Quaternary Science Advances 17 (2025) 100268

Contents lists available at ScienceDirect

Quaternary Science Advances

journal homepage: www.sciencedirect.com/journal/quaternary-science-advances

Testing the reliability of geometric morphometric and computer vision methods to identify carnivore agency using Bi-Dimensional information

Manuel Domínguez-Rodrigo [a,b,c,*], Marina Vegara-Riquelme [a,b], Juan Palomeque-González [a], Blanca Jiménez-García [a,b], Gabriel Cifuentes-Alcobendas [a,b], Marcos Pizarro-Monzo [a,d], Elia Organista [e], Enrique Baquedano [a,f]

[a] Institute of Evolution in Africa (IDEA), University of Alcalá de Henares, Madrid, Spain
[b] Area of Prehistory, Department of History and Philosophy, University of Alcalá de Henares, Alcalá de Henares, Spain
[c] Department of Anthropology, Rice University, 6100 Main St., Houston, TX, 77005-1827, USA
[d] Department of Geodynamics, Stratigraphy and Paleontology, Complutense University, Madrid, Spain
[e] Osteoarkeologiska forskningslaboratoriet och antikens kultur Stockholm Universitet, Stockholm, Sweden
[f] Archaeological and Paleontological Museum of Madrid, Alcalá de Henares, 28801 Madrid, Spain

ARTICLE INFO

Keywords:
Taphonomy
Artificial intelligence
Carnivore
Human evolution

ABSTRACT

Bidimensional information of tooth marks and other bone surface modifications (BSM) presents limitations, as highlighted in this study. Here, we establish a methodological comparison on a controlled experimentally-derived set of BSM generated by four different types of carnivores, using geometric morphometric (GMM) and computer vision (CV) methods. We highlight that previous generalizations of high accuracy on tooth marks using GMM are heuristically incomplete, because only a small range of allometrically-conditioned tooth pits have been used. Biased replication and exclusion of the most widely represented forms of non-oval tooth pits from such analyses have compromised the published results and their ensuing generalizations. Here, we document bidimensionally a much wider range of tooth pits, using their outlines (and not a limited set of non-reproducible *idem locus* semi-landmarks), through Fourier analyses. The resulting tooth mark sets show low accuracy (and resolution) in the classification of tooth marks per carnivore modifying agent. This low resolution is also reproduced when using a semi-landmark approach. In contrast, our study demonstrates that CV approaches, through Deep Learning (DL), using convolutional neural networks (DCNN), and Few-Shot Learning (FSL) models, classify experimental tooth pits with 81% and 79.52% accuracy, respectively, being equally efficient at classification. However, a limitation in CV methods occurs when applied to the fossil record, as BSM undergo dynamic transformations over time. The most impactful processes occur early in taphonomic history, altering the original BSM properties. Consequently, no objective referents exist for marks combining original and subsequent diagenetically or biostratinomically modifying processes. However, in well-preserved contexts, such as the 1.8 Ma tooth marks from some of the Olduvai sites, confidence in interpretations can be high with convergent CV models indicating high agent attribution probability. While GMM shows potential in 3D, its current bidimensional application yields limited discriminant power (<40%). Thus, future research should utilize complete 3D topographical information for more complex GMM and CV analyses, potentially resolving current interpretive challenges. Despite necessary cautions, these new methods offer an unprecedented objective means of classifying BSM to taxon-specific agency with confidence indicators. Continued research should refine these approaches, enhancing the reliability of prehistoric interpretations.

1. Introduction

In the past few years, there has been an increasing awareness among taphonomists that generic agency (e.g., carnivore bone modification) was insufficient to address not only site formation, but also hominin behavior (e.g., hominin-carnivore interaction types) (Jiménez-García et al. 2020a, 2020b; Abellán et al., 2021; Abellán et al., 2022; Domínguez-Rodrigo et al., 2024; Courtenay et al., 2019; Courtenay et al.,

* Corresponding author. Institute of Evolution in Africa (IDEA), University of Alcalá de Henares, Madrid, Spain.
E-mail address: m.dominguez.rodrigo@gmail.com (M. Domínguez-Rodrigo).

https://doi.org/10.1016/j.qsa.2025.100268
Received 12 August 2024; Received in revised form 10 January 2025; Accepted 17 January 2025
Available online 30 January 2025

2021; Cobo-Sánchez et al., 2022; Brugal and Fourvel, 2024). For this reason, the study of some skeletal modification patterns (namely, bone surface modifications [BSM]) has undergone a transformation involving: a) the use of multivariate approaches and, b) the implementation of objective statistical methods, which can provide some confidence in taxon-specific agency identification. Almost synchronically, two separate approaches have been developed for the study of carnivore tooth marks: bi- and tridimensional geometric morphometrics (GMM) (Courtenay et al. 2019, 2021), and computer vision (CV) through deep convolutional neural networks (DCNN) (Jiménez-García et al. 2020a, 2020b; Abellán et al., 2021; Abellán et al., 2022; Domínguez-Rodrigo et al., 2024). The latter is fully derived from the implementation of artificial intelligence methods in taphonomy. GMM methods have been used in conjunction with either traditional statistical classification methods (e.g., linear discriminant analysis or canonical covariate analysis) or with classificatory machine learning algorithms. CV methods have been used mostly with deep learning classification algorithms (i.e., deep neural networks). The application of both methods have independently boosted our confidence in BSM identification, by providing experimentally-controlled models that were able to classify testing data sets with high accuracy. The limited application of these models to the archaeological record have provided, in some cases, high-confidence classifications of BSM (Domínguez-Rodrigo et al., 2021a,2021b; Cifuentes-Alcobendas and Domínguez-Rodrigo, 2019). However, it must be emphasized that the heuristics of accurate classification of experimentally-controlled datasets and of archaeological datasets are not the same. After an initial optimistic (and we would say, even euphoric) adoption of these methods (which undoubtedly are a major improvement over the classical human expert-dependent identification methods (Domínguez-Rodrigo et al., 2017)), there comes a moment for reflection and caution. The present work will illustrate several showcases about this.

Not only have these mathematically-founded models succeeded in classifying multiple carnivore agencies (at the taxon-specific level) with high accuracy; both GMM and CV have even excelled at differentiating among BSM generated by different types of felids (Jiménez-García et al. 2020a, 2020b; Jiménez-García et al. 2020a, 2020b) or different types of canids (Yravedra et al., 2019). Given the overall similar morphology in the dentition of each of these closely related carnivore groups, one wonders how discriminating among them so efficiently is even possible; especially given that frequently (e.g., among canids), the morphology of their dentition differs only by size, and in some cases, also by restricted allometry. An additional caveat is that all the modern referential datasets have been obtained on fresh bone, and that the marks analyzed are in pristine state of conservation. This is rarely the case of BSM on fossil bones (Pizarro-Monzo et al., 2022; Pizarro-Monzo and Domínguez-Rodrigo, 2020; Gaudzinski-Windheuser et al., 2010). Only occasionally have taphonomists addressed that the BSM that they study are the result of palimpsestic dynamic processes, and that both the microscopic features imprinted on marks and their morphology, as generated by the primary biotic agent, rarely remain intact. Only two CV studies have analytically addressed these issues, considering the biostratinomic morphing of BSM caused by time-averaged abrasion (Pizarro-Monzo and Domínguez-Rodrigo, 2020) (see also Gaudzinski-Windheuser et al., 2010), and the diagenetic modification caused by sedimentary chemical conditions (Pizarro-Monzo et al., 2022). An additional problem is that the same DL method can provide equally accurate classification models that result in divergent classification of marks; this has been noticed in ensemble analyses. What happens when two successful experimental models provide different identification of the same archaeological/paleontological mark?

Although GMM and CV have provided similarly accurate results in the classification of their validation and testing datasets, their approach has been very different, and so is their reliability. CV studies have included all types of tooth marks in their analyses. GMM, in contrast, has been highly selective in the type of tooth pits used, focusing mostly on oval-shaped pits displaying some specific degree of relief. This implies that GMM achieves its seemingly good performance using only a selective fraction of experimentally modified BSM. If both methods are to be trusted when applied to archaeological BSM, they should initially display comparatively similar convergence in their identification of experimental BSM. Secondly, they should also converge in the classification of archaeological BSM. Until now, no attempt has been made to combine both methods on the same experimental dataset.

Here, we will use GMM and CV on the same four-carnivore experimental assemblages. We will then proceed to use the resulting resolutive method(s) on a limited number of fossil tooth marks, with different types of post-sedimentary modifications. The main goal is to test the degree of confidence that can be laid on these highly-accurate experimental methods. Our first testing hypothesis is that if both are reliable on modern assemblages, they should provide identical or very similar results. Our second hypothesis is that if the dynamic modification of BSM is not an obstacle, both methods should equally classify the selected sample of tooth marks, given their very different configurational and mathematical approaches. In this work, we will look critically both at GMM and CV methods, with special emphasis on their misapplication when using highly (commonly diagenetically) modified fossil BSMs.

2. Sample and methods

2.1. Sample and tooth pit selection

We took advantage of the tooth mark experimental data set of four types of African carnivores (lions, leopards, hyenas and crocodiles) initially analyzed by CV methods (Domínguez-Rodrigo et al., 2024) to create the analytical sample used in the present study. Given that complete mark shape by GMM methods had only been applied to tooth pits (scores have been studied using their cross-section morphology), we subsampled Domínguez-Rodrigo et al. (2024) dataset and selected only the tooth pits for the present analysis. We excluded only the Dar es Salaam leopard subsample, because when the images provided any issues about doubtful identification of mark contour, we could not access the original tooth marked bones to solve them, whereas the remaining collection was immediately accessible at the Institute of Evolution in Africa (IDEA, Madrid). For a complete description of the experiments with the four carnivores, we refer to the original study (Domínguez-Rodrigo et al., 2024). We will describe the combined comparative analytical methods below. The complete data set and code used for the present analysis can be located in the public repository: https://doi.org/10.7910/DVN/77MZDL.

Here, we used a sample of 549 tooth pits divided by carnivore as follows: lions (n = 132), leopards (n = 122), hyenas (n = 207), crocodiles (n = 88). All tooth marks made by lions, leopards and hyenas were made by adult individuals. Tooth marks made by the Faunia crocodiles (providing the bulk of marks) were also made by large individuals (>1.8 m) of several ages. Given that all the eight individuals were fed with our experimental carcasses in the same enclosure at the same time, only the dominant adults (>2.3 m) accessed the carcass parts provided. Most of the time, it was the dominant female (>3m) who monopolized the feeding. We therefore attribute most (if not all) the resulting tooth marks to the two oldest individuals (Domínguez-Rodrigo et al., 2021). Tooth size per carnivore taxon according to age and allometric variation through growth have, therefore, not impacted each carnivore subsample in any important way. Deciduous and definitive dentitions must leave potentially morphologically different tooth marks; the same as they result in dimensionally very different tooth mark sizes (e.g., spotted hyenas) (Andrés et al., 2012). These differences are not reported here, and future experimentation should incorporate them to improve the current referential dataset.

