ARTICLE

Check for updates

A deep-learning model for predictive archaeology and archaeological community detection

Abraham Resler[1], Reuven Yeshurun[2], Filipe Natalio [iD] [3✉] & Raja Giryes[1✉]

Deep learning is a powerful tool for exploring large datasets and discovering new patterns. This work presents an account of a metric learning-based deep convolutional neural network (CNN) applied to an archaeological dataset. The proposed account speaks of three stages: training, testing/validating, and community detection. Several thousand artefact images, ranging from the Lower Palaeolithic period (1.4 million years ago) to the Late Islamic period (fourteenth century AD), were used to train the model (i.e., the CNN), to discern artefacts by site and period. After training, it attained a comparable accuracy to archaeologists in various periods. In order to test the model, it was called to identify new query images according to similarities with known (training) images. Validation blinding experiments showed that while archaeologists performed as well as the model within their field of expertise, they fell behind concerning other periods. Lastly, a community detection algorithm based on the confusion matrix data was used to discern affiliations across sites. A case-study on Levantine Natufian artefacts demonstrated the algorithm's capacity to discern meaningful connections. As such, the model has the potential to reveal yet unknown patterns in archaeological data.

---

[1] School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel. [2] Zinman Institute of Archaeology, University of Haifa, Mt. Carmel, Haifa, Israel.
[3] Kimmel Center for Archaeological Science, Weizmann Institute of Science, Rehovot, Israel. ✉email: filipe.natalio@weizmann.ac.il; raja@tauex.tau.ac.il

Introduction

Archaeology, broadly defined, is the study of the human past through material remains: artefacts of various materials (e.g., stone, bone, pottery, metal, glass) that were manufactured, used, and discarded by ancient societies (Murray and Evans, 2008; Renfrew and Bahn, 2013). The first and most basic task of the field's practitioners is to properly classify the numerous artefacts they encounter, determining their date, cultural attribution, form, function, socio-economic significance, and other features (Arkadiev, 2020; Dunnell, 1993; Hermon et al., 2004; Krieger, 1944; Whittaker et al., 1998). Such classifications often depend on prior knowledge, expertise, and preference for certain visual criteria over others (Barcelo, 1995).

In order to automate this process and utilise computers' excellent pattern recognition capabilities, efforts have been made to incorporate computer applications into the processes of archaeological classifications (Derech et al., 2021; Tal, 2014). Notable among these are experimentations with machine learning models—computer algorithms that learn from data how to automatically detect patterns and make accurate decisions (Mitchell, 1997; Bishop, 2006; Duda and Hart, 1973). Several attempts were made to apply machine learning to archaeological materials (Barcelo, 2008, 2016; Barceló and Bogdanovic, 2015; Díez-Pastor et al., 2018; Macleod, 2018). However, at first, they relied on hand-crafted feature extraction, resulting in relatively poor performance measures (e.g., Boon et al., 2009). More recently, machine learning algorithms have been used to extract relevant features automatically. Thus, for instance, Agam et al. (2020) combined Raman spectroscopy with machine learning algorithms to quantitatively estimate different degrees of thermal alteration on flint artefacts.

Of particular interest is deep learning, and more specifically, Deep Convolutional Neural Networks (CNNs), which are commonly used to analyse images. CNNs were successfully applied to various computer vision tasks, as they can automatically extract features from input images (Cifuentes-Alcobendas and Domínguez-Rodrigo, 2019; He et al., 2016; Krizhevsky et al., 2017; Taigman et al., 2014). These features, also known as embeddings, are a set of numbers (1536 numbers in this case), that are later used by other computational layers, to classify/infer other useful information from input data. The features do not necessarily correspond to a realistic measure of the data, such as colour or shape. Applied to archaeological problems, CNNs have shown promise, successfully fulfilling tasks of ceramic classification (Itkin et al., 2019), periodic discrimination of lithic assemblages (Grove and Blinkhorn, 2020), and differentiation of bone surface modifications (Domínguez-Rodrigo et al., 2020). However, these experiments with CNNs focused on narrow ranges of materials and contexts, consequently failing to seriously confront the bewildering diversity of the archaeological circumstances and record.

Thus, in this paper, we seek to develop a CNN model able to navigate the full gamut of temporal and cultural diversity archaeology has to offer (Fig. 1). To do so, large publicly accessible repository of artefact photographs managed and maintained by the Israel Antiquities Authority (http://www.antiquities.org.il/t/default_en.aspx) was used. It presents archaeological items that span a million and a half years of Levantine hominin history. The base CNN was initially trained to classify everyday objects on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset (Russakovsky et al., 2015), which is a large dataset of natural images. Then, following some modifications to the CNN, the model was trained to identify archaeological artefacts according to period and site (Fig. 1a, b). Next, drawing on

the model's acquired capacity to correctly classify artefacts, it was determined whether it can be effectively used to detect communities (Fig. 1c)—cohorts of classes with a meaningful common denominator. Finally, a case-study is offered on communities found from Natufian culture (ca. 15,000–11,700 years ago) classes in the Levant, showing that this method found archaeologically meaningful similarities between different sites.

In this manner, CNN was applied to this diverse archaeological dataset. First, it was assessed whether it could predict artefact's site and period using its image. Second, the possibility of finding other similar objects for a query image was investigated. Third, based on the results that the model made "correct confusion" (e.g., confusion between two different sites that are dated to similar archaeological period), the possibility of finding similarities between few sites was examined— which can potentially open up new avenues of analysis, research, and cultural interactions.

Below an account of archaeological dataset, procedures, and estimates of the model's performances is provided. Towards the end of the paper, detailed account of the methods employed in various parts of the workflow is presented. Additional experiments, information, and technical details are provided in the supplementary material.

Dataset

The dataset is publicly accessible on the Israel Antiquities Authority (IAA) website (http://www.antiquities.org.il/t/default_en.aspx). It comprises 12,364 photographs of 6770 artefacts that derive from across the southern Levant and span the Lower Palaeolithic (1.4 million years ago; Bar-Yosef and Goren-Inbar, 1993) and the Late Islamic (fourteenth century AD) periods. They include stone tools (e.g., blades, flakes, bifaces), bone tools (e.g., awls, beads, and pendants), metal objects (e.g., spearheads, coins), pottery vessels, and figurative art. Most of the artefacts presented are complete, and every item is designated according to its site and period of origin. The attribution of periods was provided by archaeologists working for or within the Israel Antiquities Authority (IAA) and available at their website.

