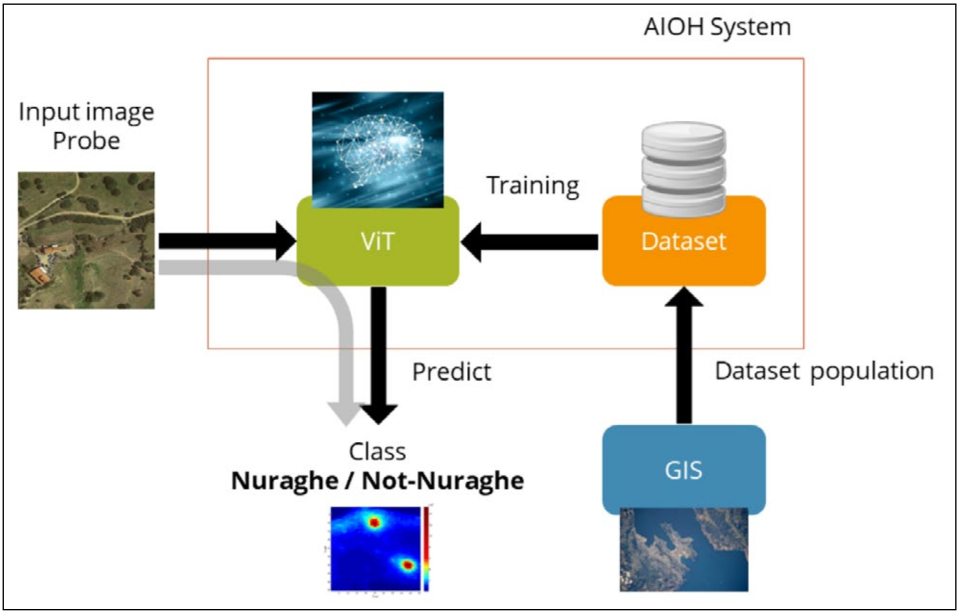


Fig. 1 – Pipeline of the proposed method.

related to the relevance of factors characterizing the archaeological features for the case study of nuraghe detection. The pipeline is shown in Fig. 1. The core of the system is a ViT-based classifier, explained in Section 3.1, which is trained on our dataset of labeled RGB images retrieved through GIS: the NuragAI dataset. The model can thus make inferences on probe data and return a probability that the image contains a nuraghe alongside a heatmap that defines the focus points in the image that support the system's decision.

## 2. The nuraghe challenge

The term 'nuraghe' traditionally refers to a particular kind of monument, typical of the island of Sardinia (Italy), where such architectural items are diffusely (and exclusively) present. The nuraghes' shapes are not unique: detailed classifications have been created by archaeologists over time. Nevertheless, some common features may be stated to reach a satisfactory definition. Nuraghes are complexes consisting of one or more truncated cone-shaped towers, with or without a series of lower rooms, round-shaped and connected to them. They are built using a dry stone technique by connecting regularly (generally squared) shaped stones. Such complexes may result in a quite wide range of structures, which has led to the definition of 'nuragic village'. Such

Fig. 2 – Four examples of nuraghes seen from the ground, in different areas of Sardinia: a) Nuraghe Albucciu; b) Nuraghe Burghidu; c) Nuraghe La Prisgiona; d) Nuraghe Losa.

archaeological remains are widely diffused all over the island, although with different concentrations, and are traditionally considered very specific cultural and geographical landmarks, characterizing Sardinia and representing the visible sign of a specific civilization (Fig. 2).

Although such a phenomenon has been deeply studied from the end of the XIX century (Centurione 1886; Lilliu 1982a, 1982b), lacking writings and a sufficient number of well-dated excavations, nuraghes' classification and chronology are mainly the result of formal analysis and individual considerations. Furthermore, the reuse of the structures over time and their deterioration due to natural events as well as human action, may have caused the reduced recognisability (or the total disappearance) of some of them. Generally, the Nuraghe civilization is considered to date back between the XVIII and VII centuries BC (Cossu *et al.* 2018), although it is difficult to define a reasonable time span, both for the exposed reasons connected to archaeological records and because of the uninterrupted use of the structures as herder shelters, storehouses, etc. There is no official census of the monuments or a defined number of items. Different sources generally report a count of 8000-10,000 elements for the whole island (Vacca *et al.* 1998; Cossu *et al.* 2018; Maxia 2020).

In recent years, some careful studies have thoroughly attempted to reach inclusive databases, the main one being probably the 'SardegnArcheologica' portal (https://www.sardegnarcheologica.it), and they represent a relevant basis to set up a wide survey project. The proposed system, AIOH, has been trained with custom data collected and structured specifically for this classification task.

The research plan has been set up considering the need for a consistent number of images, both containing and not containing the desired item (nuraghes), to reach a satisfactory training and verifying database. Nevertheless, the two sets required different approaches to be collected.

## 3. The proposed method

### 3.1 *Artificial Intelligence Operator for Heritage*

This framework's heart is a predictive model, also called the backbone, based on the ViT architecture (Dosovitskiy *et al.* 2021). It analyzes images and assigns a probability score indicating the likelihood of an image being a positive sample, with a higher score signifying greater confidence in the presence of a nuraghe. Achieving accurate predictions necessitates a training process where the model learns to distinguish between the two categories. The training involves exposing the model to a training set comprising images labeled as containing or not containing a nuraghe, as annotated by human experts. The model processes these images, attempting to predict their labels, and adjusts its parameters based on the accuracy of these predictions. This process is conducted over multiple epochs, each containing several training steps where the model processes subsets of the training data. The entire training set is shown to the model on each epoch, allowing for progressive refinement of its predictive capabilities.

Our framework uses data augmentation techniques to enhance the model's training process, including random cropping and jitter, contrast, and saturation adjustments. Data augmentation is a strategy used to increase the diversity of training data without collecting new images. It involves systematically modifying the existing images and creating different versions from which the model can learn. For instance, random cropping alters the focus and perspective by selecting different portions of an image, which helps the model learn to recognize nuraghe structures regardless of their position or size.

Similarly, adjustments in jitter, contrast, and saturation introduce variations in color and brightness, preparing the model to handle real-world scenarios where lighting and color conditions can vary significantly. Resizing images to a fixed square dimension alongside these augmentations ensures consistency in the input data, allowing the model to process images uniformly. These augmentation techniques collectively expand the range of data the model is exposed to. This is vital for enhancing its generalization and accuracy across diverse and unforeseen image conditions.

Each epoch also encompasses validation steps, where the model is evaluated against a separate dataset not seen during training. This is critical for assessing the model's ability to generalize to new data and preventing a phenomenon known as overfitting. The validation set samples are solely used to quantify the model's performance and not for refining its parameters.

### 3.2 *NuragAI dataset*

The NuragAI dataset, conceived for the system training, was obtained by collecting two satellite image categories, roughly defined 'nuraghe' (image containing a nuraghe) and 'non-nuraghe' (image not containing a nuraghe). Different methods were used to collect images in the two categories, through a brief and easy workflow made in QGIS which generated the picture datasets (Sherman 2014).