The four carnivores used have been fed with pigs, boars, sheep, cows (crocodiles), sheep (leopards), cows (lions), and deer (hyenas) (Domínguez-Rodrigo et al., 2024). It could be argued that marks might

vary according to carcass type, but to restrict such variation, we have used mostly long limb bones in the present study. We assume tooth marks on axial and more cancellous elements may display a wider range of variation. They certainly do so in size and we infer that potentially also in morphology, but this needs to be explored with additional experiments. Dense bone from limb elements is a more controlling material to how diverse tooth marks can result, because these depend mostly on the tooth morphology and the amount of force imprinted. In our observations, the variation expressed on long bone shafts is mostly dimensional and not so much morphological. Given that tooth marking is just a physical process depending on tooth morphology and force applied, we do not think that cortical bone from different taxa can affect the resulting tooth morphology. This has been tested in a controlled manner using different animal taxa and carcass sizes in butchery experiments, where the resulting butchery marks are morphologically indifferentiable (Maté-González et al., 2019).

Before proceeding, it is important to remark on the high morphological variance of pitting. Tooth pits appear in different formats. Binford's (2014) original tooth mark classification differentiated punctures from pits on the basis of bone collapse. Denting of dense cortical bone through carnivoran tooth impact or pressure, regardless of how shallow or deep the tooth cusp imprint is, was described as pitting by Binford. Individual tooth pits thus generated adopt a wide variety of formats (Fig. 1) (*contra* current GMM studies on tooth marks). An abridged summary of tooth pit types would involve: the marking of the cortical surface with minor cracking of the upper cortical layer and a diffuse contour (Fig. 1a); the cracking and exfoliation of the upper cortical layer in patches, with a balance between the original crushed cortical layer and the shallow patches where the former has been removed (Fig. 1b); the generation of a shallow depression combining both crushing of the uppermost cortical layer and removal thereof, with a defined outline (Fig. 1c); generation of a proper pit through the action of crushing several layers of bone, but with enough asymmetry to make the contour incomplete and the pit depression uneven (Fig. 1d); pit generated by partial bone collapse with a clearly defined shoulder contour (Fig. 1e) and, pit generated by the complete collapse of the inner surface inside the contour, with a transition from pit to puncture, *sensu* Binford (2014) (Fig. 1f). Here, we will use all tooth pit (and puncture) morphologies generated by the carnivores analyzed. We also included a couple of hybrid morphologies resulting from the addition of small scores ending in proper pits, generated by a tooth cusp that slides over the bone cortical surface prior to penetrating through the cortex.

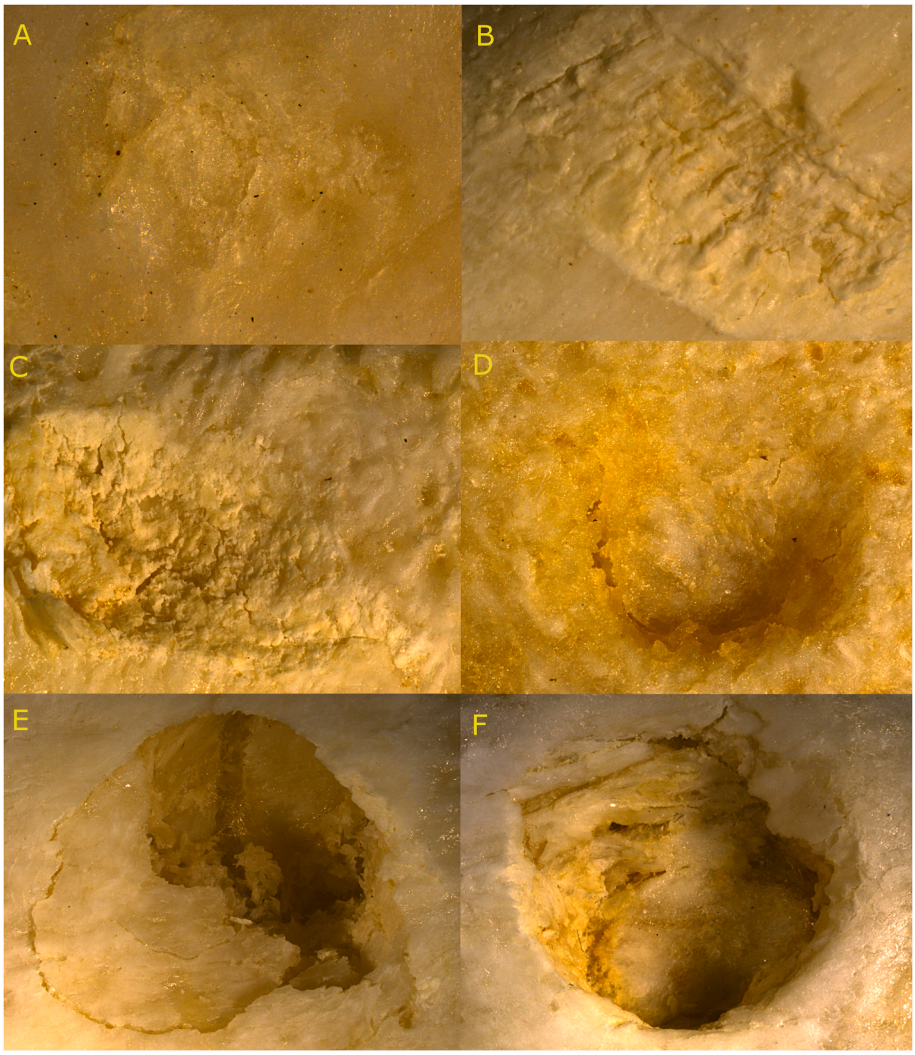

Fig. 1. Different types of tooth pits. A, shallow irregular pit involving minimal loss of upper cortical layer. B, exfoliation of the upper cortical generating a shallow pit whose outline is incomplete. C, tooth pit with complete outline exhibiting breaking and removal of the uppermost cortical layer, with a shallow incipient depression. D, proper pit involving the crushing of several cortical layers, but with asymmetry and incomplete shoulder outline. E, crushing of the upper cortical layer forming a deep depression in part of the pit and marginal collapse on a section of it. F, complete tooth pit with deep bowl shape involving the crushing of successive cortical layers and collapse of the surface area, which Binford referred to as puncture.

2.2. GMM method

Here, we will use a Fourier method to study tooth marks. The method uses an abundant number of semi-landmarks that adapt to the mark outline and reproduces its intricacies. Classical semi-landmark approaches to tooth marks use a more limited number of semi-landmarks along the outline perimeter, creating an approximation to the shape that is functional for analytical purposes, but it is an ideal recreation, which contrasts with the real shape. Therefore, in the present work, we used an outline approach instead of a classical semi-landmark approach, because for non-angular morphologies, classical semi-landmarks create an artificial comparison of non-homologous points. In addition, the semi-landmark approach does not capture bidimensional morphology (especially of irregular shapes) as well as complete contours do. This is why outline analysis constitutes a more solid approach to these types of shapes. From the original sample of tooth pits (including all their morphological diversity) used for the CV analysis (see below), we selected those marks whose outline could unambiguously be delineated. The criterion to define an outline was to focus on the uppermost contour generated through cortical layer crushing by the tooth cusp, involving breaking and deletion of the upper cortical surface, regardless of the depth of the resulting pit. This captured the original morphology of the mark on its shoulder. Pitting frequently does not produce clear-cut boundaries from the external surface to the immediate inner topography, and transitional zones at the microscale level are produced. These zones are determined by crushing, polishing and lifting of successive cortical layers as the tooth cusp penetrates bone tissue. Here, we define the uppermost boundary as the contour that was defined by a single (or multiple) upper layer(s) of bone whose outline appeared complete and uninterrupted. These might coincide with the most concave part of the mark (i.e., the pit *sensu stricto*) or not (i.e., the shoulder). Only those marks with a clearly defined external boundary were selected for GMM analysis. This generated a sample of 284 tooth marks divided by carnivore type as follows: leopards (n = 72), lions (n = 73), spotted hyenas (n = 87), and crocodiles (n = 52) (Figs. 2–4). We are aware that such a sample is a statistically small sample for four different classes (i.e, taphonomic agencies), as we criticized earlier (Domínguez-Rodrigo et al., 2024); however, we carried out the analysis in the context of seeking the extent of classification resolution, as previous similarly small samples analyzed using GMM methods have done (see Table 5). As a complement to the Fourier analysis, and more directly related to previous GMM studies on tooth marks, we also implemented a semi-landmark approach, but sampling the outline morphology at equal intervals of the main axis, instead of just sliding semi-landmarks at equal distance ot the outline perimeter. The method and results are described in the Supplementary Information.

2.3. Elliptical Fourier Analysis

Fourier analysis (FA) is a mathematical method used to understand patterns in morphological data, like the movement of waves, longitudinal (i.e., time-series) data, or the shape of bidimensional objects. There are three main methods of FA: radial, tangent, and elliptical. In radial FA, the goal is to determine how much a pattern (i.e., morphological outline of a shape) varies as we move away from its center (or centroid). This method is useful for studying patterns that have a circular or radial symmetry, but becomes less powerful when analyzing asymmetrical shapes with irregular outlines, as is the case with the tooth pit sample used in the present study. Tangent FA analyzes morphological changes as we move along a surface, similar to drawing the edge of a shape. It is especially efficient in classifying patterns that have straight or curved edges, but becomes less accurate when dealing with irregular or highly variable shapes. The third method, Elliptical FA, combines the best qualities of radial and tangent methods, since it detects how a pattern varies both in distance from its center and along its surface. This method is great for shapes that are more complex, like the tooth marks in the present study, and it is the best option to analyze shape outlines (Claude, 2008).

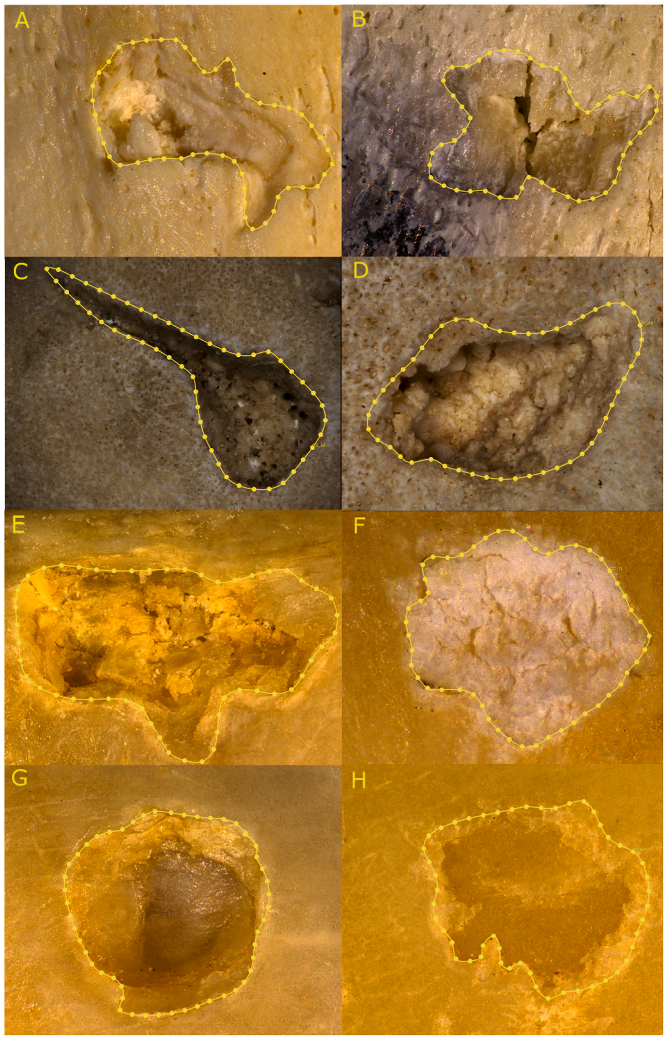

Fig. 2. Examples of some tooth pits displaying irregular outlines made by crocodiles (A–D) and leopards (E–H). Tooth mark C is a mixed score-pit tooth mark.