Artefact categorisation by site and period produced a total of 555 classes (e.g., Early Bronze II Jericho, Iron II Akhziv) of various sizes. While some were hundreds of artefacts large, others comprised merely two or three (Fig. S1). In order to maintain a balanced dataset and provide sufficient conditions for statistical manipulations, the dataset was narrowed to the 200 largest classes, encompassing a total of 9909 images of 5450 artefacts, constituting 80.1% of the photographs and 80.5% of the artefacts (Table S1). Next, the dataset was split in two: one for training, comprising 8031 images (81%) of 4428 artefacts, and another for validation, comprising 1878 images (19%) of 1020 artefacts.

Standard image classification relies on visual similarities (dogs, cats, cars, or faces of different identities). However, in this case, similar artefacts may belong to different classes (Fig. 2a), and visually distinct artefacts may belong to the same class (Fig. 2b). Furthermore, note that temporally adjacent periods are likely to incorporate visually similar artefacts (e.g., Early Roman and Roman amphorae). Therefore, to test this model, two levels of temporal discrimination were established: rough- and fine-period groups. The fine temporal classification consisted of 21 groups, while the rough classification observed 13 (Table S1).

In order to facilitate the training process, the images' background and scale were standardised. All photographs were furnished with homogeneous white background, and the scale was removed (see Methods section for more details).
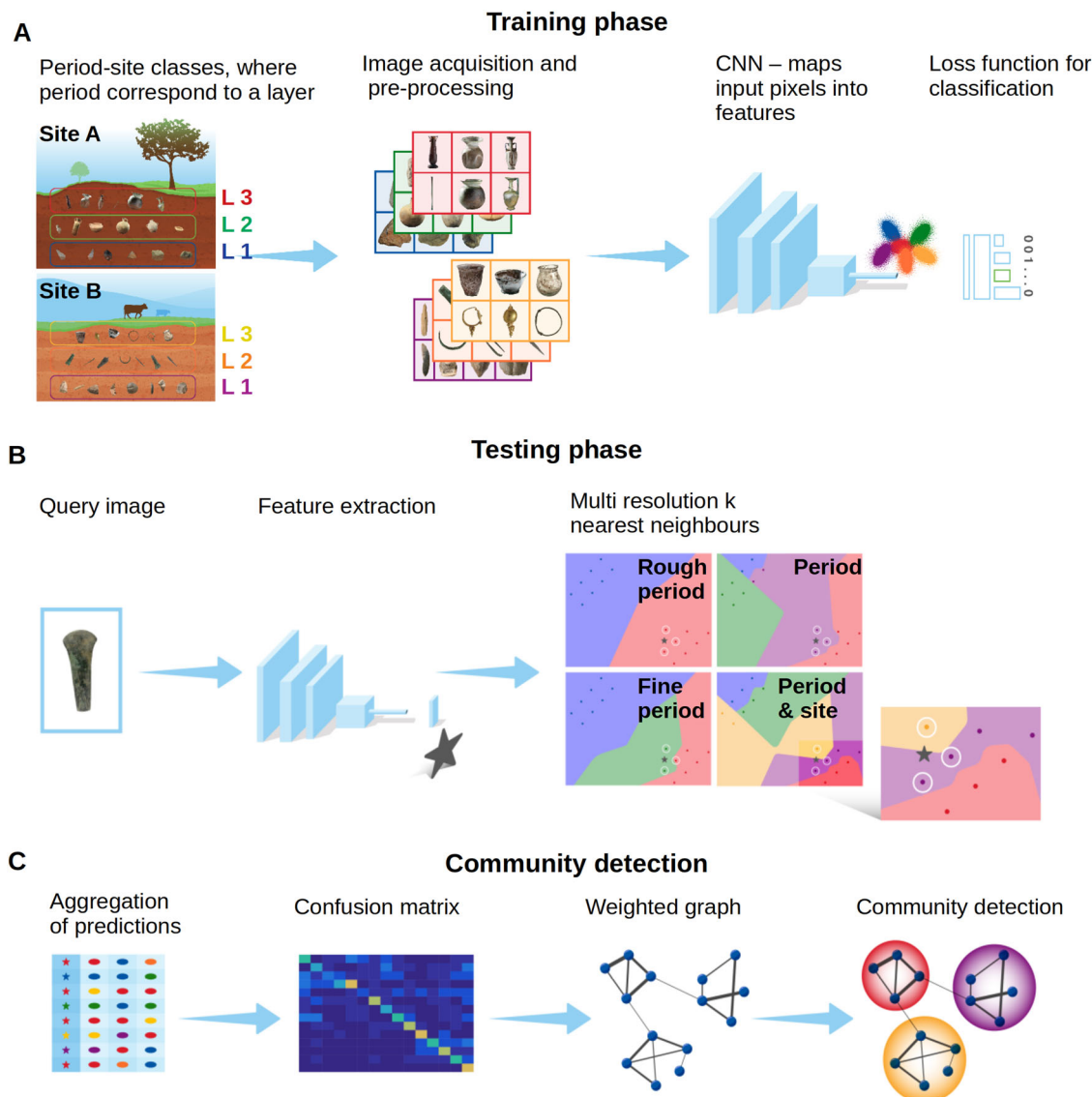
Fig. 1 A schematic representation of the machine learning-based workflow. A The training phase: a dataset of images of archaeological artefacts were grouped according to period and site, pre-processed, and used to train a Convolutional Neural Network (CNN). **B** The testing phase: the trained CNN was used to extract features from a query image and predict its class by identifying *k*-nearest neighbours in the training set. **C** Community detection: validation set predictions are aggregated in a confusion matrix that is later transformed into a weighted graph and fed to a community detection algorithm.

Model construction

In order to optimise the CNN to the task of archaeological classification, the standard transfer learning procedure was followed. Transfer learning is usually used when the available database size for the target application is relatively small. In this case, in order to improve performance, a pre-trained CNN on another larger (unrelated) database is used as the starting point for the training process.

The CNN model was based on the ImageNet (Russakovsky et al., 2015) pre-trained image classification model EfficientNetB3 (Tan and Le, 2019), which was chosen for its superior performance (see below, methods). It was built by stacking many (hence deep) basic computation layers (convolutions, non-linearities, pooling, skip connections, etc.), striving to achieve the best balance between computation complexity and prediction accuracy. The model was pre-trained on the ImageNet ILSVRC dataset to predict an image's category (class) out of 1000 possibilities, and reached 81.6% Top-1 and

95.7% Top-5 prediction accuracy (Top-k classification score computes the number of times the correct label is among the top k labels predicted). More details on this model are found in (Tan and Le, 2019).