The SardegnArcheologica database represents the most complete geo-referenced dataset on existing nuraghes. For this reason it was chosen as the starting point for the generation of training images. Nevertheless, some of the thousands of monuments recorded may be not precisely located. The original dataset from SardegnaArcheologica.it[1] includes about 8250 site points, downloaded as a CSV; this file was adapted to be loaded on QGIS as a point layer. The geographical script is set to automatically extract the nuraghes sample images in JPEG format, with the coordinate 'World File' (thus meaning that every image is georeferenced). The training images used are taken from the Google Satellite layer (Google Satellite License is free for 'Fair use, https://www.google.com/intl/en-GB_ALL/permissions/geoguidelines).

Moreover, a simple survey by satellite images was planned to measure the largest example of known nuraghes, such as Barumini, Genna Maria, and Arrubiu. This passage was useful in determining the minimum export image size, which was established to be 422,500 square meters (650x650 m). The 'nuraghes' point layer was chosen as geometry layer index for the QGIS 'Atlas Generator' to create the images automatically.

[1] The use of the dataset is free for study and research (https://sardegnarcheologica.it/text/1000/it).

Fig. 3 – Four examples of 'visible nuraghes': a) La Prisgiona; b) Genna Maria; c) Is Orbais; d) Nuristene (Google photos, elaboration by AIOH project).

Although the dataset creation process started according to the information of the 'SardegnArcheologica' website, for our purpose, a start 'from scratch' must be pursued, avoiding referring to existing classifications of nuraghes (nuraghe 'a corridoio', 'complesso', etc.), influenced by many arbitrary and misleading criteria, typical of classifying actions in archaeology (Palombini 2001), and trying to focus on formal aspects that are statistically recognized. The categories not defined as 'nuraghe', present in the SardegnArcheologica database, have been flagged out, because they are completely different structures. Once selected and downloaded, satellite images were divided into three macro-groups through a human optical analysis (data labeling).

Fig. 4 – Four examples of 'hardly-visible nuraghes': a) Montumarci; b) Perdas Albas; c) Cambulas; d) Monte Sirai (Google photos, elaboration by AIOH project).

Considering this case study, a manual and internal approach was preferred (Fredriksson *et al.* 2020). The task of selecting, classifying, and manually labeling the data of nuraghes took approximately five months, during which satellite images were optically scrutinized (Fig. 3a, b).

The classification of nuraghes is the following:

– 1. Visible (nuraghes are recognizable at sight):

– 1.1 Well-preserved nuraghes, clearly recognizable at sight (Fig. 3a, b), such as Nuraghe Nuraxi, Mulinu, La Prisgiona (Antona 2013; Fadda 2016), and Genna Maria (Van Dommelen 1998).

Fig. 5 – Two examples of 'invisible nuraghes': a) Punta Spinosa; b) San Sergio (Google photos, elaboration by AIOH project).

– 1.2 Nuraghes in particular environmental and geo-morphological contexts (stone quarries, shores, woods, etc.) but still clearly recognizable at sight (Fig. 3c), as Nuraghe Is Orbais (Cicilloni, Migaleddu 2008).
– 1.3 Nuraghes inside towns or cities but still clearly recognizable at sight (Fig. 3d), like Nuraghe Muristene (Moravetti 2000).

– 2. Hardly visible (nuraghes are hardly recognizable at sight):

– 2.1 Nuraghes in particular environmental and geo-morphological contexts (stone quarries, shores, woods, etc.) and not clearly recognizable at sight (Fig. 4a), as is the case with Nuraghe Montumarci (AA.VV. 1922).
– 2.2 Nuraghes inside towns or cities but not distinguishable from modern buildings (Fig. 4b), as Nuraghe Perdas Albas (AA.VV. 1922).
– 2.3 Vegetation areas which may cover hidden nuraghes (Fig. 4c). In most cases, we are in front of isolated forests or trees, such as Nuraghe Cambulas (Velli 2020).
– 2.4 Different period structures that are backed against nuraghes or reused (Fig. 4d). In the example of Monte Sirai, a Phoenician village reused and developed around the nuraghe (Bartoloni 2000; Finocchi 2005).

– 3. Invisible (nuraghes are not recognizable at sight):

– 3.1 Nuraghes in particular environmental and geo-morphological contexts (stone quarries, shores, woods, etc.) and not recognizable at sight (Fig. 5a), as in the case with Punta Spinosa (Melis 2007).
– 3.2 Nuraghes which have been obliterated by urbanization processes (Fig. 5b). A clear example is Nuraghe San Sergio (Depalmas 1998).

To make the training process as efficient as possible both the nuraghe belonging to the 'visible' and 'hardly visible' categories were included in the 'nuraghe' sample, as the 'non-visible' items were excluded.

### 3.3 *The 'non-nuraghe' subset*

About the non-nuraghe sample, the selection was pseudo-random, as it was performed automatically through QGIS, creating buffer zones around the known nuraghes and sampling the surrounding areas. Images were then checked by an archaeologist, confirming no nuraghes were visible in each frame. We also tried to get different environmental areas, from trees to grass, to urban contexts and even seaside, to improve generalization and to make the system more robust; after the initial trial tests, the subset was enriched with some images clearly representing false-positive cases (e.g. containing round-shaped modern buildings, water towers, etc.). Tiles were made uniform to the 'nuraghe' subset in terms of size and resolution. At the end of the collection procedure and the selection of the best samples, the total number of images containing nuraghes was 1612, while the images without nuraghes ('non-nuraghe') were 6751 (Figs. 3-5).

## 4. Experiments

### 4.1 *Experimental setup*

The model exploited in our study was pre-trained on a large public dataset. To apply transfer learning successfully and make use of the powerful features the model has already learned, it is essential that our input data is processed in the exact same way. Using different statistics would create a domain shift, significantly hurting the model's performance. Thus, we applied the same normalization procedure on our data: each of the three image channels was normalized to achieve a mean and standard deviation of 0.5, also ensuring consistency in lighting and color scales across different images.

As previously mentioned, we applied data augmentation techniques such as random horizontal and vertical flips at a 50% probability to enhance the model's robustness by simulating variations in the input data. Additionally, we employed TrivialAugment (Müller, Hutter 2021), further diversifying the training dataset. The choice of a 384-pixel dimension and specific normalization parameters was made to align with the configurations used during the initial training phase of the backbone architecture, ViT-B-16, ensuring compatibility and optimal performance (to preserve accuracy while reducing computational resources requirements). ViT-B-16 (Vision Transformer - Base 16) is a foundational machine learning model designed for computer vision tasks, leveraging the Transformer architecture originally developed for natural language processing. 'ViT' stands for Vision Transformer; it adapts the self-attention mechanism of Transformers to process images by dividing them into smaller patches. 'B' (Base) denotes a specific size/configuration within the ViT family. '16' refers to the patch size of 16×16 pixels. A smaller patch size (e.g., 16 vs. 32) captures finer details but increases computational load.

For model optimization, we utilized the AdamW optimizer (Loshchilov, Hutter 2019), combining the advantages of Adam optimization with a strategy for managing weight decay, aiding in regularizing the model and preventing overfitting. The implementation was done in Python using the PyTorch framework. Computations were performed on an NVIDIA RTX 3060 GPU with 12GB of VRAM.

### 4.2 *Evaluation scheme*

Several metrics have been employed to evaluate the model in this binary classification task:

1) 'Precision' measures the model's accuracy in identifying only relevant instances among those labeled as positives. This metric is particularly important when the cost of a false positive is high. However, precision alone doesn't account for all relevant instances in the dataset, potentially overlooking some true positives.