The Elliptical Fourier Analysis (EFA) method is a mathematical technique that creates representations of shapes by using a large series of evenly-spaced semi-landmarks along the shape contour. These are digitized as a series of points in a two-dimensional space. These points trace the boundary of the shape, capturing its overall morphology (Fig. 5). Then, the EFA method applies Fourier transform techniques to these digitized points. These techniques decompose a complex waveform into a series of simpler sinusoidal (sine and cosine) waves, called harmonics. These harmonics describe shape variation in both "x" and "y" directions. This step is denominated EFA Transform. Once the shape has been decomposed into its Fourier harmonics, EFA approximates these harmonics as ellipses that are fitted to the actual shape, and try to reproduce it. Each ellipse represents one of the harmonics, with its size, orientation, and position capturing different aspects of the shape's structure. This step is denominated Elliptical Approximation. Subsequently, EFA calculates coefficients for each harmonic, representing the size, orientation, and position of the corresponding ellipse on the plane. These coefficients quantify how much each harmonic contributes to the overall shape. Finally, EFA reconstructs the original shape by combining the ellipses corresponding to all the harmonics, using their coefficients. By adding these ellipses, EFA generates an approximation of the original shape, allowing for detailed analysis and comparison (Claude, 2008).

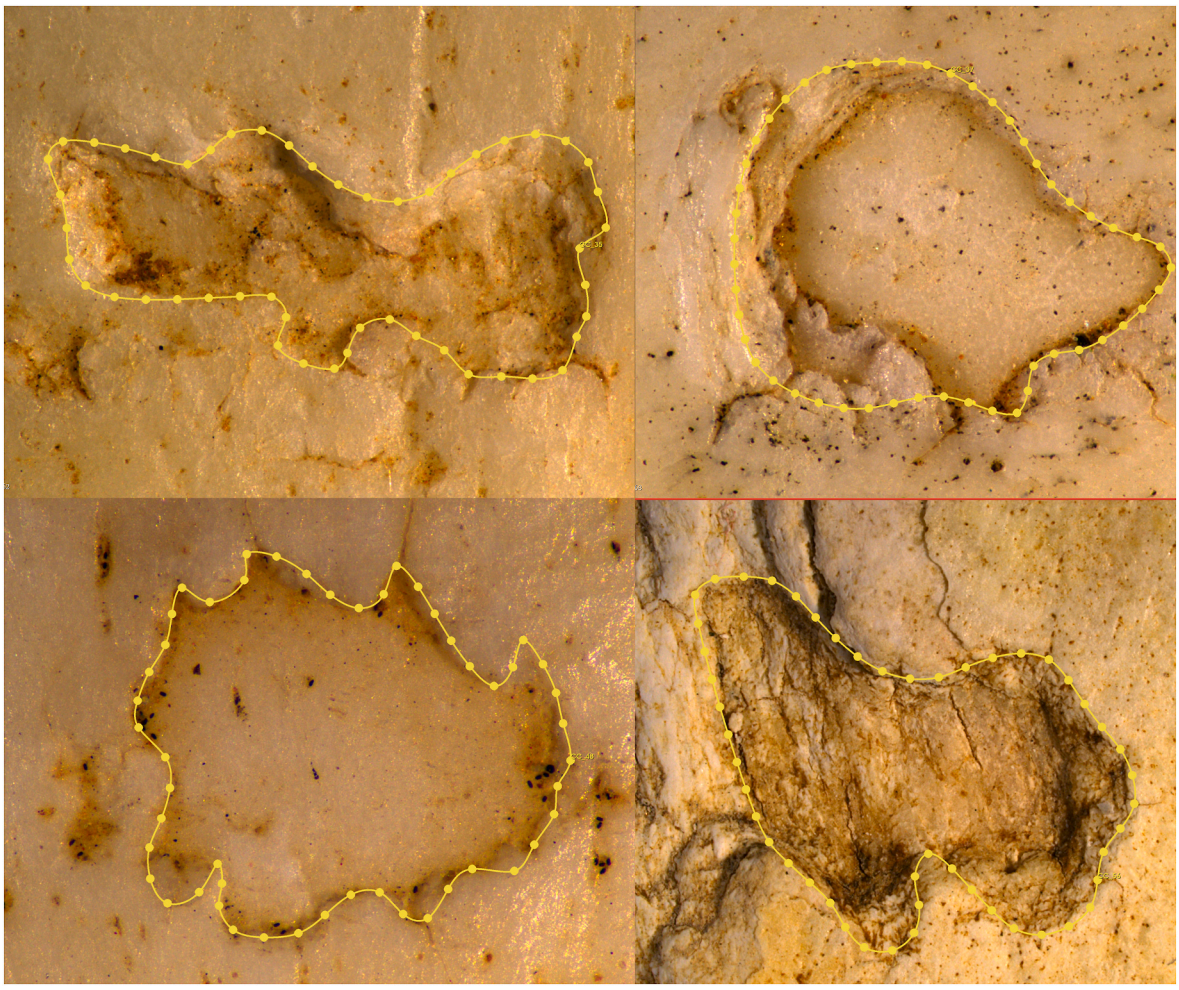

Fig. 3. Examples of some tooth pits displaying irregular outlines made by spotted hyenas.

The outline GMM analysis was carried out using R (www.r-project.org). EFA was performed on the sample selected using the Momocs R library (Picq et al., 2014; Bonhomme and Claude, 2020). First, to generate the tooth pit contours, images from the marks taken following the same protocol (see description below in CV method) were uploaded in 3D Slicer. Then the mark contour was semi-landmarked manually, and subsequently resampled 50 times, so that each mark was composed of 50 evenly-spaced semi-landmarks (Figs. 2–4). This enabled the correction of the initial complete contour and the tuning of each semi-landmark to adapt to the irregularities of each mark outline (Figs. 2–4). The resulting list of coordinated marks was uploaded in R for statistical analysis. Then, we proceeded to select the most adequate number of harmonics for Fourier transformation. In EFA, four coefficients per harmonic are created (two for each coordinate). These are normalized using the best-fitting harmonic to make them stationary to rotation and size. This also enables to make alignments of outlines of different shapes following either a single homologous point or a minimum radius to the centroid. This latter procedure enables stacking images to get a visual inspection of shape variability and outline trend (Fig. 6). Prior to EFA, we analyzed the cumulative sum of harmonic power to select the right number of harmonics, explaining 99.9% of the shape. We experimented with a range of one to 25 harmonics and probability thresholds of 90%,95%, 99% and 99.9% (Fig. 7). Then, we proceeded to assess the Fourier coefficients on the shape sub-assemblages of the four carnivores analyzed (Fig. 8). EFA involves the decomposition of shapes into a series of ellipses, each representing a harmonic component of the shape. There are four main coefficients associated with each ellipse. Each of them provide valuable information about the size, orientation, and position of the ellipses. which is an indirect way of describing shape. The first coefficient (here labeled with letter A) represents the size of the ellipse along its major axis. It indicates how elongated or stretched the ellipse is. A larger value indicates a longer major axis and thus a more elongated ellipse. The second coefficient (here denominated B) represents the ratio of the ellipse's minor axis to its major axis (indicating how much it differs from a circular shape). The third -theta (θ)- coefficient represents the orientation of the ellipse in relation to the x-axis. It indicates the angle at which the major axis of the ellipse is oriented with respect to the horizontal axis. This coefficient provides information about the rotational orientation of the shape. In contrast with the first three coefficients, that indicate the size, shape and orientation of the outline, the fourth -phi (ϕ)- coefficient represents the phase shift in the Fourier series. This means the degree of distortion between the shape outline and the elliptical trajectory in the Fourier series. Then, reliability in the outline reconstruction using the Fourier harmonics was tested with several independent marks (see example in Fig. 9). Using the Fourier transform, we obtained mean shapes for each of the four carnivores analyzed. The pairwise similarities of the two pairs of carnivores suggested that subtle variation should be approached using deformation grids through thin plate splines (Fig. 10).

After Fourier transformation, we performed a principal component analysis (PCA). The resulting components were then used for a multiple analysis of variance (MANOVA) to test if shapes from different carnivores were statistically similar or different. For that purpose, the Hotelling-Lawley test was used. Then, a pairwise MANOVA was made to

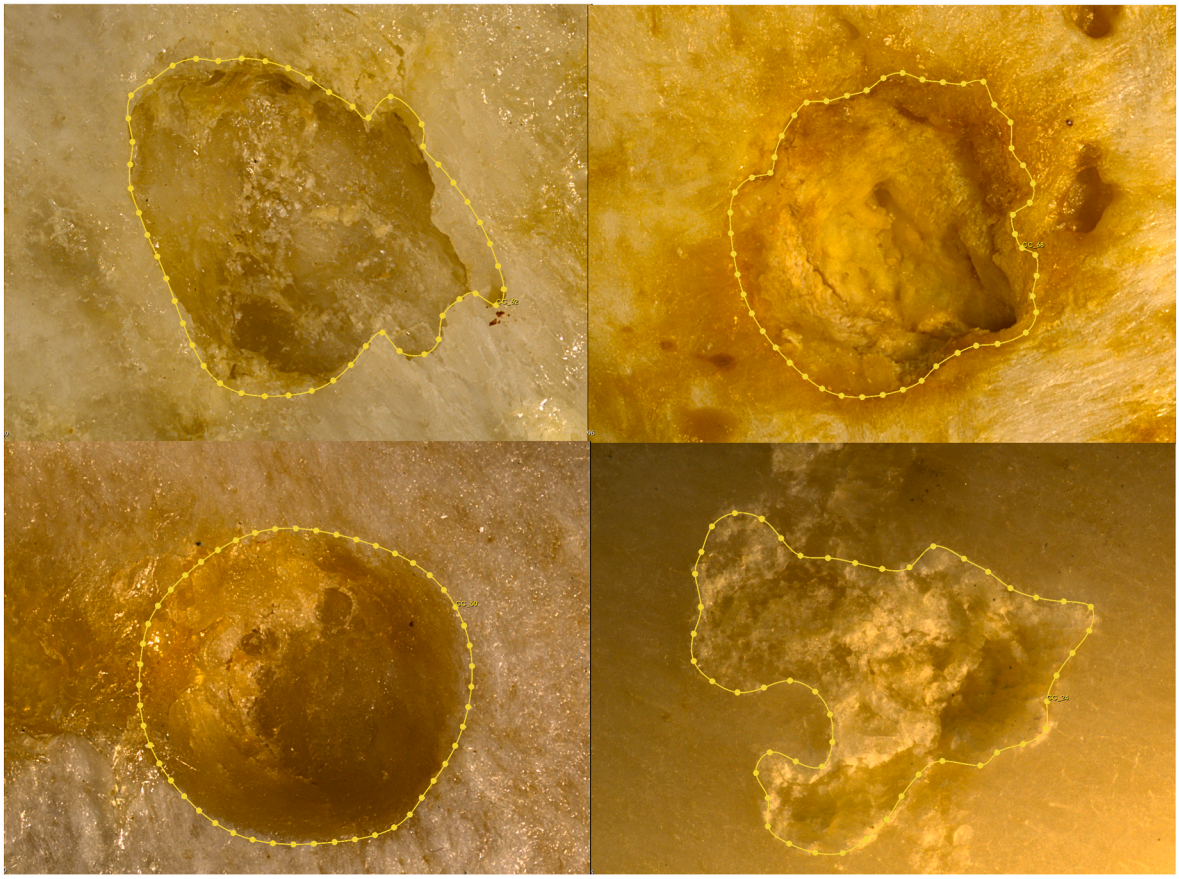

Fig. 4. Examples of some tooth mark pits displaying irregular outlines made by lions. Here, we also include a circular pit (lower left), which would have been difficult to correctly trace using a semi-landmark system.