To perform the transfer learning, the original classification layers were removed and a customised classification layer was added (a fully connected layer, that transformed EfficientNetB3 embeddings, of size 1536, to 200 classes). To optimise the training, five models were trained with the same ImageNet initialisation, each generating a different feature vector, which we then used to produce a final feature vector. To improve robustness and enrich the database, a standard data augmentation techniques was applied. These include: random rotations, spatial shifts, zoom, and horizontal flips. All CNNs' layers were trained for 25 epochs (in each epoch, the model is trained on the entire training set), using the categorical cross-entropy loss function (most common loss function for classification tasks). Additional details can be found in the methods section below.
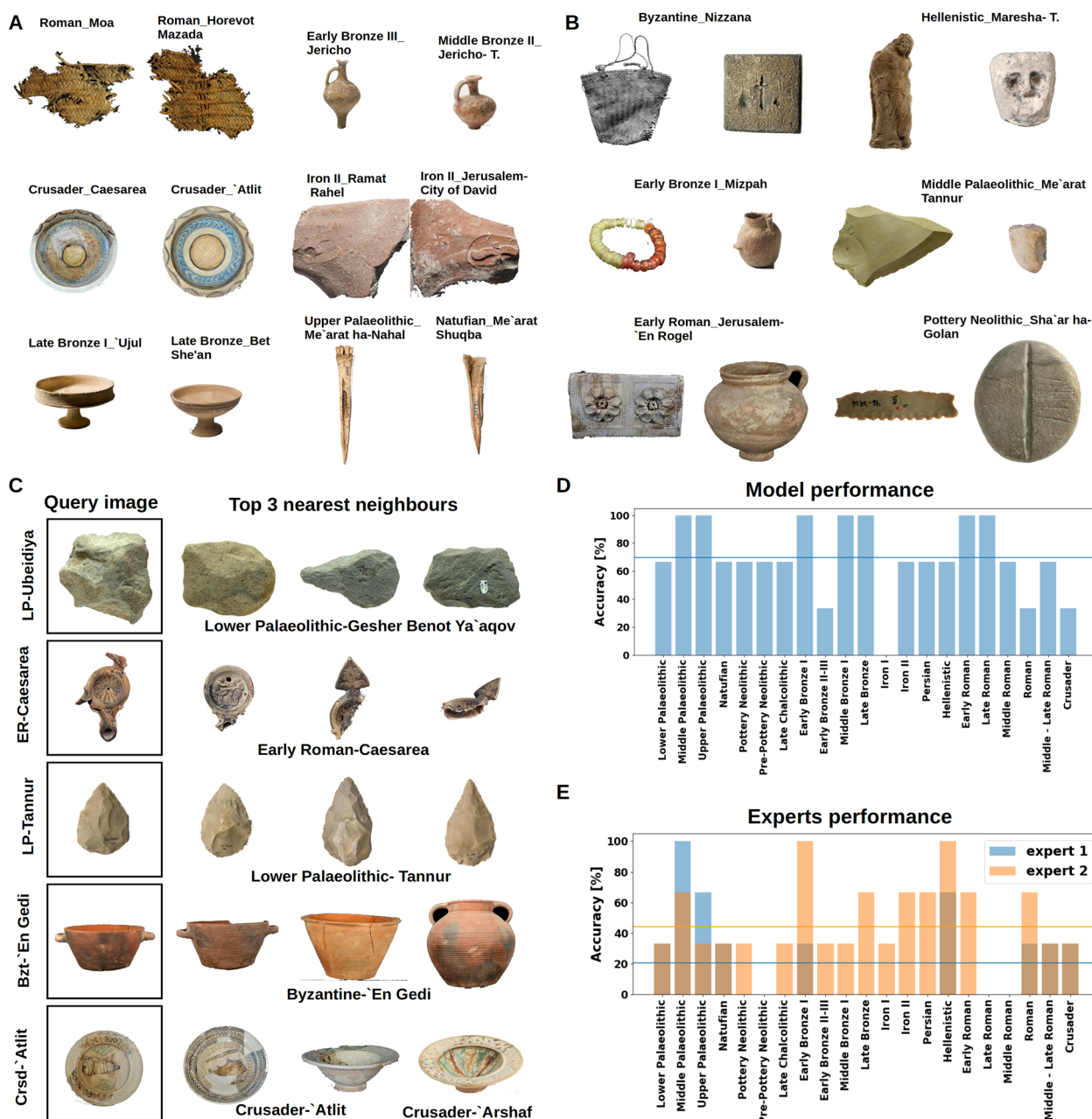
Fig. 2 Classification and CNN model performance. A Nearest neighbour pairs of artefacts from different classes (the image on the left derives from the validation set, and the image on the right derives from the training set). **B** Pairs of distinct images that derive from the same class. **C** Validation set query images (left column) and the top-3 training set nearest neighbours. **D** A histogram of model performance for fine-period prediction on 63 randomly picked images (3 images per period); the straight horizontal line marks the average prediction accuracy (69.84%). **E** A histogram of two archaeologists' performances (blind experiments) for the same 63 artefact images used for D; the horizontal lines mark the average prediction accuracies for each archaeologist (44.44, 20.63%).

Results

In the testing phase (Fig. 1b), the CNN was use as an "archaeological" feature extractor, and measured the archaeological dissimilarity between artefacts by calculating the cosine similarity distance (see methods below) between their feature vectors. Predictions were made by looking at query image's nearest neighbours, from the labelled training set. This procedure is illustrated in Fig. 2c that presents the three nearest neighbours for five query images. Interestingly, three sorts of outcomes are notable: (1) a complete match between the query image and the top three nearest neighbours (Early Roman Caesarea, Lower Palaeolithic Tabun, and Byzantine En Gedi), (2) a proximal

match between query and prediction, pertaining to site or period but not both (Lower Palaeolithic Ubeidiya), and (3) mixed results where some of the neighbours are a full match and others are proximal (Crusader Atlit) (see also Fig. S5).

The procedure above was used to measure accuracy on the validation set and can be used to classify other artefacts in the future (there was no other test set in the setup). Since each item in the dataset set had few labels: period/site/period-site/rough, fine-period group, accuracy on each one of these options is reported, regardless of the training process.