2) 'Recall' gauges the model's capability to capture all relevant cases, or true positives, from the dataset. It is essential that, when a positive instance is missing, it is addressed, as it carries significant consequences. The downside of recall is that it does not penalize the model for leading to many false positives.

3) 'Accuracy' provides an overall measure of the model's performance across both positives and negatives. It is straightforward and intuitive, representing the ratio of correct predictions to total ones. However, accuracy can be misleading in datasets where class distributions are imbalanced, as it may favor the majority class.

4) 'AUROC' (Area Under the Receiver Operating Characteristic) curve represents the model's ability to discriminate between classes at various threshold settings. It is particularly useful for evaluating models on imbalanced datasets. A high AUROC value indicates a good model, but it can be overly optimistic when assessing the performance of the minority class in extremely imbalanced scenarios.

5) The F1 *s*core is computed as the harmonic mean of precision and recall, providing a balanced measure between finding relevant instances and avoiding labeling negatives as positives.

Moreover, to ensure a more robust evaluation of the model, a 10-fold cross-validation scheme is employed. In this setting, the dataset is divided into ten equal parts, each used once for testing across ten different training cycles.

This approach guarantees that every dataset image contributes to training and testing, allowing for an extensive assessment of the model's performance. The results are averaged, providing a reliable measure of the model's effectiveness (Lu 2010). For each fold, the performance metrics are calculated based on the confusion matrix, which consists of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The metrics are then averaged across all folds to obtain the final estimates.

For each fold i, precision Pi is calculated as:

$$Pi = \frac{TPi}{TPi + FPi}$$

The final precision *P* is the average of precision values across all folds:

$$P = \frac{1}{k}\sum_{i=1}^{k} Pi$$

For each fold i, recall Ri is calculated as:

$$Ri = \frac{TPi}{TPi + FNi}$$

The final recall R is the average of recall values across all folds:

$$R = \frac{1}{k}\sum_{i=1}^{k} Ri$$

For each fold i, the F1 score F 1i is calculated as:

$$F1i = 2\frac{Pi \cdot Ri}{Pi + Ri}$$

The final F1 score F 1 is the average of F1 scores across all folds:

$$F1 = \frac{1}{k}\sum_{i=1}^{k} F1i$$

For each fold i, the ROC curve is generated, and the AUROC Ai is calculated. The final AUROC A is the average of AUROC values across all folds:

$$A = \frac{1}{k}\sum_{i=1}^{k} Ai$$

### 4.3 *Results*

The results obtained with ViT-B-16 on the NuragAI dataset are highly promising, as shown in Tab. 1. Across all ten cross-validation folds, the model demonstrated exceptional performance in all the tested metrics. The precision of the model achieved near-perfect scores in every fold, with an average of

| Metric | #Fold | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
| Precision (%) | 100,0 | 100,0 | 99,65 | 100,0 | 100,0 | 100,0 | 100,0 | 99,65 | 100,0 | 100,0 | 99,93 |
| Recall (%) | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| Accuracy (%) | 100,0 | 100,0 | 99,99 | 100,0 | 100,0 | 100,0 | 100,0 | 99,99 | 100,0 | 100,0 | 99,99 |
| AUROC (%) | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| F1 score (%) | 99,62 | 99,37 | 99,37 | 99,62 | 99,62 | 99,37 | 99,00 | 99,75 | 99,37 | 99,62 | 99,47 |

Tab. 1 – Benchmark results for the model on the NuragAI dataset across all cross-validation folds. Higher scores indicate better performance.

99.93%. This suggests that almost every image classified by the model as containing a nuraghe was correct, showing minimal false-positive errors.

Such strikingly high results undoubtedly require a cautious approach and reflection: it may indeed mean that the two selected datasets (presence and absence) exhibit excessive internal uniformity. However, although field conditions may differ from those of an experimental dataset, such a high result on the analysis sample is certainly promising for the model.

Analyzing the parameters in detail: Recall maintained a perfect score of 100% across all folds; this indicates that the model successfully identified all nuraghes in the images, missing no true positives. The Accuracy of the model, reflecting the overall correctness of both positive and negative predictions, was consistently high with an average of 99.99%. The AUROC, a metric that assesses the model's ability to distinguish between classes, stood at a perfect 100% across all folds. This indicates exceptional model performance in effectively separating the classes. Lastly, the F1 score, which balances Precision and Recall, averaged at 99.47%. While slightly lower than the other metrics, it still indicates a strong harmonic mean between precision and recall. In this specific case, it could be considered one of the most relevant metrics due to the unbalanced dataset.

While the absolute F1 difference appears small (<1%), k-fold cross-validation and modern statistical methods help validate meaningful improvements, as proved in numerous works (Kohavi 1995; Dietterich 1998; Bengio, Grandvalet 2004). In summary, we could interpret metric interplay in context according to these 3 findings:

– AUROC vs. Accuracy/Recall/Precision: The perfect AUROC reinforces that the model's class separation is optimal, even as precision and recall individually achieve saturation. This suggests the model not only captures all positives but also maintains minimal overlap with negatives, a rare combination in imbalanced tasks.

– F1 as a critical metric for imbalance: While accuracy (99.99%) appears near-perfect, it can be misleading in highly skewed datasets due to dominance by the majority class. The F1 score's focus on harmonic mean ensures that both precision and recall are explicitly prioritized, a necessary approach when missing positives or falsely flagging negatives could have domain-specific consequences (e.g., false alarms in archaeological surveys).

– Cross-validation as a safeguard: the consistency of metrics across all ten folds strengthens confidence in these results, mitigating concerns about sampling bias or data leakage. Modern statistical validation techniques further support the conclusion that improvements are not spurious.

Also, while the individual metrics (Precision, Recall, AUROC) highlight near-ideal performance, their collective behavior underscores ViT-B-16's robustness under extreme class imbalance. Notably, this combination of metrics, simultaneous saturation in AUROC and class-imbalance-aware F1, sets a new benchmark for archaeological object detection, where balancing false negatives (missed structures) and overinterpretation is paramount.

### 4.4 *Ablation study*

In this work, an ablation study to evaluate and select the most effective model architecture for our framework has been conducted. This study is crucial for ensuring that the final framework employs the optimal model for the task. The findings of the ablation study are presented in Tab. 2. It displays the performances of various well-known models on the NuragAI dataset across a 10-fold cross-validation scheme. Each column under '#Fold' corresponds to one of the ten parts of the dataset used in the cross-validation process. The rows list the different AI model architectures tested: VGG-11 (Simonyan, Zisserman 2015), MobileNet v3 (Howard *et al.* 2019) in both Small and Large versions, ResNet-18, ResNet-101 (He *et al.* 2015), and ViT-B-16 (Dosovitskiy *et al.* 2021).