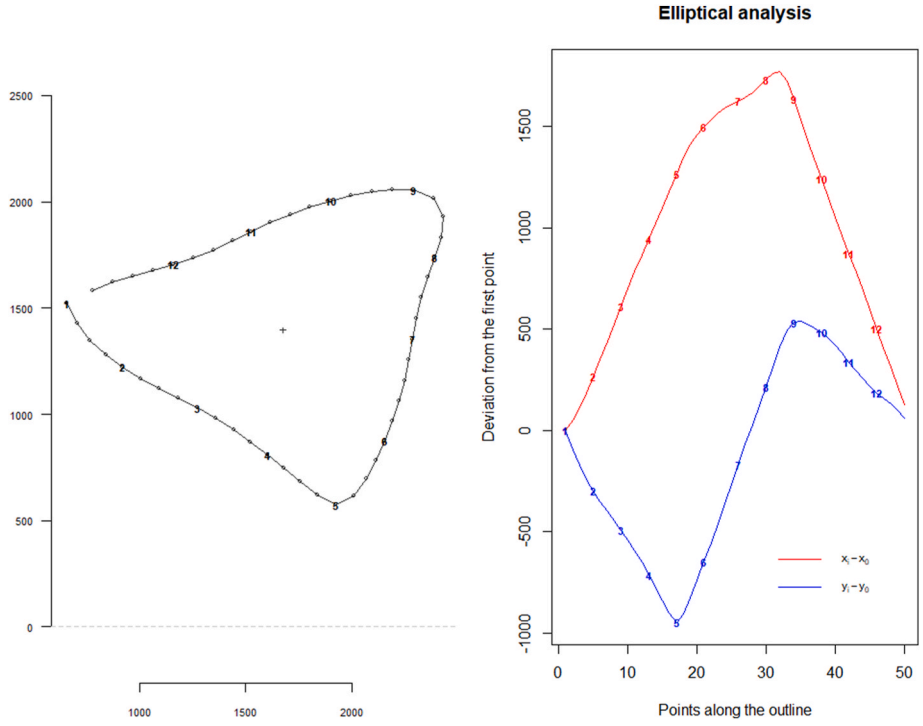

Fig. 5. Example of Elliptical Fourier Analysis of a triangular tooth pit made by a leopard.

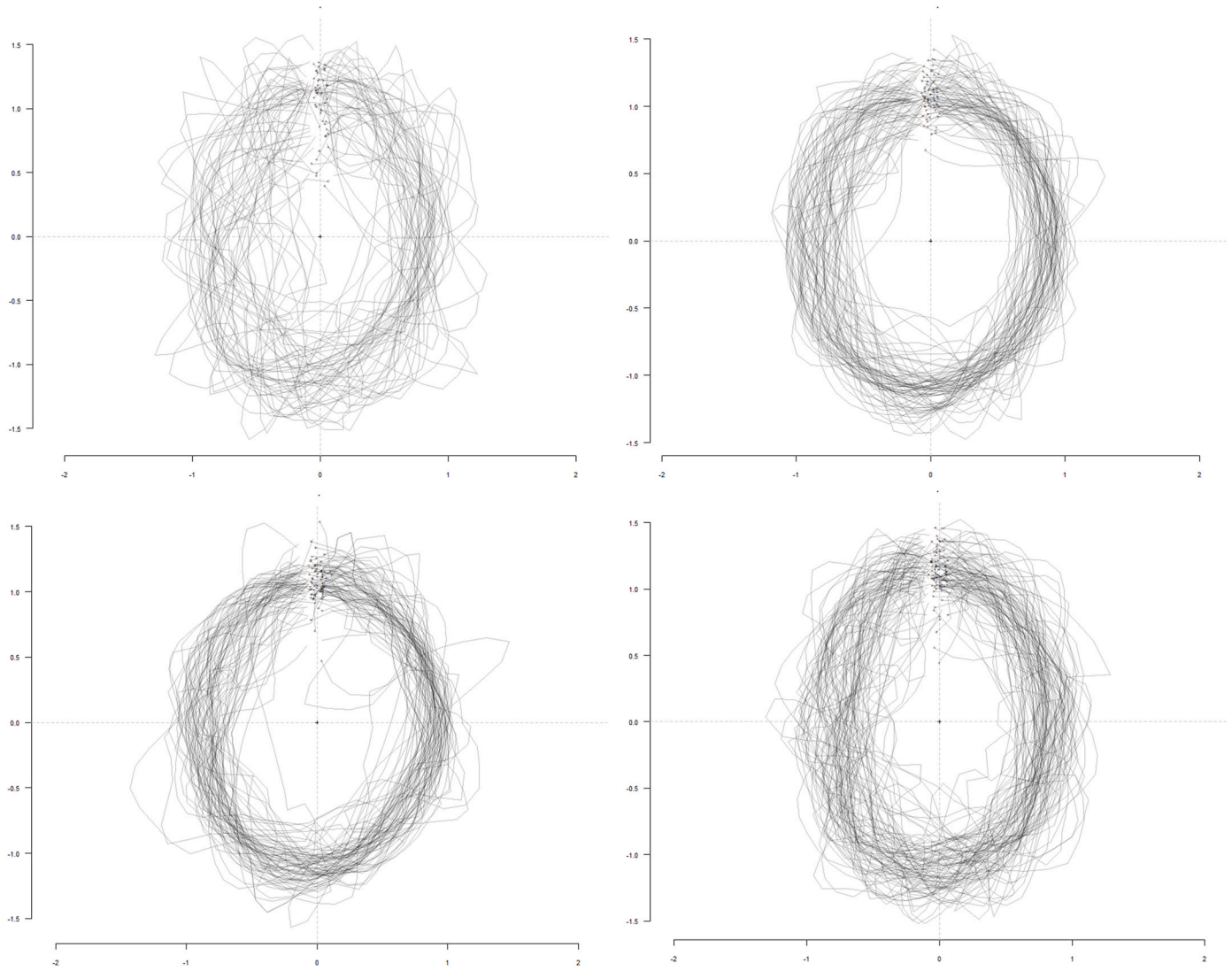

Fig. 6. Stacking of tooth mark outlines for each of the four carnivores, following minor distance to centroids.

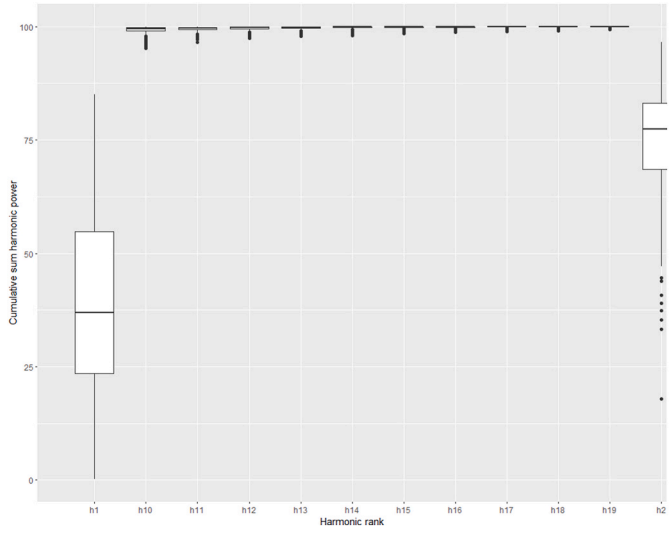

Fig. 7. Boxplot of the cumulative sum of harmonic power according to the number of harmonics.

detect significant differences between carnivore types.

Subsequently, the PCA (with 11 components explaining 99% of variance) were also used for a linear discriminant analysis (LDA) with the goal of classifying marks per taxon. We also implemented a machine learning (ML) approach as a comparative framework. In ML, we divided the sample into training and testing sets (with a balanced representation of the four taxa), and then, we implemented a PCA transformation on the training set. In order to prevent overfitting, we used a repeated (n = 5) cross-validation method. Then, we used three methods: a LDA, a random forest (RF), and a radial support vector machine (SVM). For this purpose we used R's "caret" library (Kuhn and Johnson, n.d.). We used the "Kappa" values for evaluating the performance of the classification models and for tuning them. We also used the "tuneLength" function so that we could automatize the number of different combinations of the model's tuning parameters (hyperparameters) to implement during the training process. We also provided a classification report (See Supplementary information) with the classification metrics, including accuracy, precision, recall, and F1-score.

2.4. CV method

Traditionally, computer vision methods have relied on Deep Learning (DL) approaches. DL is renowned for its capability to uncover complex patterns within large datasets. Methodologically, it is primarily

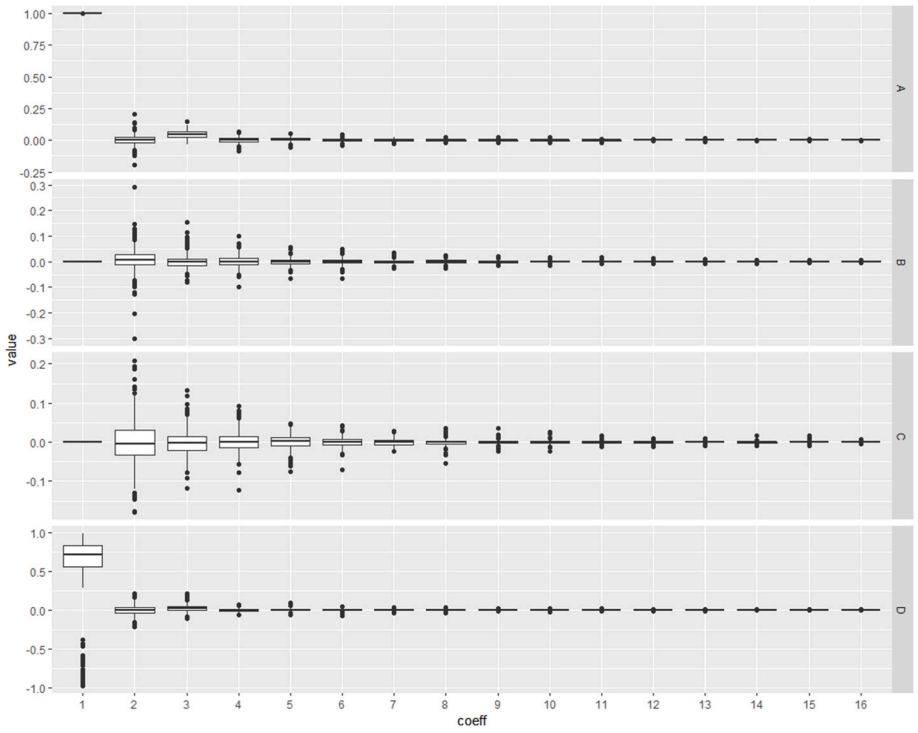

Fig. 8. Coefficients of the Fourier transform, according to the number of harmonics.