Table 1 shows the model's prediction accuracy for all possible labels on the validation set (i.e., period-site, site, period, fine, and

	Period-site	Site	Period	Fine-period grouping	Rough-period grouping
Table 1 Prediction accuracy [%] for period-site, site, period, fine-, and rough-period grouping.					
Top-1	58.10	63.58	67.79	71.03	76.36
Top-3	64.22	68.96	74.18	77.96	82.70
Top-5	67.36	71.89	77.69	81.47	85.41

rough-period accuracy grades). Accuracy values in this table were obtained after training with the standard period-site classification objective. Specifically, prediction accuracy was [%] of 58.10 (Top 1), 67.36 (Top-5) for period-site classes, and 76.36 (Top1), 85.41 (Top5) for rough-period groups. Model accuracy for each fine/rough-period group can be found in Fig. S2. The resulted confusion matrix and embeddings t-SNE visualisation (Van der Maaten and Hinton, 2008) can be found in Fig. S3 and Fig. S4, respectively.

Another evaluation strategy entailed pitting the trained model against two archaeologists. Sixty-three query images of different artefacts were selected, three for each of the twenty-one fine-period groups. These images were then presented to two archaeologists, and the model to be assigned their appropriate temporal designations. The results indicate that the model performed as well as these two archaeologists within their field of expertise, and had a higher average accuracy level, when considering all possible periods. Thus, the model achieved an average accuracy score of 69.84% (Fig. 2d), while the archaeologists scored 44.44 and 20.63% (Fig. 2e).

Having attained these results, the best classification choice in the archaeological dataset was determined. To do so, an experiment was devised that entailed training the model with few classification objectives—only sites, only periods, or a combination of sites and periods—and compared their performance (see the additional classification experiments section in the supplementary material). It was found that (1) When trained on period-site classes, the model achieved the highest accuracy levels for all three parameters (period-site, period, and site); (2) When trained on periods, the model's periodic attributions remained unchanged (compared to 1), while the precision of its period-site and site attributions dropped; (3) When trained on sites, the model's accuracy levels were nearly as good as in 1.

On these grounds, it can be proposed that information about artefacts' sites of origin carries significant weight for effective network learning. Therefore, it should be used in future works for classification with the periodic data (see the supplementary material for further details, Table S2). This is also the reason that network was train with period-site data also when it is tested only on the period information.

Community detection

A close review of the model's prediction accuracy presented above suggests that most errors entail the confusion of neighbouring periods (e.g., a Pre-Pottery Neolithic A artefact mistakenly attributed to the Pre-Pottery Neolithic B). The propensity for such errors is readily illustrated by a chronologically sorted confusion matrix (Fig. S3), demonstrating that most errors clustered along the main diagonal (i.e., they occurred between nearby periods). While this observation can be read as indicating an inherent weakness in the model, it also indicates the model's response to an actual condition: that visually similar artefacts often derive from temporally adjacent contexts. On these grounds, model's ability to discern associations among classes (i.e., period-site designations) that can

correspond to meaningful archaeological categories was explored. Technically, such clusters are termed 'communities.'

The archaeological community detection method is illustrated in Fig. 1c. It starts by converting the confusion matrix into a network (i.e., graph) that consists of nodes and edges (i.e., links). In this case, each node represents a class, and each edge represents the confusion between classes that was registered in the confusion matrix. Next, the edges were weighted—they were given numerical values to capture their different strengths. An edge's weight was computed as follows:

(1) Let $A \in \mathbb{R}^{C \times C}$ be the normalised confusion matrix. C is the number of classes and A_{ij} is the relative number of cases, where the true label is i and the predicted label is j. Note that this matrix is not necessarily symmetric, i.e., it may have $A_{ij} \neq A_{ji}$.

(2) Let $B = \frac{1}{2}(A + A')$ be the symmetrical version of A.

(3) B_{ij} or B_{ji} is the weight of the edge that connects nodes i and j.

Next, the Louvain community detection algorithm (Blondel et al., 2008) was applied to the network (Fig. 1c, Fig. S6, see methods section for more details), producing clusters—communities—of similar period-site classes. Twenty-eight communities were detected with a modularity score—a measure of the network's division into communities—of 0.77.

In an attempt to achieve better communities, two further adjustments were introduced. The first consisted of rebuilding the confusion matrix to include ten nearest neighbour predictions for each query image instead of one. This modification resulted in more confusion and, by extension, a denser network with more edges. The second adjustment was to use only certain part of the confusion matrix, with several neighbour periods, before applying community detection (e.g., Palaeolithic–Epipalaeolithic periods; Bronze–Iron Ages). In this manner, irrelevant confusion is precluded, and a way is paved to explore more nuanced relations among classes. Thus, for instance, Table S3 presents the communities detected for three periodic groups: Palaeolithic–Natufian, Bronze–Iron Ages, Hellenistic—Byzantine periods.

Setting out to render these community detection procedures relevant for archaeological practice, an interactive computer application was developed geared to visually present classes and communities against their geographical setting (Fig. 3a). Thus, for instance, Fig. 3b offers an overview of the communities detected, Fig. 3c demonstrates the application's node selection mode, where the user is presented with all community members associated with a specific node, and Fig. 3d presents a community of nine members (archaeological sites) —eight Roman and one Byzantine—around the Dead Sea.

The resulting communities' validity may be tested against their members' periodic attributions. If the community comprises one or two successive periods, we may consider the community valid. However, if the community includes outliers—i.e., members whose periodic attribution is inconsistent with the rest of the group— a problem may be assumed, or that there are interesting similarities that need to be further explored.

For example, community 1 in Table S3, in the Bronze–Iron ages, has the following members: Early Bronze I Megiddo, Early Bronze I Mizpah, Early Bronze I 'Ai, Early Bronze II-III 'Ai, Early Bronze III Jericho, Middle Bronze I Megiddo, Iron II Bet Mirsham. Iron II Bet Mirsham, is considered a-priori as an outlier because the periodic assignment is different from the rest Bronze classes. Therefore, it would be interesting to look at the confusion between artefacts in this community.