| | **#Fold** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **Mean** |
| **Model** | **F1 score (%)** | | | | | | | | | | |
| VGG-11 (Simonyan, Zisserman 2015) | 97,91 | 97,75 | 97,15 | 98,72 | 97,21 | 97,27 | 98,70 | 98,02 | 97,25 | 97,05 | 97,70 |
| MobileNet v3 S (Howard *et al.* 2019) | 92,80 | 90,43 | 90,48 | 92,39 | 91,04 | 93,36 | 85,25 | 88,35 | 89,65 | 90,19 | 90,39 |
| MobileNet v3 L (Howard *et al.* 2019) | 97,86 | 97,28 | 97,23 | 98,75 | 97,81 | 97,14 | 98,92 | 98,01 | 97,71 | 97,38 | 97,81 |
| ResNet-18 (He *et al.* 2015) | 97,90 | 96,28 | 97,66 | 98,65 | 98,64 | 98,85 | 98,46 | 98,84 | 97,28 | 99,03 | 98,16 |
| ResNet-101 (He *et al.* 2015) | 98,74 | 98,14 | 98,26 | **99,90** | **99,89** | 98,68 | **99,96** | 99,60 | 98,29 | 98,80 | 99,03 |
| ViT-B-16 (Dosovitskiy *et al.* 2021) | **99,62** | **99,37** | **99,37** | 99,62 | 99,62 | **99,37** | 99,00 | **99,75** | **99,37** | **99,62** | **99,47** |

Tab. 2 – Ablation study on the F1 score of different models on the NuragAI dataset on all the cross-validation folds. Higher is better. Bold values are the best in their columns.

The Table's cells reveal the F1 score of each model in each fold. Notably, the last column, 'Mean', provides the average score of each model across all folds, offering a summarized view of the model's overall performance. The bolded values in the table highlight these top performances. ViT-B-16 stands out with its consistently high accuracy across the entire cross-validation process. As indicated by its highest overall mean F1, this consistent performance suggests that ViT-B-16 is particularly well-suited for this task.

## 5. AIOH takes the field

The satisfactory results achieved on the original dataset require further experimentation at a broader territorial scale in order to assess the algorithm's ability under real-world conditions. However, extending the analysis to the entire regional territory would entail excessive computational effort and substantial verification work by archaeologists. For this reason, a randomly selected sample area was chosen as an intermediate step between controlled experimentation and full field application. This transition introduces several complexities: environmental variability, the large number of known archaeological remains, an even greater number of cells identified as potentially favorable, and the uncertainty surrounding the exact location of many nuraghes. Addressing these factors is nonetheless essential to evaluate the system as a practical support tool for archaeological fieldwork.

To minimize bias, the selection of the test area was entrusted to a researcher not involved in the project. A sector of north-western Sardinia measuring

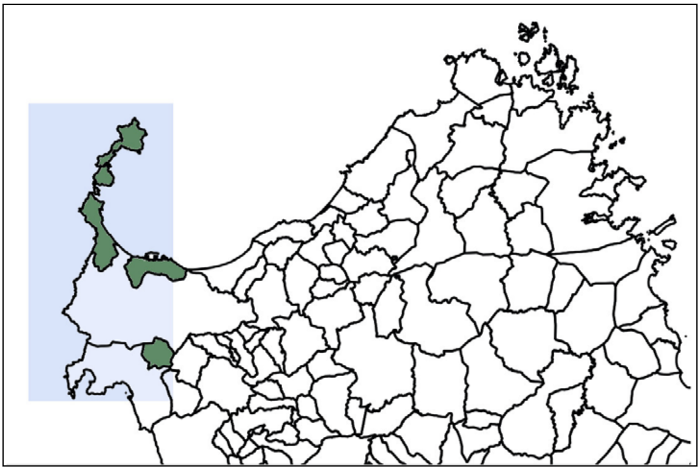

Fig. 6 – Map of the Northern Sardinia municipalities, with the bounding box of the square sample test. The three areas chosen for the analysis (Olmedo, Stintino and Porto Torres) are filled in green.

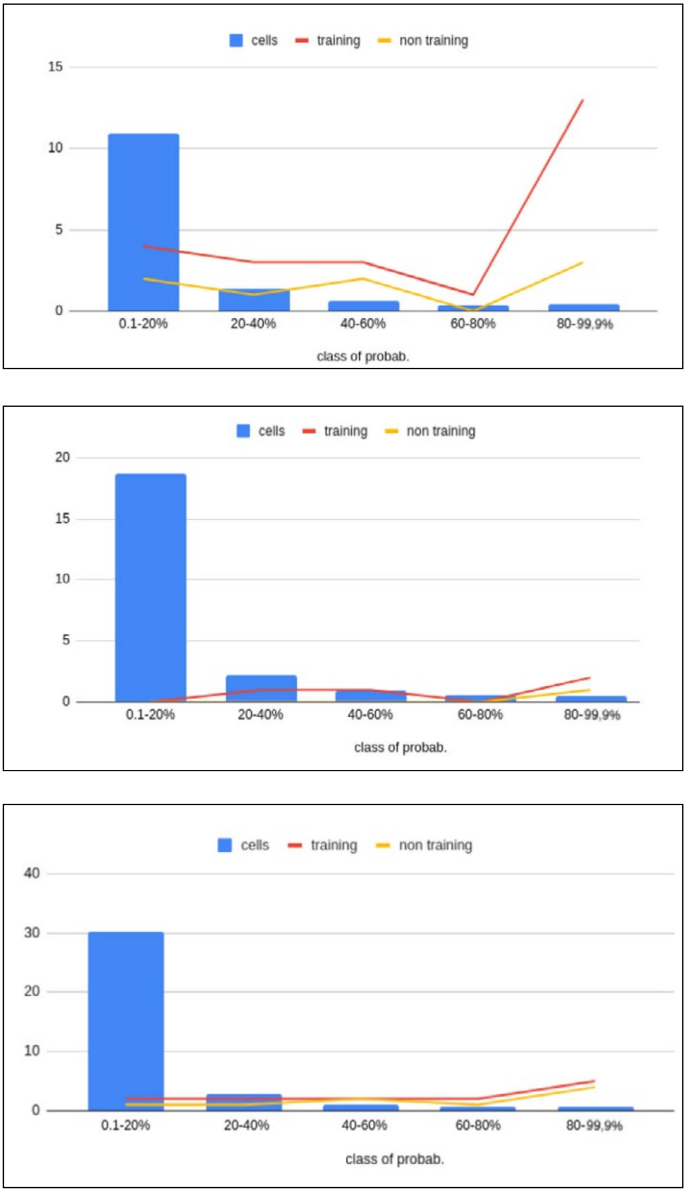


Fig. 7 – Distribution of nuraghes known in literature for the 3 districts (from left to right: Olmedo, Stintino, Porto Torres), including (red line) or non-including (yellow line) the ones used for training, in comparison to the amount of cells (blue bars) for each probability category. The distribution seems to be proportional to the probability, despite the number of cells.

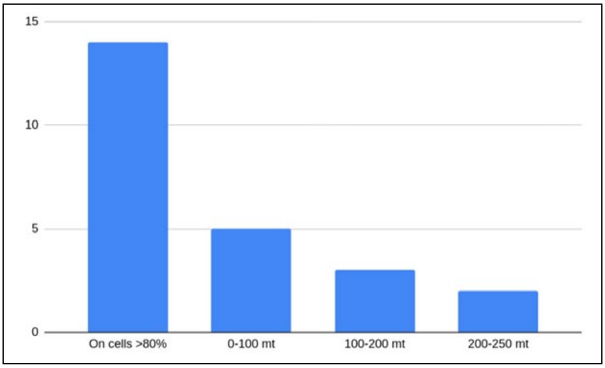


Fig. 8 – Distribution of nuraghes precisely located in the Municipalities of Olmedo, Stintino and Porto Torres, in relation to cells > 80% probability: one cell, up to 100, 200, 250 m.

approximately 60×30 km was identified. Within this quadrilateral, three municipalities were selected – Olmedo, Porto Torres, and Stintino – chosen for their manageable size and substantial inclusion within the designated area (Fig. 6). Unlike previous phases, where training and test datasets were strictly separated, this large-scale application inevitably includes cases previously used for training. This overlap is acknowledged and considered in the interpretation of results.