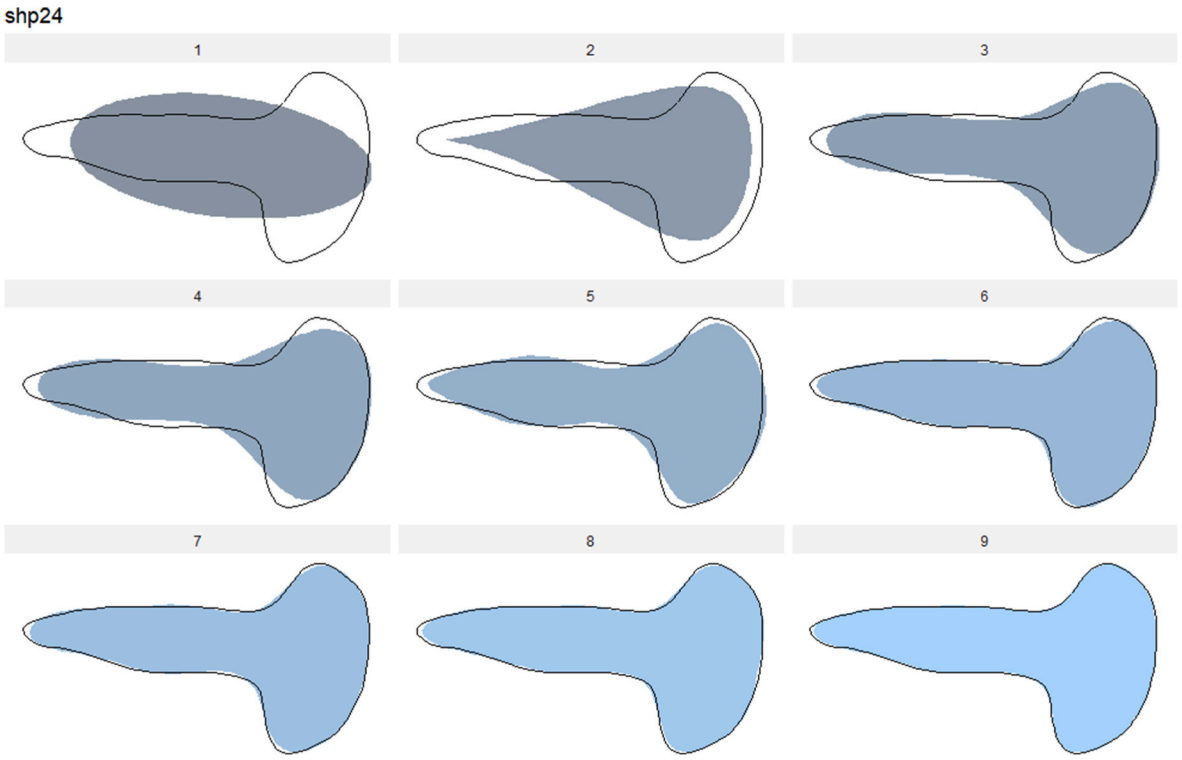

Fig. 9. Example reconstruction of the original shape of a tooth mark using 9 harmonics. This particular tooth mark, which is a score-pit hybrid, has been specifically selected for its complicated morphology as a good testing ground to the efficiency of shape reconstruction by different harmonic numbers.

based on artificial neural networks. These networks comprise interconnected layers of nodes, with each layer performing computations on incoming data and passing the results to subsequent layers. Data inside each node is transformed through the application of weights, and then each node applies a function to the weighted sum of its inputs. Data are successively transformed throughout each hidden layer. One of the distinguishing features of DL lies in its ability to automatically learn hierarchical representations of data through successive layers of abstraction. By processing extensive labeled datasets, DL models can identify intricate patterns, extract meaningful features, and use all of

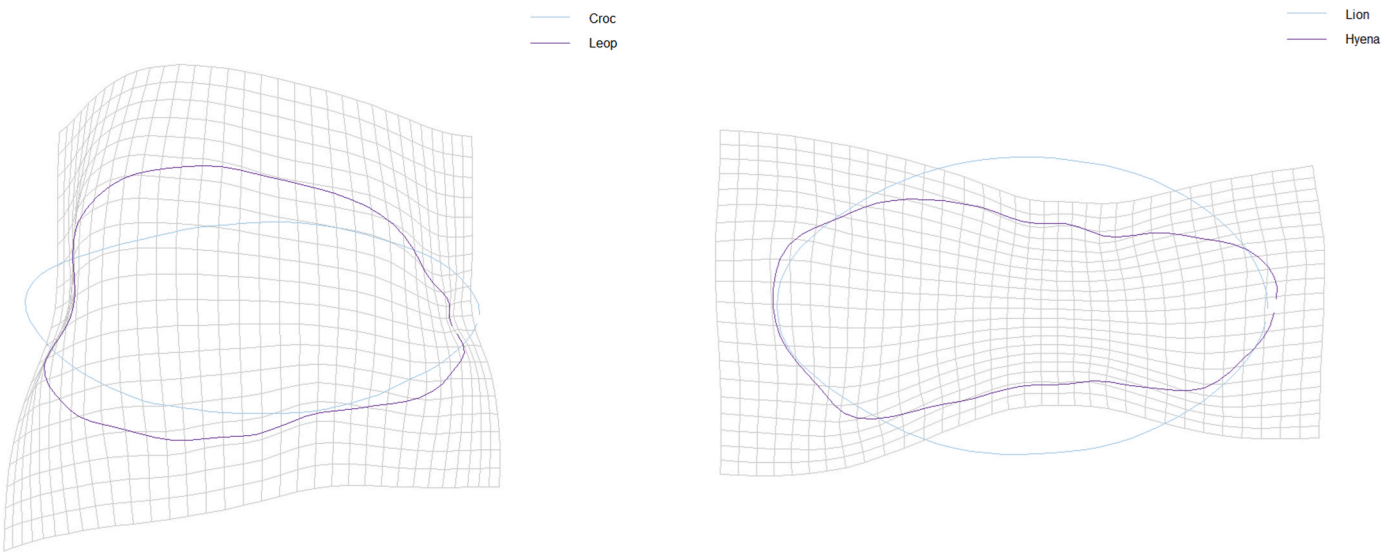

Fig. 10. Thin-plate splines of mean-shape tooth pits (see Fig. 17), comparing lion and hyena (right) and crocodile and leopard (left).

this to make accurate predictions across various tasks, being particularly successful in image classification.

One of the key advantages of DL is its scalability and adaptability to diverse datasets. Deep neural networks can handle vast amounts of labeled data and utilize powerful computational resources (such as graphic processing units [GPUs]) to expedite training and enhance performance of the models on testing sets or new data. This process is known as generalization. However, DL also has its limitations, particularly in scenarios where labeled data are scarce or limited. Training deep neural networks requires substantial computational resources and extensive libraries of images or datasets, posing challenges in applications where such resources are limited or unavailable. In contrast to the data-intensive nature of DL, few-shot learning (FSL) presents a more efficient and cost-effective approach to model training. FSL strategies aim to enable models to generalize from a minimal number of labeled examples, often as few as one or a few instances per class. This is what provides the name to the method (i.e., one-shot or few-shot analysis). This technique offers promising opportunities for machine learning applications, particularly in domains where data limitation poses significant challenges.

The primary obstacle in few-shot learning lies in facilitating models to extract generalized knowledge from limited examples and effectively apply it to unseen scenarios. It could be argued that the major drawback of FSL compared to DL is its more limited generalization. To address this challenge, few-shot learning algorithms employ various techniques, including generative modeling, transfer learning, or meta-learning (Vinyals et al., 2016; Snell et al., 2017; Finn et al. 2017; Flood et al., 2017; Wang and Hebert, 2016). Generative learning involves data augmentation techniques, such as Generative Adversarial Networks (GAN), so popular also in DL analyses. Transfer learning (TL) entails transferring knowledge acquired from one task or domain to another, allowing models to leverage pre-existing representations and adapt them to new tasks with minimal additional training data. It is, in essence, the same strategy used for a long time by DL analyses. Meta-learning, in contrast, focuses on "learning to learn" by presenting a diverse range of tasks to models, and enabling them to acquire meta-knowledge (commonly through meta-parameters) that facilitate rapid adaptation to new tasks, thereby increasing their performance at generalization. By training on a meta-learning dataset comprising multiple tasks, few-shot learning models can learn to infer underlying patterns and generalize from limited examples more effectively. The key here is to establish a balanced proportion according to the dataset between the number of examples (i.e., number of shots) and the number of training tasks.

In the present study, we will apply both CV approaches, with the goal of testing: a) if DL (DCNN) outperforms FSL in generalization, when data sets are extensive, and b) if FSL is a good replacement for DL when data sets are limited.

2.5. DL analysis

As described in more detail in the comprehensive analysis of the carnivore tooth marks (pits and scores), which served to generate models of direct application to the archaeological record (Domínguez-Rodrigo et al., 2024), here we used the same deep learning approach. The only difference is that we only used tooth pits and excluded tooth scores, focusing on capturing microscopic features within marks and their boundaries. The deep convolutional neural network (DCNN) format employed TL, utilizing pre-trained architectures, which had undergone prior training on a vast array of images. This pre-training enhanced the network's ability to recognize the intricate features of bone surface modifications. Comparative tests between TL architectures and raw convolutional neural networks demonstrated the superior performance of pre-trained networks. These pre-trained networks used ImageNet as a training ground. It is an image dataset containing millions of categorized images involving thousands of classes, that has been important for the evolution of DL networks through the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Images contained therein come in several resolutions, but are typically pre-processed to 224 × 224 pixels. ILSVRC contains 1000 classes of objects, but ImageNet contains many more.

For TL, we proceeded to load the pre-trained models using the convolutional base without the top classification layers. All layers were frozen to retain the features of the pre-trained models, The architecture of the TL networks consisted of the following custom classifier layers: a "flatten" layer to convert each TL model into a 1D array, a fully connected Dense layer with 128 units, a Dropout layer to reduce overfitting by randomly setting a fraction (30%) of the input units to zero during training, and an Output Dense layer with 4 units (for the four classes) using a Softmax activation.

We employed sequential and residual TL architectures, including ResNet 50 (version 1.0), VGG19, and Densenet 201, which have shown success in previous analyses of BSM. Individual models were generated, followed by the implementation of ensemble learning (EL)

methodology. EL involved using the three models as base learners, followed by a stacking process utilizing a random forest (RF) and an extra-gradient boosted tree (EGBT) as the meta-learner. Hyperparameter tuning for RF involved using 100 estimators as a fixed hyperparameter value during training, given our success in using such a parameter in previous models (Domínguez-Rodrigo et al., 2021a,2021b).