Notably, the number of outliers per community is a function of the range of periods included in the confusion matrix, that was used in the community detection. Therefore, the results should be carefully analysed and validated with
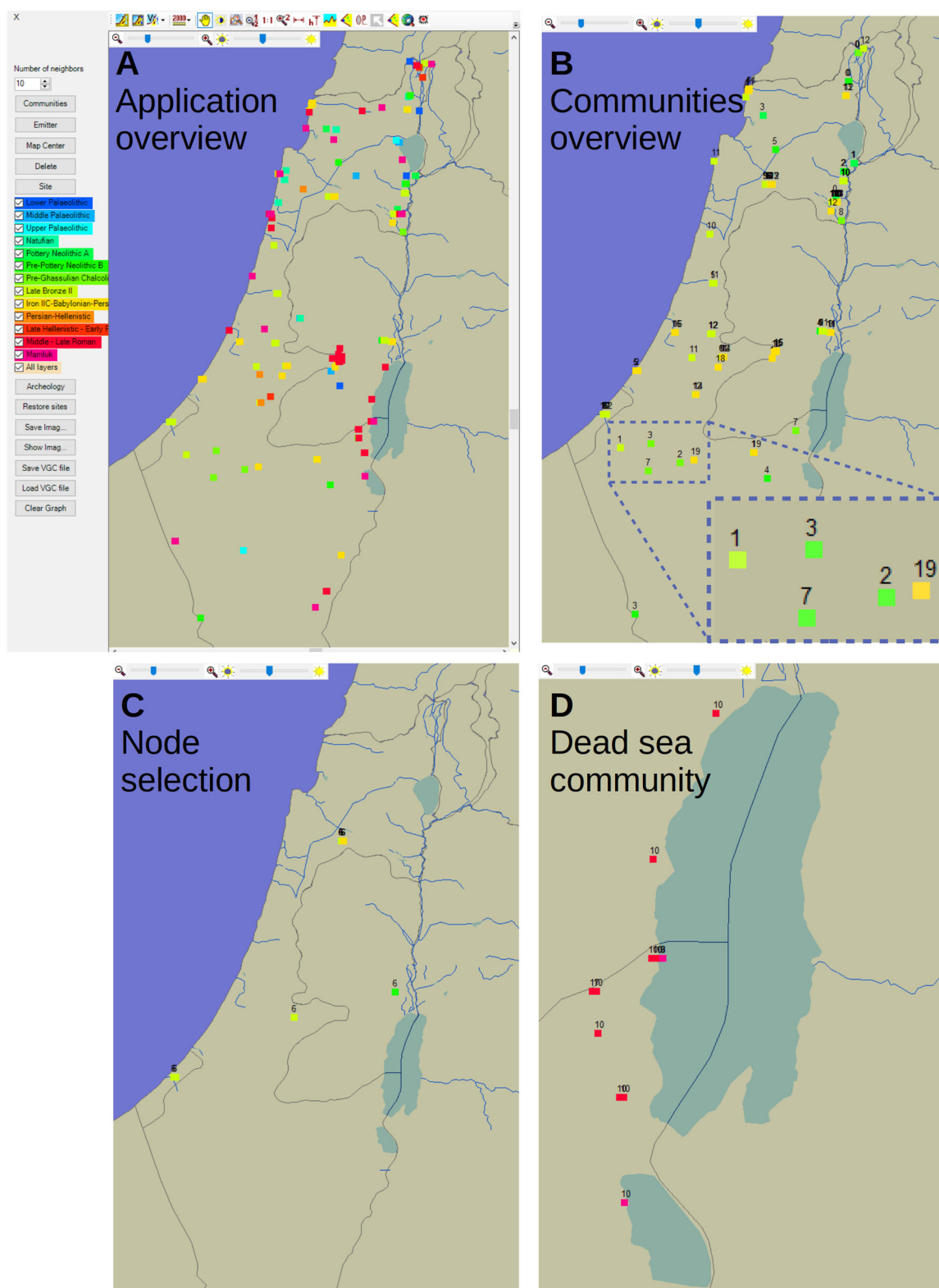
Fig. 3 Map application for interactive community detection. A Classes in the database are represented by coloured nodes, where the colour represents period. The menu on the left allows the user to alter the period groups presented and find communities of interest. **B** Community detection of some period groups based on ten nearest neighbours. The number above each node represents its community. **C** Node selection mode: displays the community members of a selected class; In this example, they include (in period-site format) Iron I-Megiddo, Late Bronze Age II-Bet Shemesh, Late Bronze Age II-'Ujul, Late Bronze Age II-Megiddo, Late Bronze Age-Megiddo, Late Bronze Age-'Ujul, Middle Bronze Age II–Late Bronze Age-Megiddo, and Pre-Pottery Neolithic B-Jericho. **D** An example of a community clustered around the Dead Sea; it consists (in period-site format) of Early Roman-Horevot Mazada, Early Roman-Qumran Caves, Early Roman-'En Gedi, Roman-Horevot Mazada, Roman-Wadi Murabba, Roman-Mezad Rahel, Roman-'En Gedi, Roman-Nahal Mishmar Cave of the Treasure, Byzantine-Mesad Boqeq, and Roman-Nahal Hever.

archaeologists to compensate for insufficiently diverse or imbalanced datasets.

Community detection—Natufian case-study

To explore the potential of the community detection method, a case-study of the Natufian culture is presented here. Since its definition in the 1930s, the Natufian culture of the Levantine late Epipalaeolithic period (ca. 15,000–11,700 years ago) attracted considerable scholarly attention. There are two main reasons for this. First, the Natufian archaeological record suggests a shift from small nomadic human groups to sedentary hamlets in the Mediterranean zone, a unique event of settling down shortly before the transition to farming in the Neolithic Period (Bar-Yosef, 1998; Bar-Yosef and Valla, 2013). Second, while many Natufian artefact types resemble those of the early Epipalaeolithic and Upper Palaeolithic periods (e.g., pointy implements made of bone), many others are novel, producing unprecedentedly diverse assemblages that include abundant worked-stone and worked-bone items, art items, and personal ornaments. Consequently, Natufian artefacts can be found in museum and web exhibits, such as this database. A dataset of five rough-period groups was constructed, spanning the Middle Palaeolithic and the Pre-Pottery Neolithic B, thus constituting a temporal range up to two steps removed from Natufian elements (rough-period groups 2–6; Table S1). In this manner, it may be expected that a query of Natufian classes will find close ties with other Natufian classes, weaker ties with classes that are one step removed, and nearly none with classes two steps removed. Five communities were detected (all site designations follow the labels of the IAA picture database): (1) Natufian_Me'arat Kebara, Natufian_Me'arat ha-Nahal, Natufian_Magharat Shuqba, Pottery Neolithic A_Jericho-T. (2) Upper Palaeolithic_Me'arot Hayonim, Natufian_Me'arot Hayonim, Pre-Pottery Neolithic B_Nahal Hemar (3) Middle Palaeolithic_Me'arat Tannur, Middle Palaeolithic_Har Qedumim, Natufian_'Enot' Eynan, Pre-Pottery Neolithic A_Har Harif (4) Upper Palaeolithic_Me'arat Kebara (5) Natufian_Me'arat Oren.