For each cell in the grid, the algorithm produces a probability value (ranging from 0.1% to 99.9%) estimating the likelihood of the presence of a nuragic monument, or more broadly, a nuragic area. Importantly, conventional performance metrics from the training phase are not applied here, since the aim is not to present this territory as a fully validated benchmark. Rather, the objective is to demonstrate the method's capacity to identify areas of archaeological interest in previously unseen contexts, interacting with expert archaeological evaluation.

An initial comparison between predicted probabilities and the recorded distribution of nuraghes (based on the SardegnArcheologica dataset) requires caution. In Olmedo, 10 of the 24 recorded nuraghes had been included in training; in Stintino, 3 of 4 monuments were part of the training dataset (in the latter case, anyway, the only independent monument 'Monte Atene' falls within a cell assigned a 95% probability); these results are therefore not entirely independent validations. More informative is the case of Porto Torres, where only 4 of 13 nuraghes were used in training. Here, the distribution of monuments tends to cluster within higher-probability classes regardless of training inclusion (Fig. 7). A Pearson-Bravais correlation test between the number of cells and the number of nuraghes yields values between 0.3 and –0.3, indicating no strong linear correlation but suggesting a non-random distribution.

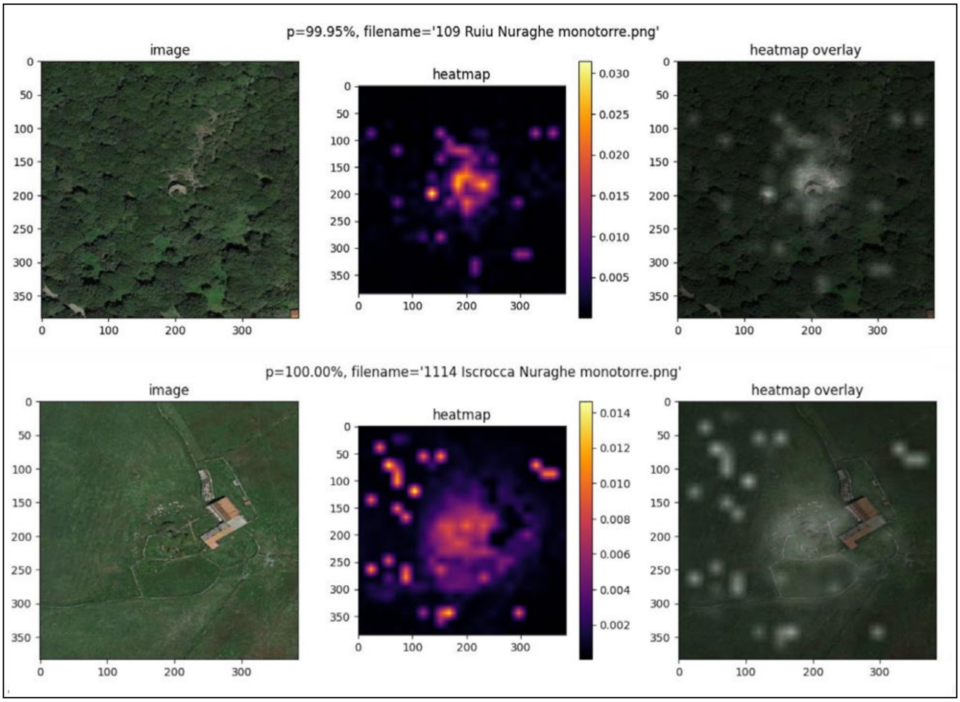


Fig. 9 – Sample results on test data (output heat map): the upper part shows the output heatmap of the ViT on an image containing a monotorre nuraghe. The model focuses on the scene's specific features. The lower part shows the output heatmap of the ViT on an image containing a nuraghe near a building. The main magnitude is on the correct location of the ancient ruins, ignoring the building.

Some monuments appear in lower-probability cells; in many such cases, however, the structures are poorly preserved or inconsistently geolocated across sources, raising doubts about positional accuracy. A particularly significant observation emerges from spatial proximity analysis: all precisely located nuraghes within the three municipalities, even those situated in low-probability cells, are found within 250 m of cells with probability values exceeding 80% (Fig. 8). This pattern suggests that the algorithm may capture not only the physical monument, but also broader environmental characteristics associated with nuragic presence.

This interpretation is supported by qualitative analysis of the Vision Transformer (ViT) outputs. Heatmaps generated for selected cases indicate that classification decisions are influenced not solely by the monument itself, but also by its surrounding landscape (Fig. 9). The system thus appears sensitive to environmental and contextual features, reinforcing the hypothesis that nuraghes were strategically positioned within broader territorial networks.

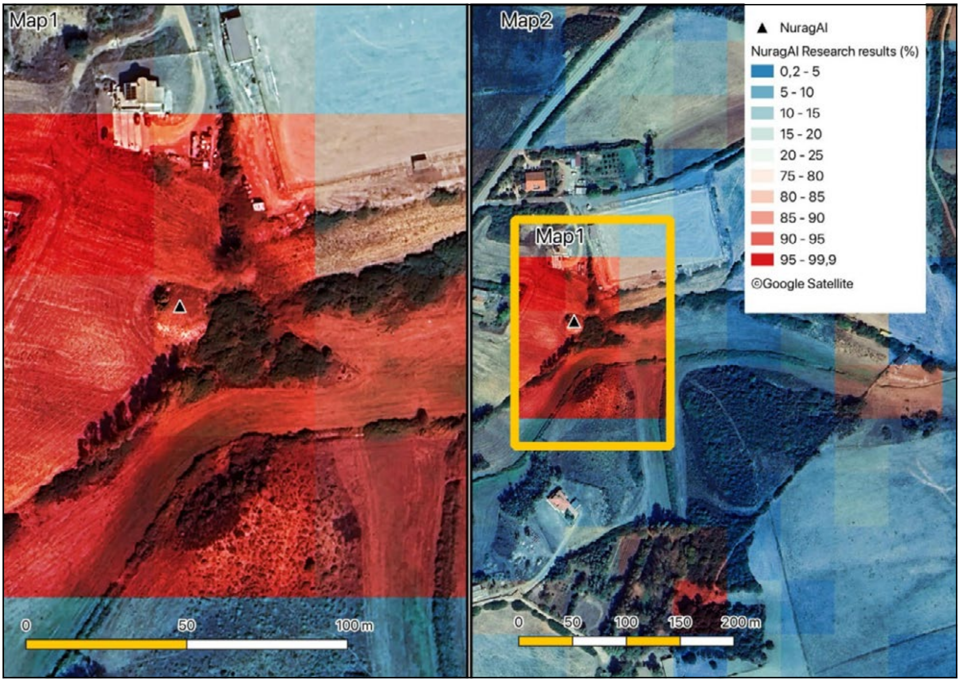


Fig. 10 – NuragAI research results.