The majority of the tooth pits were documented and photographed with a microscope Leica Emspira 3. A few crocodile tooth marks (from the Altamira Zoo) were documented with a Hirox digital microscope (optics HR-2016). A few lion tooth marks (from Cabárceno) were documented with a an Optika binocular microscope. Prior to DL analysis, the original set of 549 tooth pit images was split into training (70%) and testing (30%) sets, as is customary in machine learning analysis. In the present work, we did not use validation/testing splits because of the very small size of the dataset. DL requires a large portion of images to be devoted to training (in this case 70%), but it also requires an extensive validation dataset. A validation/testing split would have produced a very small validation set (<100 images divided by four classes), which could have negatively impacted the training feedback. This would not be a problem for FSL methods, and this is why validation/testing splits were made in the FSL models (see below). Small validation sets may not adequately represent the diversity of the full dataset, potentially leading to a noisy or misleading feedback. This can lead to incorrect tuning parameters that may hinder model performance. If the validation set is too small, the model may overfit to the specific examples of the validation set. In this case, it might seem like the model is performing well (low validation loss), but in reality, the validation set does not provide a good measure of the model's generalization capability. A larger validation set guarantees more stable validation loss measurements, especially in high-variance datasets. With a larger validation set, estimates of the model's performance are more reliable, and the model can better detect overfitting during training.

Images were standardized by using each TL model's pre-processing functions, which also include color normalization to avoid potential biases introduced by differential coloring of images. Image augmentation techniques were applied to the training dataset to enhance training and mitigate overfitting risks. Augmentation procedures included random shifting, shear and zoom modification, horizontal flipping, and rotation. Image standardization and resizing to 250 × 200 pixels were conducted to facilitate algorithm discrimination of microscopic features within tooth marks. Tooth marks were taken at different magnifications, given that our CV method wanted to avoid biases determined by size, which in DL models would also introduce the potential bias of features from the surrounding bone cortical zone (Cifuentes-Alcobendas and Domínguez-Rodrigo, 2019). Therefore, by showing all tooth marks at similar sizes and minimizing the surrounding surface outside the pits, we are forcing the models to train by capturing features inside the tooth mark and its shape.

In addition to color normalization, a special protocol to determine potential biases from different colors of the images from the tooth mark dataset was implemented posteriorly as a control. In order to do that, we designed a function to perform color-based data augmentation on the image data set, following a protocol designed by Cifuentes-Alcobendas (2025). This function randomly adjusts the hue, saturation and brightness of the image in the HSV color space to introduce tone variations, making the model more robust to changes in lighting and color conditions. First, we input the images from their RGB color space. Then, we covert them to HSV (Hue, Saturation, Value) color space for its simplicity to operate tone-modifying operations on images. Hue (H) represents the color tone (red, green, blue) on a circular scale of 0–180°. The code is set to randomly modify the red-green-blue balance of the image to shift the original color to any of these three channels. Saturation (S) determines the intensity of the color by making it more vivid or pale. Changes in saturation can also provide benefits when working with images obtained from different imaging equipments. Value (V) indicates the brightness level, making the specific color darker or lighter. Changes in brightness also improve the resilience of the CNN when faced with images of varying contrast. The HSV conversion carried out by this function allows for independent assortment of hue, saturation and brightness of each augmented image. The randomization process involves the random selection of the adjustment (H,S or V). If the adjustment is made on hue, a random value between −10 and 10° is added to the hue channel. If the adjustment is made on saturation, the saturation channel is scaled by a random factor between 0.8 (decrease saturation) and 1.2 (increase saturation). If the adjustment is made on the brightness, the value channel is scaled by a factor between 0.8 (darker) and 1.2 (lighter). The range of the modifications is constrained to avoid sudden extreme changes to the images that may impact the CNN performance. After applying the randomly selected color augmentation, the image is converted back from HSV to RGB to be able to work with the CNN requirements. Then the image is further processed through the color normalization required by the network and re-scaled to fit the input size to be used.

In DL, there is always the uncertainty of how much the surrounding environment to the targeted object is also conditioning the classification results. In the particular case of BSM, the surrounding cortical area could also play a role beyond different coloration, by adding information on texture and cortical preservation that may bias the identification of agency. For this purpose, we carried out a cropping of the tooth pit sample by using most of the tooth marks included in the primary DL analysis and making sure that at least two sides of their outlines are tangential or quasi-tangential to the frame of the cropping window. With this procedure, we estimate approximately more than 30–40% of the original contextual information contained in the surrounding cortical area to each mark has been eliminated in the cropped image data set. If bone cortical information were a bias to the DL results, one would expect the accuracy-loss values to vary dramatically between the original and the cropped data sets. Lack thereof would imply marginal to no impact of contextual spatial information surrounding each mark. In addition to image cropping, color augmentation was also applied to this sample. Only when assessing the impact of these two important potential biasing processes (image color and surrounding cortical information), did we proceed to assess the utility of the original DL models on the complete unmodified tooth mark data set. The complete data set and code used for this control analysis using partially-cropped images can be located in the public repository: https://doi.org/10.7910/DVN/TDIEIT.

It could still be argued that even if the fraction of the surrounding cortical area was significantly reduced, it could still induce to over-reliant classification because of the potential of spatially located texture-preservation biases. In order to test this possibility, we carried out an ultimate test, consisting of completely cropping the tooth marks by encircling them in oval shapes that removed the surrounding cortical area completely. In absence of peripheral contextual information, model performance was conditioned exclusively by each mark's characteristics, especially given that color augmentation was also applied to this extremely cropped image sample. The complete data set and code used for this control analysis using completely-cropped images can be located in the public repository: https://doi.org/10.7910/DVN/BL5YSE.

DCNN models were developed using the Keras API with a Tensorflow backend, and computation was performed on a GPU HP Z6 Workstation within a CUDA computing environment. Activation functions and optimizers were selected through an exploratory phase, with the "relu" function and SGD optimizer yielding optimal results. The final fully connected layer utilized a "softmax" activation, with categorical cross-entropy as the loss function. Accuracy and F1 score values were evaluated to assess classification performance, considering the imbalanced nature of the dataset. Training employed mini-batch kernels of size 32, with testing utilizing kernels of size 20, and weight updates were performed using backpropagation over 100 epochs.

Regularization techniques, specifically Dropout with a rate of 30%, were implemented to mitigate overfitting during training. Training graphs for accuracy and loss were monitored to identify over- and

underfitting processes, ensuring the robustness of the trained models. Overall, the study aimed to develop DL models capable of accurately classifying tooth marks, contributing to the field of taphonomy and forensic anthropology.

2.6. FSL analysis

There is a wide array of FSL methods, several of them very different and divergent in their philosophy and structure: from siamese networks, to prototypical networks, relation networks, matching networks and model agnostic networks (Ravichandiran, 2018; Jadon and Garg, 2020). Model Agnostic Meta-Learning (MAML) will be the method of choice for the present work, because it is not dependent on any particular model architecture. MAML is designed to be model agnostic and it can be applied to various types of model architectures without significant modifications (Ravichandiran, 2018; Finn et al. 2017; Liu et al., 2024). The model's parameters are updated and fine-tuned through a meta-learning process, rendering them highly adaptable to new models and tasks. At the core of MAML are "tasks"", which represent individual learning problems or datasets that the model must solve. In the context of the few-shot learning method used here, each task might involve a small dataset containing only a few examples, and the goal is for the model to learn from these limited data. A task could be something like classifying images into classes, using a variable number of subsamples for different numbers of classifying tasks, or making predictions based on a set of features. Different tasks could involve the use of the same number of classes, but variable number of examples for each class. The meta-learning process in MAML works across multiple tasks, training the model to adapt quickly to any new task it is forced to resolve. This meta-learning process in FSL differs from the meta-learning approach in EL applied to DCNN, as described above.

MAML operates through a two-stage learning process. The first stage requires task-specific learning, where the model's parameters are updated based on data from a particular task. Such adaptation is done using gradient descent, which prompts the model to fine-tune its parameters to fit the task at hand. For each task, the model starts with a set of shared parameters, which are then adjusted through one or more steps of gradient descent using the task's support set (training data). In the second stage, MAML uses the updated task-specific parameters from multiple tasks to make additional adjustments to the previous model parameters using a query (i.e., evaluation) set. The aim of this process is to optimize these initial parameters so that the model becomes increasingly efficient, understanding by efficiency its ability to adapt to new tasks. Based on the results from the query set, the meta-learner adjusts the original parameters to improve the model's overall performance. This process ensures that the model learns an initialization that allows it to be functional at generalization. When the number of tasks increases, the model's parameters are modified in a way that maximizes performance. In summary, in the first stage, the model is trained using a limited set of examples and the parameters adapt to minimize the value of loss. This is done through the application of optimization methods. The second stage is the meta-learning process properly, whereby the model is tested against a validation dataset, which again updates the parameters to optimize its performance across a new set of tasks other than the initial training task. This allows the model to generalize effectively to unseen tasks with minimal fine-tuning or with additional training data.

The MAML method applied here is based on transfer learning, as in the DL method described above. We used the two best performing DL models (Resnet50 and Densenet201) as the MAML base models. All layers were frozen to retain the features of the pre-trained models. The MAML model was created with a sequential structure, using the TL model and a series of additional overlaid hidden layers, consisting of a convolutional layer with 512 filters and kernel size (3,3) (this layer applies convolutional operations to the features extracted from the base model), a global average pooling layer (which reduced the dimensions of the feature maps to a single value per feature map, resulting in 1D array per feature), a fully connected Dense layer with 512 units following a pooling layer, a Dropout layer, a Batch Normalization layer and a final Output Dense layer with Softmax activation. Two regularization measures were adopted. One was implemented inside the MAML model, consisting of Dropout, which is discarding a portion of nodes in a single layer to avoid overfitting. In this case, the MAML model used a 60% dropout, twice as much as the previous DL models. The other regularization method was "early stopping", which was monitored through the validation loss and a patience interval of 15 epochs, including weight restoration.

Images were normalized using each TL model preprocessing functions. These functions also incorporate color normalization to avoid potential biases introduced by differential coloring of images.The original dataset was divided into training (70%), validation (15%) and testing (15%) sets. This resulted in 384 training images and 166 images evenly split for validation and testing. As in the use of DCNN, augmentation procedures included random shifting, shear and zoom modification, horizontal flipping, and rotation. Image standardization and resizing to 250 × 200 pixels were also conducted. The meta-optimizer used was Adam (Adaptive Moment Estimation), with a learning rate of 1e-03. During model compilation, loss was measured using "sparse categorical cross-entropy". We initiated the model with 100 epochs, a validation interval of 1, and variable numbers of tasks and shots. We carried out three sets of shot-task combinations, according to the number of shots: low (5 shots and 10 tasks), medium (10 shots and 9 tasks), and high (15 shots, 6 tasks). The latter involved a substantial amount of computation, and involved 360 images (15 images x 4 classes x 6 trials), which was below the number of images in the training dataset. This is why the configuration of the task-shot module was done with "replace = False". If True, we could have amplified the number of shot-task by resampling images, despite the limitation of the dataset. Previously, we had used a resampling method (with replace = True) that enabled the use of high numbers of shots and tasks, involving the use of hundreds and thousands of images. The accuracy results were similar to those obtained here using the restricted approach (i.e., limiting the number of shot-task to the number of training images), but their loss values were higher and so were their overfitting ranges. For this reason, we opted for a restricted approach, limiting the number of shots and tasks, as described above. This yielded much more stable models with low or no overfitting depending on the set. The use of these three sets (low, medium, high) was based on the comparison of which factor increased accuracy: low shot-high task (low set-medium set), or high shot-low task (low/medium sets-high set). The analysis was carried out with a GPU HP Z6 Workstation within a CUDA computing environment, and using the Keras API with a Tensorflow backend.