These communities demonstrate few interesting insights: first, in communities 4 and 5 there is only one class. It means that probably there was no confusion between this class to others, resulting in self-loops in the network. Second, in community 1, Pottery Neolithic A_Jericho is most likely an outlier, because it doesn't belong to the Natufian period, like the rest of the members. Third, in community 2, there are two classes from the same archaeological site (Me'arot Hayonim), one dated to Upper Palaeolithic, and the second to the Natufian culture.

A close review of the details demonstrates that many of the affiliations among artefact images, upon which communities are subsequently established, were both visually similar and archaeologically significant. Figure 4 presents some examples of confusion between artefact images. Thus, Community 1 includes similar Natufian bone implements from different sites (Fig. 4a), Community 2 encompasses worked animal teeth from Upper Palaeolithic and Natufian Me'arot Hayonim (Fig. 4c), Community 3 contains flint tools from Middle Palaeolithic Har Qedumim and Me'arat Tannur (Fig. 4e), Community 4 consists of Upper Palaeolithic bone awls from Kebara Cave (Fig. 4g), and Community 5 includes Natufian worked-stone items from Nahal Oren (Fig. 4h).

However, on several occasions, visual similarities among artefacts produced archaeologically false (or problematic) associations. In Community 2, Natufian bone awls were grouped with Pre-Pottery Neolithic B flint arrowheads, which were of similar shape and colour (ca. 10,000 years ago; Fig. 4d). In Community 3, Natufian implements made on ungulate long bones from 'Eynan (Hula Valley, northern Israel) were grouped

with similar Pre-Pottery Neolithic A artefacts from Har Harif (Negev Desert, southern Israel) (Fig. 4f).

The analysis above is only the tip of the iceberg, as it examined thoroughly some examples of confusion between community members. Researchers are encouraged to follow this procedure with other communities in the dataset (e.g., Table S3), or apply the community detection workflow on other archaeological databases.

Methods

This section provides additional technical details for particular parts of this work. Each subsection is concerned with a specific methodological or procedural component and does not communicate directly with the others.

Image pre-processing. The images that populated the database were collected without an image capturing protocol. Consequently, image capturing conditions varied considerably from one case to the next, mainly pertaining to issues of background and scale. To overcome this, homogeneous white background was implemented and removed the scale following one of two procedures: (1) automatic contour retrieval (Suzuki, 1985) performed on the output of the Canny edge detector (Canny, 1986) on the input image, or (2) the interactive GrabCut method (Rother et al., 2004). The second procedure is comparatively manual and used whenever the first procedure failed. To fit images to the model input spatial dimensions, the images were resized to 300x300 pixels.

Base network. To choose the base network, three ImageNet pre-trained models were evaluated. These include VGG (Simonyan and Zisserman, 2014), InceptionResNetV2 (Szegedy et al., 2017), and EfficientNetB3 (Tan and Le, 2019). We found that EfficientNetB3 was 1% more precise than the other two and needed fewer epochs for training.

Loss functions. Large Margin Cosine Loss (Wang et al., 2018) and cross-entropy loss functions resulted in similar classification accuracy measures while further training with online triplet mining and triplet loss (Schroff et al., 2015) improved results by around 1% on VGG and InceptionResNetV2. The final model was trained with the cross-entropy loss alone on EfficientNetB3.

Distance metric. Predictions for the query images were generated by determining their k-nearest training-set neighbours ($k = 1$ in this setup). For this purpose, the cosine similarity distance measure was used:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|},$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$ is the feature vectors of two different input images, and D is the embedding vector length.

Voting of five CNNs. To optimise these results, five models with the same ImageNet initialisation were trained, each generating a different feature vector, which was then used to produce the final feature vector (Z^{RP}). To do so, (1) the five feature vectors were concatenated, achieving $Z \in \mathbb{R}^{5D}$, where D is the single model feature vector length, and (2) randomly projected Z to a lower-dimensional space (due to memory limitations) by multiplying it with the random Gaussian matrix

$$Z^{RP}_{D \times 1} = G_{D \times 5D} Z_{5D \times 1}$$

where Z^{RP} is Z projected onto a lower D-dimensional subspace, and $G_{D \times 5D}$ is a random Gaussian matrix.
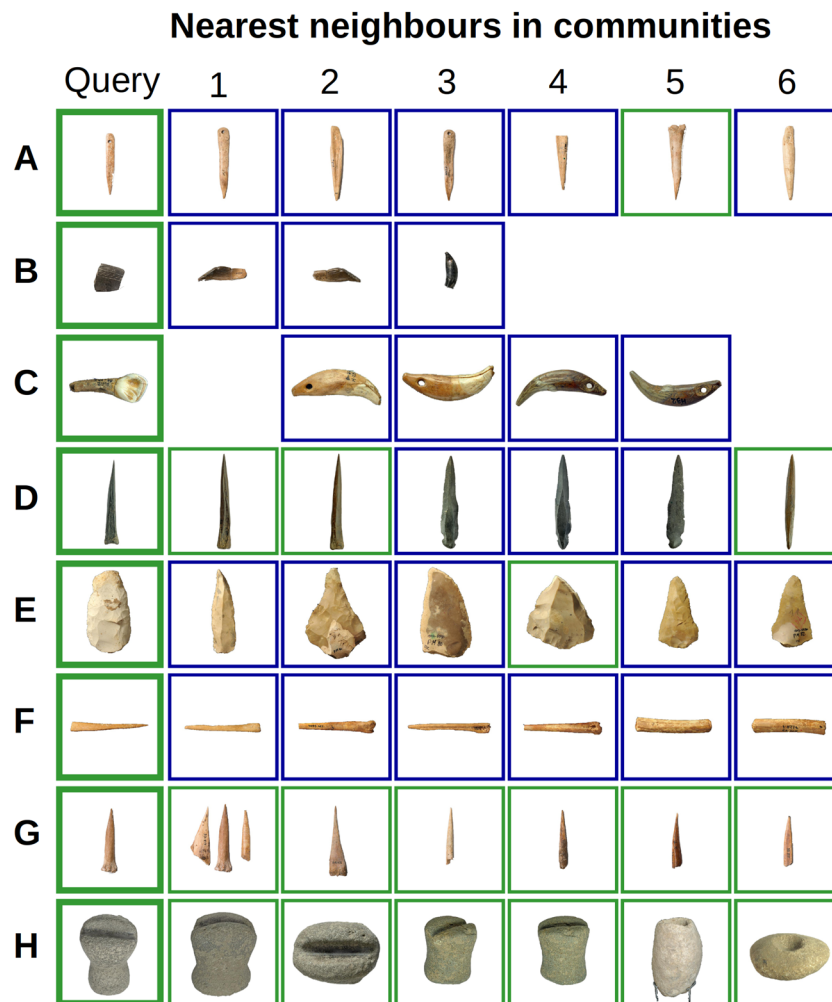
Nearest neighbours in communities