This aligns with established spatial analysis models in archaeology (Hodder, Orton 1976) and with recent studies emphasizing environmental determinants in settlement selection (Cicilloni *et al.* 2015; Minchilli, Tedeschi 2017; Mariani *et al.* 2021; Malavasi *et al.* 2023) (Fig. 10).

These considerations lead to the proposal of adopting the concept of 'nuragic area', defined as a zone characterized by environmental conditions favorable to the presence of nuragic monuments. The algorithm's performance, therefore, may reflect its capacity to identify such environmentally suitable areas rather than exclusively detecting architectural remains.

The difference between the very high accuracy achieved in the controlled experimental phase (99%) and the more moderate results observed in this field-oriented test can be attributed to multiple factors: the heterogeneity of nuragic evidence, varying states of preservation, partial visibility in satellite images, and inconsistencies in documentation. Furthermore, several high-probability cells correspond to indications found in historical cartography produced by the Italian Military Geographic Institute (IGM), although not recorded in the SardegnArcheologica dataset. These cases highlight the importance of integrating multiple documentary sources and suggest promising directions for future verification through comparison with historical maps and on-site investigation.

Overall, this experiment demonstrates the methodological robustness of the approach while emphasizing the need for careful interpretation when moving from controlled datasets to complex archaeological landscapes.

## 6. On-site experiment

This work is intended to describe the approach to researching archaeological monuments through AI algorithms, exploring all theoretical aspects and postponing field application until the model's robustness has been strengthened. However, the opportunity for an initial, indicative, and absolutely preliminary test seemed important, also for comparison with practical problems. To this end, a series of 24 cases were selected from those contained in the high-potential cells, in the provinces of Alghero, Olmedo, Sassari, Stintino, and Porto Torres (Fig. 11), based on a desk-based optical assessment; and an experimental field test was conducted. The field investigation has thus far been a quick survey, limited to a surface investigation of visible evidence. Lacking specific authorization, we could not perform actions like surface excavation, material collection or entrance in private areas. As a consequence, the investigation was limited to publicly accessible areas and visible evidence. For this reason, the locations explored were first classified into a series of schematic categories based on accessibility: 'non-verifiable', which were not accessible and inspected; 'remotely verified', which were not accessible but were assessed through remote visual analysis; 'verified', where the location could be accessed.

For the same reasons, excluding unverified cases, the judgment on the actual presence of a structure of archaeological significance was expressed in terms of probability, based on an assessment of the visible external characteristics. The defined categories are therefore: 'probable'; 'possible'; 'negative'; 'negative +'. The 'negative+' refers to negative aspect but with archaeological traces in the immediate vicinity. The 'negative' label refers to cases that have been determined to be invalid, sometimes including elements that may have misled the model (for example, a pile of landfill covered with rocks or the drystone fence of a recently created garden). The 'possible' label refers to cases where no conclusive traces were found, despite the presence of elements, such as concentrations of rocks, that make the archaeological hypothesis plausible. The 'probable' cases, on the other hand, refer to contexts where elements have been found that suggest human intervention, such as the presence of traces of pottery on the ground, hammer blows on rocks, etc.

The survey result has been:

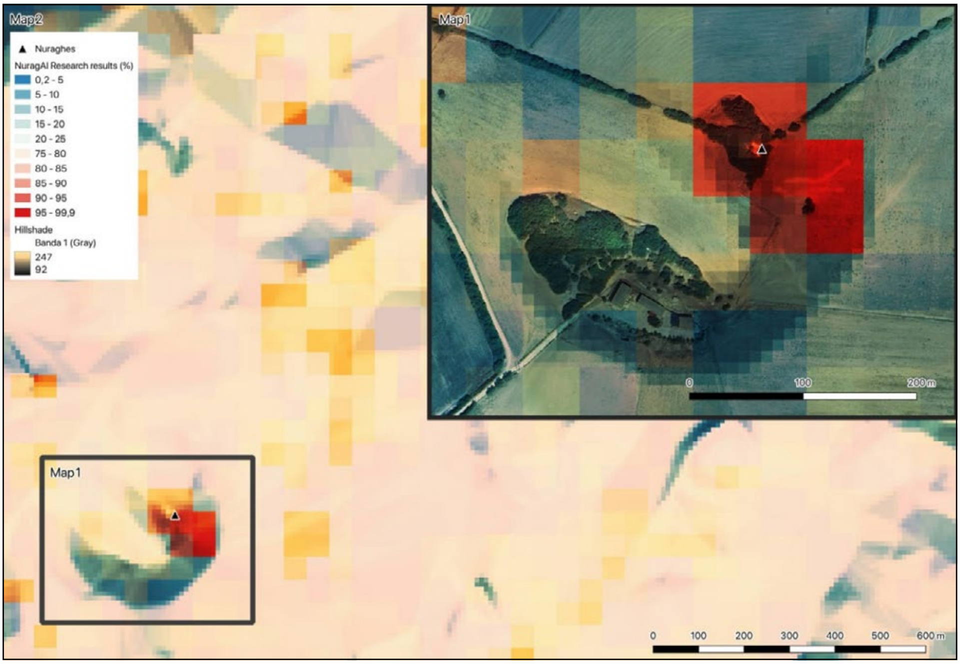


Fig. 11 – Identification of new nuraghes on particular geomorphological contexts.

– probable: 4 items;
– probable (but remotely verified): 2 items;
– possible: 3 items;
– negative: 8 items:
– negative +: 2 items;
– non-verifiable: 5 items.

Finally, narrowing the analysis to the six probable cases (4+2), none of the sites were previously recorded in the archaeological documentation of the area, at least not within the thematic maps of interest of the Regional Technical Service (https://www.regione.sardegna.it/servizi/accesso-ai-servizi-on-line/indice-dei-servizi/sitr-sistema-informativo-territoriale-regionale), nor are there any documented structures within a short distance. In only one case out of six, a nuragic structure approximately one hundred meters away ('Sa Lattara') is documented in the SardegnArcheologica database. Based on this necessarily cursory test, the method illustrated was deemed to be of interest for archaeological analysis, and discussions are underway with the relevant local authorities to further explore this line of research.