Both types of analyses (DCNN and FSL) have additional common features. For example, to address class imbalance in the training-testing split, each carnivore has contributed with 70% of tooth marks to the training set and 30% of tooth marks to the evaluation/testing sets. Data augmentation was applied equally to both sets.

2.6.1. Archaeo-paleontological analysis

We decided to test our CV models against a small set of archaeo-paleontological tooth pits, comprising both marks with quasi-optimal preservation and diagenetically modified tooth marks. The goal was not to analyze the assemblage to infer paleo-agencies, but to compare individual and ensemble model performance on a limited number of images, according to their preservation. We selected the tooth mark set from the >1.8 Ma FLK North (FLK N) site at Olduvai Gorge (Tanzania) (Vegara-Riquelme et al., 2023). FLK N stands out as a remarkable palimpsest, where both hominins and carnivores repeatedly occupied the area over numerous years, potentially spanning centuries or millennia. Carnivores predominantly contributed to bone accumulation at the site. Positioned atop Bed I and located less than 100 m north of the FLK 22 Zinjanthropus (FLK Zinj) site, FLK N boasts the distinction of

being "the thickest early Pleistocene archaeological deposit currently known" (Domínguez-Rodrigo et al., 2010). Discovered in 1960, the site's vertical sequence comprises a continuous stratigraphic sequence, including Bed I and the lower portion of Bed II. Initial excavations revealed three archaeological levels above Tuff IF and six more below it. Subsequent excavations by TOPPP (The Olduvai Paleoanthropology and Paleoecology Project) uncovered additional underlying levels (7–9). FLK N represents an exemplary case of carnivore-accumulated assemblages with minimal hominin involvement, despite the continuous presence of stone artifacts throughout the deep vertical deposit. The co-occupation of the same space by hominins and carnivores has been suggested to occur independently or with minimal interaction. Felids were initially identified as the primary accumulators and consumers of carcasses at the site, with intermittent hyenid intervention observed over extensive temporal periods (Domínguez-Rodrigo et al., 2007). Previous taphonomic investigations utilizing conventional taphonomic methods have established bone modifications by both felids and hyenids across multiple levels at FLK N. A more recent analysis indicates a stronger hyena input, especially in most broken long bones (Vegara-Riquelme et al., 2023). Here, we selected four well-preserved tooth pits and three diagenetically-morphed tooth pits as a contrasting set. They were initially classified using the original images, taken at x30, using a Leica S9i microscope. Prior to analysis, images were standardized using the pre-processing framework for each of the three models (Resnet50, Densenet201 and VGG19) independently.

The classification was made using the best performing DL and FSL models. We then proceeded to classify the selected well-preserved and the diagenetically modified tooth pits from FLK N. For DL, the analysis was carried out by submitting every tooth pit to each individual model, and later to the ensemble generated by the three models. We use a dual ensemble approach. The first one involved joint classification by the model weights. The second approach (weighted ensemble) involved using weighted transformation of the accuracy rates of each model, in order to have the most accurate models have a greater weight in the final classification decision. It is this latter model that should generally be used when several models are simultaneously used as a joint classifier. For FSL, we simply used the best perfoming Resnet50 and Densenet201 models resulting from the different combinations of shots and tasks, and then we also used an ensemble model.

Here, we establish a minimum confidence threshold based on the probability of classification (>70%), with optimal confidence when the probability is >90%. We apply these thresholds to individual models, but the reliability of the classification was determined by convergence of different models on the same classification decision and with overall probabilities >70% and >90% respectively.

3. Results

3.1. GMM analysis

The normalization of tooth mark outlines per taxon following an alignment along the minimum radius showed that crocodiles have the most diverse morphological range (Fig. 6). The contour analysis clearly shows that symmetrical oval shapes are a minority in all the four carnivores studied (Figs. 11–14). This underscores the need to include all shapes of tooth pits prior to claiming that discrimination among agents is feasible. Oval-circular shapes are more common in the lion sample and least common in the crocodile sample. The large variety of asymmetrical and irregular shapes in crocodiles, when compared to the other taxa, could be the result of crocodile behavior in pulling, jerking and biting their prey during hunting or competitive bouts against other individuals.

The calibration of harmonic power on the complete tooth mark sample yielded a range of 5 (threshold = 90%) to 16 harmonics (threshold = 99.9%) (Fig. 7). The latter was used for the Fourier transformation. The coefficients indicate that capture of length was achieved with 12 harmonics; width with 15 harmonics; rotation with 16 harmonics and positioning with 12 harmonics (Fig. 8). Outline reconstruction was very accurate using 6 harmonics (95% confidence) and higher (Fig. 9). A PCA yielded a two-dimensional solution explaining 94.1% of the sample variance. The overwhelming majority of variance was explained by the first component (91.5%) (Fig. 15).The point cloud showed two poles along the first component axis, explained by a majority of marks displaying irregular outlines close to open-shaped ovals. At the other end, a minority showed more elongated shapes. The confidence ellipses and the hull shapes for each taxon show intense overlap, as has also been the case for several previous GMM analyses of tooth pits.

The subsequent MANOVA analysis yielded significant differences in the outline shapes of the four carnivores (H-L = 0.271, df = 3, p = 0.000124). A pairwise MANOVA test found significant differences (<0.01) between crocodile and felids, and hyena and felids. The comparison between crocodiles and hyenas showed weak significance (p = 0.08), which was similar to the significance observed when comparing

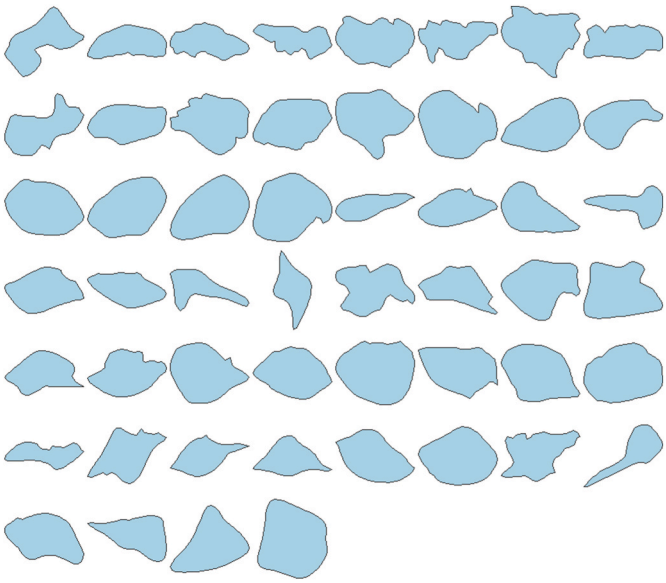

Fig. 11. Range of outlines of the tooth pits documented in the crocodile sample.

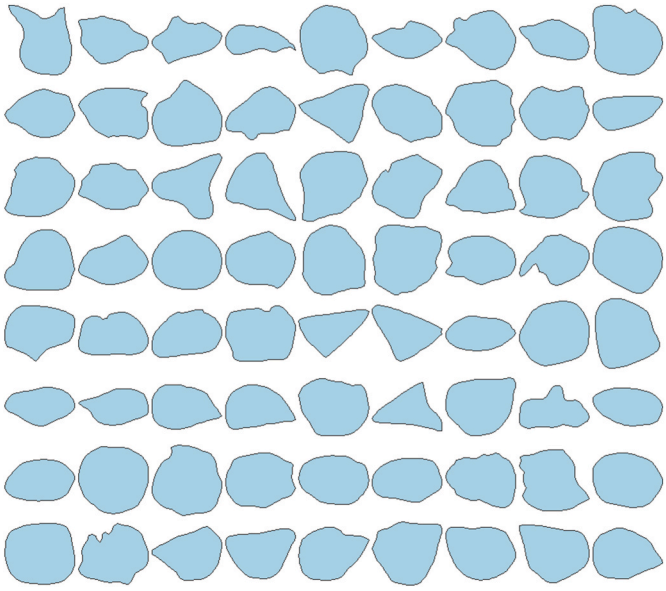

Fig. 12. Range of outlines of the tooth pits documented in the leopard sample.

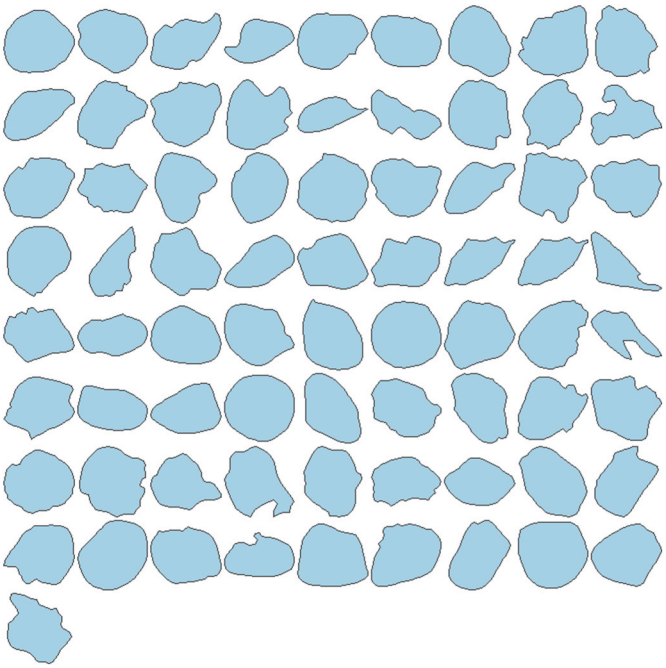

Fig. 13. Range of outlines of the tooth pits documented in the lion sample.

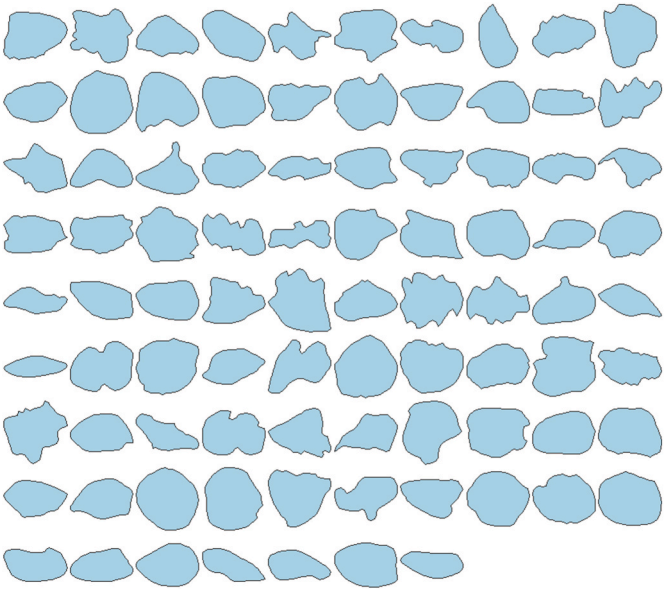

Fig. 14. Range of outlines of the tooth pits documented in the hyena sample.

leopards and lions (p = 0.07) (Table 1).