Fig. 4 Natufian artefact image confusion in community detection case-study. The query images are presented in the left column, while, to their right, the nearest training-set neighbours are presented in order. If the neighbour is of the same class as the query (i.e., of the same site and period), it is placed in a green frame. Otherwise, a blue frame is used. A blank space indicates that the neighbour image detected by the model was assigned to a different community. A more detailed description of this figure can be found in supplementary material (Additional information for Fig. section). **A** Community 1, archaeologically meaningful confusion. **B** Community 1, wrong confusion. **C** Community 2, archaeologically meaningful confusion. **D** Community 2, wrong confusion. **E** Community 3, archaeologically meaningful confusion. **F** Community 3, confusion. **G** Community 4, correct predictions. **H** Community 5, correct predictions.

Training details. CNN weights were optimised by the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of 0.0001 divided by 10 when validation loss was not improving. The hardware used throughout these experiments is a single Nvidia GeForce 2080 Ti GPU, and the batch size was 20.

Community detection. A modularity score measures the quality of a network's partition into communities (Blondel et al., 2008; Newman and Girvan, 2004). A high score indicates dense connections within communities and sparse connections between them. It is defined as the fraction of edges within communities minus the expected fraction had their distribution been random.

Derivation of the modularity formula starts with two nodes, v and w. The difference between the actual and expected weight between nodes v and w is calculated as follows:

$$A_{vw} - \frac{k_v k_w}{2m}$$

where (1) A_{vw} is the weight between v and w, (2) $k_i = \sum_j A_{ij}$ is equivalent to the degree of node i, and (3) $m = \frac{1}{2}\sum_{ij} A_{ij}$ is the sum of all weights in the graph (number of edges in a uniform-weights graph).

Summation over all pairs that belong to the same community will yield the modularity score Q:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v, c_w)$$

where c_i is the community of node i, and $\delta(c_v, c_w)$ equals one or zero if nodes v and w belong to the same or different communities, respectively.

Based on this metric, Blondel et al. (2008) introduced a popular community detection algorithm (Fig. S6). It is based on the iteration of two phases. First, each node is assigned to a different community, and the modularity gain of node i is calculated, should it be found to be in the same community as its neighbour j. After considering all possible neighbours, node i is placed in the community that produced the highest modularity gain. This process is repeated until no further improvement in modularity score is noted.

The second phase entails establishing a new network based on the communities detected in the first phase. Each community is

represented by a node, and edges' weights are determined by summing all the edges between communities, while edges within communities produce self-loops.

The combination of these two phases is called a "pass," and it is repeated until the modularity score stabilises and maximum modularity is achieved (Fig. S6).

Prior confusion. Ambiguities concerning periodic attribution (e.g., Roman/Early Roman) may be considered a type of label noise. However, in practice, they are attributable to several closely related features of the archaeological record: (1) Most artefact types span several periods, (2) archaeological periods usually have vague boundaries, and (3) artefacts may vary in frequency across time and space while retaining their formal properties.

Motivated by Kaneko et al. (2019), attempts were performed to enhance the loss function with prior confusion knowledge. Let (x_j, y_j) be an image-label pair; given x_j, the probability for label y_j will be

$$p\left(y_j\middle|x_j\right) = \sum_i p\left(y_j\middle|y_i\right)p\left(y_i\middle|x_j\right),$$

where $p(y_i|x_j)$ is the ith output of the neural network's final layer, when the input image is x_j, and $p(y_j|y_i)$ is the measure of ambiguity between labels y_j and y_i. For example, if there is 50% indeterminacy between Persian-Hellenistic and Hellenistic labels, it would be 0.5. The final cross-entropy loss for mini-batch with B images and C classes is

$$\text{Loss} = -\sum_j^B \sum_i^C t_{ij} \log p\left(y_j\middle|x_j\right),$$

where t_{ij} is the ith one hot encoding element of the label y_j.

Notwithstanding the method's potential, quantifying the prior ambiguity measure— $p(y_j|y_i)$—proved difficult, rendering it useless for this purposes.

Attempts were made to manage periodic indeterminacies by setting $p(y_j|y_i)$ according to a Gaussian function. Unfortunately, this method did not improve the model's accuracy measures.

Website for archaeological predictions
A website containing the pre-trained CNN model is available[1]. Researchers are invited to upload their query images and receive images of similarly labelled artefacts from the training set.

Conclusion
Machine learning is a powerful tool to explore large datasets. This paper describes the development of a deep-learning-based model for a diverse archaeological dataset that spans more than a million years of south Levantine material culture. It is particularly well-suited for purposes of artefact classification, potentially accelerating the interpretation of archaeological contexts. Moreover, based on the model, meaningful connections across artefacts, assemblages, and sites were automatically found.

Notably, archaeological classification is uniquely challenging. It is often ambiguous, and there is considerable room for controversy over dating. Moreover, archaeological assemblages are synchronically variegated, encompassing materially and visually distinct objects, but often diachronically similar. Harnessed this inherent quality of temporal ambiguity is key to find meaningful archaeological communities, recognising that the confusion of classes can underscore real connections.

At its most basic, this CNN can help archaeologists find similar artefacts and efficiently complete some of the more tedious and humdrum tasks of the profession. At its more advanced applications, the model can help archaeologists analyse large data bodies, find new previously unknown relations, and raise new archaeological questions. This workflow presented here can be applied to other datasets worldwide and has the potential to make way for significant archaeological insights.

Note
1 See: https://github.com/aviresler/antique-gen. We recommending visiting the "Prediction of archaeological period/site" section.

References
Agam A, Azuri I, Pinkas I et al. (2020) Publisher correction: estimating temperatures of heated lower palaeolithic flint artefacts. Nat Human Behav 4:1322. https://doi.org/10.1038/s41562-020-01017-0

Arkadiev PM (2020) Morphology in typology: Historical retrospect, state of the art, and prospects. Oxford research encyclopedia of linguistics. https://doi.org/10.1093/acrefore/9780199384655.013.626.