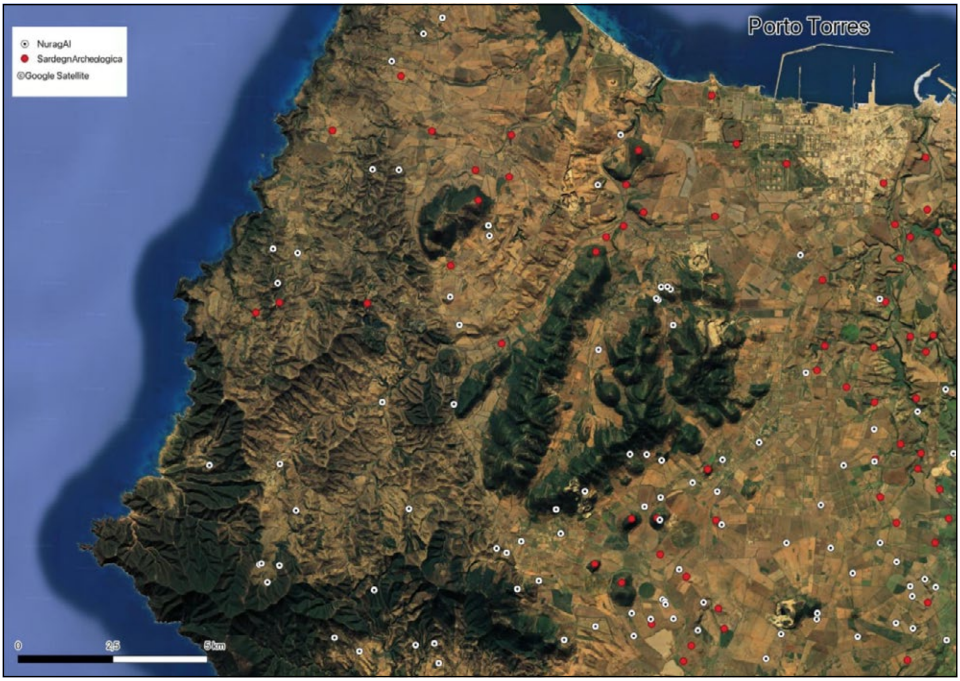


Fig. 12 – Distribution map of high-ranked 'nuraghe areas' found by AI (black) and nuraghes previously mapped by SardegnArcheologica (red) in the area of Porto Torres.

## 7. Conclusions and perspectives

The present work constitutes a first step towards creating a useful support system for archaeological research. The exceptional performance of ViT-B-16 on the NuragAI dataset, evidenced by near-saturation in precision, recall, and AUROC, demonstrates its potential for archaeological detection tasks, where minimizing false negatives is critical. The harmonious alignment of these metrics under k-fold cross-validation suggests that the model's robustness is not dependent on specific data partitions but stems from effective architectural choices and augmentation strategies tailored to class imbalance. However, the slight divergence in F1 score highlights remaining challenges in reconciling extreme precision-recall balance for rare classes, a consideration that underscores the importance of dataset diversity in future work. These results affirm ViT-B-16's adaptability to real-world archaeological constraints and provide a benchmark for evaluating domain-specific adaptations of vision models. Thus, the ViT shows significant potential to advance the field in this usage context.

Although the effectiveness of systematic application across the territory is not as high as that of the training dataset, preliminary tests have proven effectiveness in localizing documented monuments and identifying potential new, unrecorded contexts of interest. Through the algorithm results, it would also be possible to comprehend possible relations not fully considered by the archaeological scientific community. Although the DL algorithm, as a 'black box', makes it difficult to clearly identify the effective similarity elements, it was possible to observe a frequent geomorphological pattern, with elevated hills and plateau (Fig. 12). It seems to confirm environmental implications of nuragic location choices and led the authors to adopt the concept of 'nuragic area'. Some of such areas, identified by the model and checked by human sight, revealed the possible presence of not-recorded monuments whose real nature, to be correctly interpreted, requires a careful survey campaign on the territory: a first attempt in this direction, necessarily rapid and superficial, nevertheless provided interesting results and allowed the start of interactions with local institutions.

At a Mediterranean level, the methodology that the project is developing and refining may allow the declination of the recognition processes in other sectors and chronological periods, which suffer from the same problems identified for the nuraghes, through different and specific training processes. A coherent and robust methodology will also be able to accompany the recognition operations to provide for projects in order to help small local communities in enhancing their territory with a view to the creation of local and tourist valorization actions centered on the widespread of monumental heritage of the territory (Caruso *et al.* 2023).

Augusto Palombini*
Istituto di Scienze del Patrimonio Culturale - CNR
augusto.palombini@cnr.it

Edoardo Baiocchi
Independent Researcher
baiocchiedoardo@gmail.com

Saverio Giulio Malatesta
Centro interdipartimentale di ricerca DigiLab, Sapienza Università di Roma
saveriogiulio.malatesta@uniroma1.it

Paolo Rosati
Progetto ArcheoSITAR
archeorosati@gmail.com

Romeo Lanzino, Marco Raoul Marini
Dipartimento di Informatica, Sapienza Università di Roma
lanzino@di.uniroma1.it, marini@di.uniroma1.it

* Corresponding Author

## REFERENCES


AA.VV. 1922, *Elenco degli edifizi monumentali LXIX Provincia di Sassari (p. 150)*, Ministero della Pubblica Istruzione (https://www.bdl.servizirl.it/bdl/bookreader/index.html?path=fe&cdOggetto=2049#page/12/mode/2up).

Antona A. 2013, *Il complesso nuragico La Prisgjona in regione Capichera (Arzachena)*, in L. Usai (ed.), *Memorie dal sottosuolo. Scoperte archeologiche nella Sardegna centro-settentrionale*, Quartucciu, Scuola Sarda Editrice, 56-65.

Bai Y., Mei J., Yuille A.L., Xie C. 2021, *Are transformers more robust than CNNs?*, «Advances in Neural Information Processing Systems», 34, 26831-26843.

Barber M. 2011, *A History of Aerial Photography and Archaeology: Mata Hari's Glass Eye and Other Stories*, London, English Heritage.

Bartoloni P. 2000, *La necropoli di Monte Sirai*, Roma, CNR.

Bengio Y., Grandvalet Y. 2004, *No unbiased estimator of the variance of k-fold cross-validation*, «Journal of Machine Learning Research», 5, 1089-1105.

Caruso M., Leopardi L., Malatesta S.G. 2023, *AI and custom marketing for cultural initiatives in regional projects: Perspectives on the territorial integration of communities with tourism*, in *Proceeding of the Una Quantum 2022 Conference: Open Source Technologies for Cultural Heritage, Cultural Activities and Tourism (Rome 2022)* (https://doi.org/10.3390/proceedings2024096008).

Centurione A.M. 1886, *Studi recenti sopra nuraghi e loro importanza*, Cagliari, Tip. Giachetti.

Cicilloni R., Mossa A., Cabras M. 2015, *Studio dell'insediamento protostorico in un'area della Sardegna centro-occidentale tramite strumenti GIS ed analisi multivariate*, «Archeologia e Calcolatori», 26, 149-168 (https://doi.org/10.19282/ac.26.2015.23).

Cicilloni R., Migaleddu M. 2008, *Monumenti nuragici in territorio di Teulada (Cagliari): note preliminari*, in *La civiltà nuragica. Nuove acquisizioni*, in *Atti del Convegno (Senorbì 2000)*, Quartu Sant'Elena, Soprintendenza per i beni archeologici delle province di Cagliari e Oristano, 433-448.

Cossu T., Perra M., Usai A. 2018, *Il tempo dei nuraghi*, Nuoro, Ilisso.

Depalmas A. 1998, *Organizzazione ed assetto territoriale nella regione di Sedilo durante i tempi preistorici*, «Antichità Sarde», 3, 3, 33-76.

Dietterich T.G. 1998, *Approximate statistical tests for comparing supervised classification learning algorithms*, «Neural Computation», 10, 7, 1895-1923 (https://doi.org/10.1162/089976698300017197).

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. 2021, *An image is worth 16x16 words: Transformers for image recognition at scale*, «arXiv», 2010.11929v2 (https://doi.org/10.48550/arXiv.2010.11929).