Bar-Yosef O (1998) The Natufian culture in the Levant, threshold to the origins of agriculture. Evolution Anthropol 6:159–177

Bar-Yosef O, Goren-Inbar N (1993) The lithic assemblages of 'Ubeidiya: A Lower Palaeolithic site in the Jordan Valley. Qedem

Bar-Yosef O, Valla FR (eds.) (2013) Natufian foragers in the Levant: Terminal Pleistocene social changes in Western Asia. Int Monogr Prehist 19

Barcelo JA (1995) Back-propagation algorithms to compute similarity relationships among archaeological artefacts. In J Wilcock, K Lockyear (eds.) CAA 1993: computer applications and quantitative methods in archaeology (BAR International Series 598). Tempus Reparatum, pp. 165–176

Barcelo JA (2008) Computational intelligence in archaeology. IGI Global

Barceló JA (2016) The role of computers to understand the past: The case of archaeological research. It–Informat Technol 58(2):104–111. https://doi.org/10.1515/itit-2015-0034

Barcelo JA, Bogdanovic I (eds.) (2015) Mathematics and archaeology. CRC Press

Bishop CM (2006) Pattern recognition and machine learning. Springer

Blondel VD, Guillaume J-L, Lambiotte R et al. (2008) Fast unfolding of communities in large networks. J Stat Mechan 2008:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boon P, van der Maaten L, Paijmans H et al. (2009) Digital support for archaeology. Interdiscip Sci Rev 34(2–3):189–205

Canny JA (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Machine Intell 8(6):679–698. https://doi.org/10.1109/TPAMI.1986.4767851

Cifuentes-Alcobendas G, Domínguez-Rodrigo M (2019) Deep learning and taphonomy: high accuracy in the classification of cut marks made on fleshed and defleshed bones using convolutional neural networks. Sci Rep 9:1–12

Derech N, Tal A, Shimshoni I (2021) Solving archaeological puzzles. Pattern Recog 119:108065. https://doi.org/10.1016/j.patcog.2021.108065

Díez-Pastor JF, Jorge-Villar SE, Arniaz-Gonzaléz Á et al. (2018) Machine learning algorithms applied to Raman spectra for the identification of variscite originating from the mining complex of Gavà. J Raman Spectrosc 51(9):1563–1574. https://doi.org/10.1002/jrs.5509

Domínguez-Rodrigo M, Cifuentes-Alcobendas G, Jiménez-García B et al. (2020) Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. Sci Rep 10:18862. https://doi.org/10.1038/s41598-020-75994-7

Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley

Dunnell RC (1993) Archaeological typology and practical reality: A dialectical approach to artifact classification and sorting. Am Antiq 58:165–167

Grove M, Blinkhorn J (2020) Neural networks differentiate between middle and later Stone Age lithic assemblages in eastern Africa. PLoS ONE 15:e0237528. https://doi.org/10.1371/journal.pone.0237528

He K, Zhang X, Ren S et al. (2016) Deep residual learning for image recognition. In 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp. 770–778

Hermon S, Nuccolucci F, Alhique F et al. (2004) Archaeological typologies—an archaeological fuzzy reality. In Fischer- Ausserer K, Börner W, Goriany M, Karlhuber-Vöckl L (eds.) CAA 2003: computer applications and quantitative methods in archaeology (BAR International Series 1227). Archaeopress, pp. 30–34

Itkin B, Wolf L, Derishowitz N (2019) Computational ceramicology. ArXiv. https://arxiv.org/abs/1911.09960

Kaneko T, Ushiku Y, Harada T (2019) Label-noise robust generative adversarial networks. In 2019 IEEE/CVF conference on computer vision and pattern recognition. IEEE, pp. 2462–2471

Krieger AD (1944) The typological concept. Am Antiq 9:271–288

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90

Loshchilov I, Hutter F (2019) Fixing weight decay regularisation. ArXiv. https://arxiv.org/pdf/1711.05101.pdf

MacLeod N (2018) The quantitative assessment of archaeological artifact groups: Beyond geometric morphometrics. Quater Sci Rev 201:319–348

Mitchell TM (1997) Artificial neural networks. Mach Learn 45:81–127

Murray T, Evans C (2008) Histories of archaeology: A reader in the history of archaeology. Oxford University Press

Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113. https://doi.org/10.1103/PhysRevE.69.026113

Renfrew C, Bahn P (2013) Archaeology: The key concepts. Routledge

Rother C, Kolmogorov V, Blake A (2004) "GrabCut": Interactive foreground extraction using iterated graph cuts. AMC Trans Graphics 23(3): 309–314

Russakovsky O, Deng J, Su H et al. (2015) ImageNet large scale visual recognition challenge. Int J Comput Vision 115:211–252. https://doi.org/10.1007/s11263-015-0816-y

Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: A unified embedding for face recognition and clustering. In 2015 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp. 815–823

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. ArXiv. https://arxiv.org/abs/1409.1556v6

Suzuki S (1985) Topological structural analysis of digitised binary images by border following. Comput Vis Graph Image Process 30:32–46. https://doi.org/10.1016/0734-189X(85)90016-7

Szegedy C, Ioffe S, Vanhoucke V et al. (2017) Inception-v4, inseption-ResNet and the impact of residual connections on learning. In Proceedings of the thirty-first AAAI conference of artificial intelligence (AAAI '17). AAAI Press, pp. 4278–4284.

Taigman Y, Yang M, Ranzato M et al. (2014) DeepFace: Closing the gap to human-level performance in face verification. In 2014 IEEE conference on computer vision and pattern recognition. IEEE, pp. 1701–1708

Tal A (2014) Shape analysis in archaeology. In Ioaninides M, Quak E (Eds.) 3D research challenges in cultural heritage. Springer, 50–63

Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. Proc Machine Learn Res 97:6105–6114. http://proceedings.mlr.press/v97/tan19a.html

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:11

Wang H, Wang Y, Zhou Z, et al. (2018) CosFace: Large margin loss for deep face recognition. In IEEE/CVF conference on computer vision and pattern recognition. IEEE, pp. 5265–5274

Whittaker JC, Caulkins D, Kamp KA (1998) Evaluating consistency in typology and classification. J Archaeol Method Theory 5:129–164

Acknowledgements

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval were not required for this study.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-021-00970-z.

Correspondence and requests for materials should be addressed to Filipe Natalio or Raja Giryes.

Reprints and permission information is available at http://www.nature.com/reprints

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

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;

2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;

3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;

4. use bots or other automated methods to access the content or redirect messages

5. override any security feature or exclusionary protocol; or

6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com