Fadda N. 2016, *Il nuraghe Ola e i sistemi di organizzazione socio-territoriale nel Nuorese*, Sassari, Università di Sassari, 53-57.

Finocchi S. 2005, *Ricognizione nel territorio di Monte Sirai*, «Rivista di Studi Fenici», 33, 225-259 (https://doi.org/10.1400/88655).

Fiorucci M., Khoroshiltseva M., Pontil M., Traviglia A., Del Bue A., James S. 2020, *Machine Learning for Cultural Heritage: A survey*, «Pattern Recognition Letters», 133, 102-108 (https://doi.org/10.1016/j.patrec.2020.02.017).

Fredriksson T., Issa Mattos D., Bosch J., Olsson H. 2020, *Data labeling: An empirical investigation into industrial challenges and mitigation strategies*, «Lecture Notes in Computer Science», 12562, 202-216 (https://doi.org/10.1007/978-3-030-64148-1_13).

He K., Zhang X., Ren S., Sun J. 2015, *Deep residual learning for image recognition*, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas 2016)*, 770-778 (https://doi.org/10.1109/CVPR.2016.90).

Hodder I., Orton C. 1976, *Spatial Analysis in Archaeology*, Cambridge, Cambridge University Press.

Howard A., Sandler M., Chu G., Chen L.-C., Chen B., Tan M., Wang W., Zhu Y., Pang R., Vasudevan V., Le Q.V., Adam H. 2019, *Searching for MobileNetV3*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, Piscataway, IEEE, 1314-1324 (https://doi.org/10.1109/ICCV.2019.00140).

Iandola F.N., Han S., Moskewicz M.W., Ashraf K., Dally W.J., Keutzer K. 2016, *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*, «arXiv», 1602.07360 (https://doi.org/10.48550/arXiv.1602.07360).

Jamil A.H., Yakub F., Azizan A., Roslan S.A., Zaki S.A., Ahmad S.A. 2022, *A review on Deep Learning application for detection of archaeological structures*, «Journal of Advanced Research in Applied Sciences and Engineering Technology», 26, 7-14 (https://doi.org/10.37934/araset.26.1.714).

Kohavi R. 1995, *A study of cross-validation and bootstrap for accuracy estimation and model* selection, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2, Montréal, Morgan Kaufmann Publishers, 1137-1143.

Lambers K., Traviglia A. 2016, *Automated detection in remote sensing archaeology: A reading list*, «AARGNews», 53, 25-29.

Lilliu G. 1982a, *La civiltà nuragica*, Sassari, Delfino Editore.

Lilliu G. 1982b, *Sardegna nuragica*, Nuoro, Il Maestrale.

Loshchilov I., Hutter F. 2019, *Decoupled weight decay regularization*, in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019* (https://openreview.net/forum?id=Bkg6RiCqY7).

Lu Z. 2010, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Oxford, Oxford University Press.

Malavasi M., Bazzichetto M., Bagella S., Barták V., Depalmas A., Gregorini A., Sperandii M.G., Acosta A.T.R., Bagella S. 2023, *Ecology meets archaeology: Past, present and future vegetation-derived ecosystems services from the Nuragic Sardinia (1700-580 BCE)*, «People and Nature», 5, 3, 938-949.

Mariani G.S., Brandolini F., Melis P. 2021, *Landscape and geology as controls on Bronze Age human dispersal: A case study from Sardinia (Italy)*, «Journal of Maps», 18, 4, 598-609 (https://doi.org/10.1080/17445647.2021.1999339).

Maxia M. 2020, *I nomi di 8000 nuraghi*, Cagliari, NOR Edizioni.

Melis P. 2007, *Preistoria e protostoria nel territorio di Castelsardo*, in A. Mattone, A. Soddu (eds.), *Castelsardo. 900 anni di storia*, Roma, Carocci, 15-61.

Minchilli M., Tedeschi L.F. 2017, *The geographical distribution of nuraghi in north-western Sardinia: Analysis and evaluation of the influence of anthropic and natural factors*, in O. Gervasi, B. Murgante, S. Misra, G. Borruso, C.M. Torre, A.M.A.C. Rocha, D. Taniar, B.O. Apduhan, E. Stankova, A. Cuzzocrea (eds.), *Computational Science and Its Applications, ICCSA 2017*, Cham, Springer International Publishing, 95-108.

Moravetti A. 2000, *Ricerche archeologiche nel Marghine-Planargia*, Sassari, Carlo Delfino Editore.

Müller S.G., Hutter F. 2021, *TrivialAugment: Tuning-free yet state-of-the-art data augmentation*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, IEEE/CVF, 774-782 (https://doi.org/10.1109/ICCV48922.2021.00081).

Palombini A. 2001, *Miti e pregiudizi nell'uso dei sistemi di classificazione in archeologia*, «Rassegna di Archeologia Preistorica e Protostorica», 10, 45-68.

Parcak S. 2009, *Satellite Remote Sensing for Archaeology*, London, Routledge.

Sherman G. 2014, *The PyQGIS Programmer's Guide, Extending QGIS 2.x with Python*, Locate press.

Simonyan K., Zisserman A. 2015, *Very deep convolutional networks for large-scale image recognition*, «arXiv», 1409.1556v6 (https://doi.org/10.48550/arXiv.1409.1556).

Vacca S., Aru A., Baldaccini G. 1998, *Rapporti tra suoli e insediamenti nuragici nella regione del Marghine-Planargia (Sardegna centro-occidentale)*, in C. Albore Livadie, F. Ortolani (eds.), *Il sistema uomo-ambiente tra passato e presente*, Bari, Edipuglia, 27-36.

Van Dommelen P. 1998, *On Colonial Grounds. A Comparative Study of Colonialism and Rural Settlement in First Millennium BC West Central Sardinia*, Leiden, University of Leiden.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. 2017, *Attention is all you need*, in I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 30, Red Hook, Curran Associates, Inc.

Velli A. 2020, *Dinamica y variaciones en la ocupacion territorial y relacion hombre-ambiente desde el Neolitico hasta la Edad del Bronce en el Bajo Sulcis Costero (Provincia de Carbonia-Iglesias, Cerdena)*, Granada, Universidad de Granada.

## ABSTRACT

Detecting archaeological features from satellite images is an increasingly vital tool in archaeology, offering a non-invasive means to discover and protect ancient structures. Among these, nuraghes, which are tower-shaped stone structures from the Bronze Age unique to Sardinia (Italy), present a significant challenge due to their varied sizes, forms, and environmental contexts. This paper introduces a pioneering approach employing an Artificial Intelligence (AI) driven method for detecting nuraghes in satellite images, specifically employing a Deep Learning (DL) model based on Vision Transformers (ViT), which is an advanced neural network architecture that excels in processing image data. The system, called 'Artificial Intelligence Operator for Heritage' (AIOH), is trained on custom data retrieved for this work: the dataset, NuragAI, includes thousands of satellite images from Google Maps with manually annotated nuraghe presence. It is designed to train the model to identify subtle and complex characteristics of nuraghe structures. The effectiveness of this approach was quantitatively assessed through extensive validation, with very high results on the training and testing set and interesting results also on data, detecting nuraghes in various terrains and under different lighting conditions. A preliminary on-site field campaign, to test results, was also carried out, confirming the model efficiency.