The PCA components (11 were retained) were then used for a LDA, resulting in an overall accuracy (leave-one-out cross-validation) of 34.9%. Class accuracy ranges between 23% (leopards) and 44% (hyenas). Kappa is 0.11 and the average precision, recall and F1 score are 0.34 (Table 2) (Fig. 16). Despite the significant differences found in the MANOVA test, the four carnivores show a wide degree of overlap that prevents reaching a higher accuracy in differentiating their tooth marks. When using LDA through a ML framework (using all the components), the testing set was classified with an accuracy of only 25.3% (Supplementary Information). This classification was slightly higher when using a random forest (accuracy = 31.3%) and a support vector machine (accuracy = 33.7%) (see the classification reports in the Supplementary Information). This underscores the low resolution of the bidimensional information contained in the outline of tooth pits made by the four carnivores analyzed, being very similar to random classification. This is probably due to the great variation in tooth pit outlines generated by each of the four agents. The discrimination of agency using the complete outlines is lower (<35%) than when using a set of 20 semi-landmarks (51%) (see Supplementary Information). This generates an apparent contradiction resulting in higher resolution when using substantially less information about tooth mark shape. It can also be the by-product of artificial shape homogenization, which can more easily accommodate allometric information and be impacted more by forms than by shape (see more in Discussion). Given that the complete outline generated non-resiolutive models, we excluded GMM from the analysis of the selected archaeo-paleontological BSM (see below).

Mean shapes for each of the four carnivores show more elongated forms for crocodiles and leopards, and more oval-rounded forms for hyenas and lions (Fig. 17). Thin-plate splines show that crocodiles displayed a longer elongation than leopards (Fig. 10a). Likewise, lions generated more rounded marks than hyenas (Fig. 10b).

3.2. CV analysis

3.2.1. DL analysis

Training of two of the models showed clear indications of some overfitting (Resnet50, Densenet 201), whereas for VGG19, the curve of the training and validation tests was similar (Fig. 18). Despite the slight overfitting, both Resnet50 and Densenet 201 achieved high accuracy (Table 3). The most successful model was Resnet50, with 81% of accuracy, followed by Densenet 201 (77.3%) and VGG19 (73%). The overall accuracy was lower than when using the combined sample of tooth pits and scores, which enabled the same models to train with a substantially larger sample per agent (Domínguez-Rodrigo et al., 2024). In that case, Resnet50 achieved 88% of accuracy, and all models classified correctly >80% of the testing set (Domínguez-Rodrigo et al., 2024). The lower accuracy in the present work may have something to do with the significantly smaller training and testing sets. Oddly, the ensemble analysis yielded a very similar result in both cases (see below).

The classification indicators (precision, recall and F1-score) show similar classification rates in all four carnivores (See Supplementary Information). For example, for the most successful model (Resnet50), the F-1 scores for the taxa are: crocodiles (0.82), hyenas (0.85), leopards (0.75) and lions (0.82).

The EL analysis achieved an accuracy of classification on the testing set of 80.4% when using both a random forest and an extra-randomized boosted tree classifiers. In the tooth pit-score sample the random forest learner yielded 82.5% of accuracy (Domínguez-Rodrigo et al., 2024). It is interesting to note that EL is affected less by the training sample size than each model individually.

In order to test the reliability of these results for identifying agency through correct classification of the experimental marks, we reproduced the same analyses by introducing color augmentation (see Methods). In this case, color augmentation did not only not handicap the performance of the models, but boosted them instead by improving their generalization. The Resnet 50 model yielded 82.21% of accuracy in the classification of the testing set. The Densenet 201 model yielded 78.53% of accuracy, and VGG19 reproduced a similar accuracy to that obtained in the original analysis lacking color augmentation (73.62%). On average, there is only a difference of 1.12% between the original and the color augmented sample in favor of the latter, which indicates that the different coloring of images in the experimental set did not impact the results in any meaningful way (Table 3).

When introducing mark cropping (to remove potential data leakage introduced by the contextual information of the surrounding cortical area) and color augmentation, the models yielded similar accuracy to the original complete image analysis: ResNet50 (80%), Densenet 201 (76.67%) and VGG 19 (71.67%) (Table 3). There is an average of 2% difference between the complete image data set and the cropped image

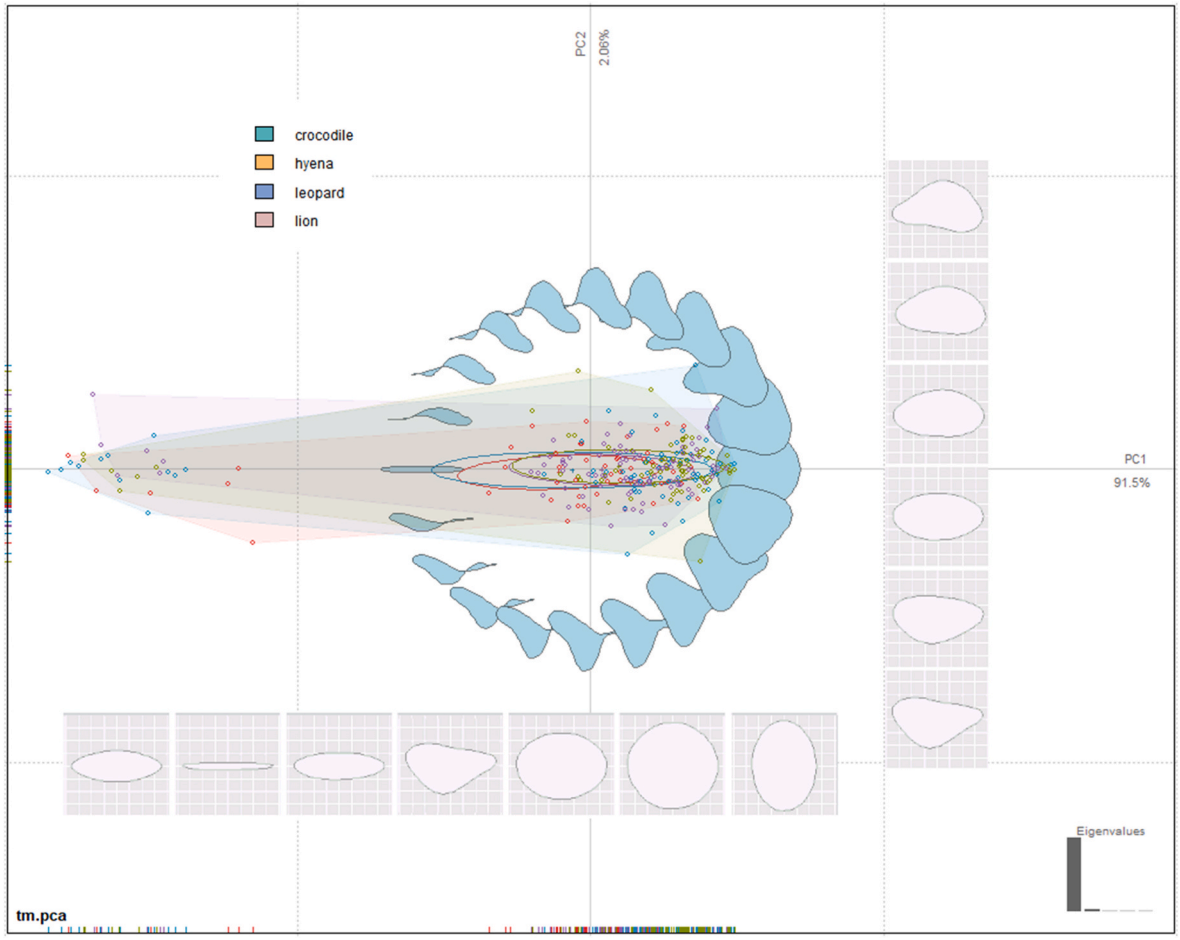

Fig. 15. PCA of outline shapes generated by the four carnivores, including 95% confidence ellipses and taxon-specific hulls.

Table 1
MANOVA scores for pairwise comparison of tooth pits, following a Fourier analysis.

	Df	Pillai	approx F	numDF	den DF	Pr(>F)
crocodile-hyena	1	0.1273	1.685	11	127	0.08399
crocodile-leopard	1	0.2545	3.476	11	112	**0.00033**
crocodile-lion	1	0.268	3.762	11	113	**0.00013**
hyena-leopard	1	0.1644	2.629	11	147	**0.00429**
hyena-lion	1	0.1403	2.195	11	148	**0.01749**
leopard-lion	1	0.1254	1.734	11	133	0.07248

Table 2
LDA metrics per class of the Fourier analysis of the tooth pits.

	precision	recall	f1-score
crocodile	0.364	0.308	0.333
hyena	0.355	0.448	0.396
leopard	0.309	0.236	0.268
lion	0.361	0.37	0.365

data set (both using color augmentation), and <1% between the original complete image data set and the combined cropped and color augmented data sets (Table 3), suggesting that contextual information did not bias in any significant way the original accuracy values reported by the three DL models.

This is further reinstated when using only the fully-cropped tooth mark data set (where tooth marks appear without the surrounding cortical area) (Table 3). In this case, the average difference in accuracy

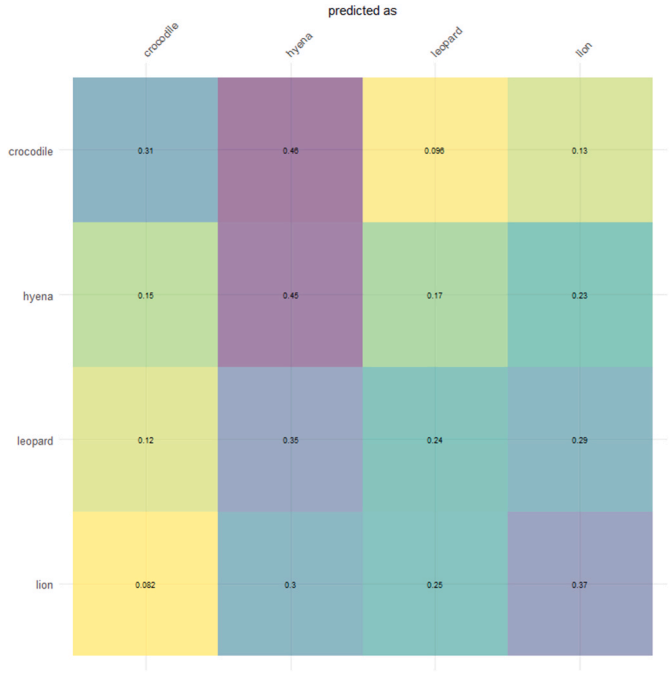

Fig. 16. Confusion matrix for the linear discriminant analysis (LDA) showing the percentage of correct classification (diagonal) and misclassification of the tooth pits attributed to the four carnivores.

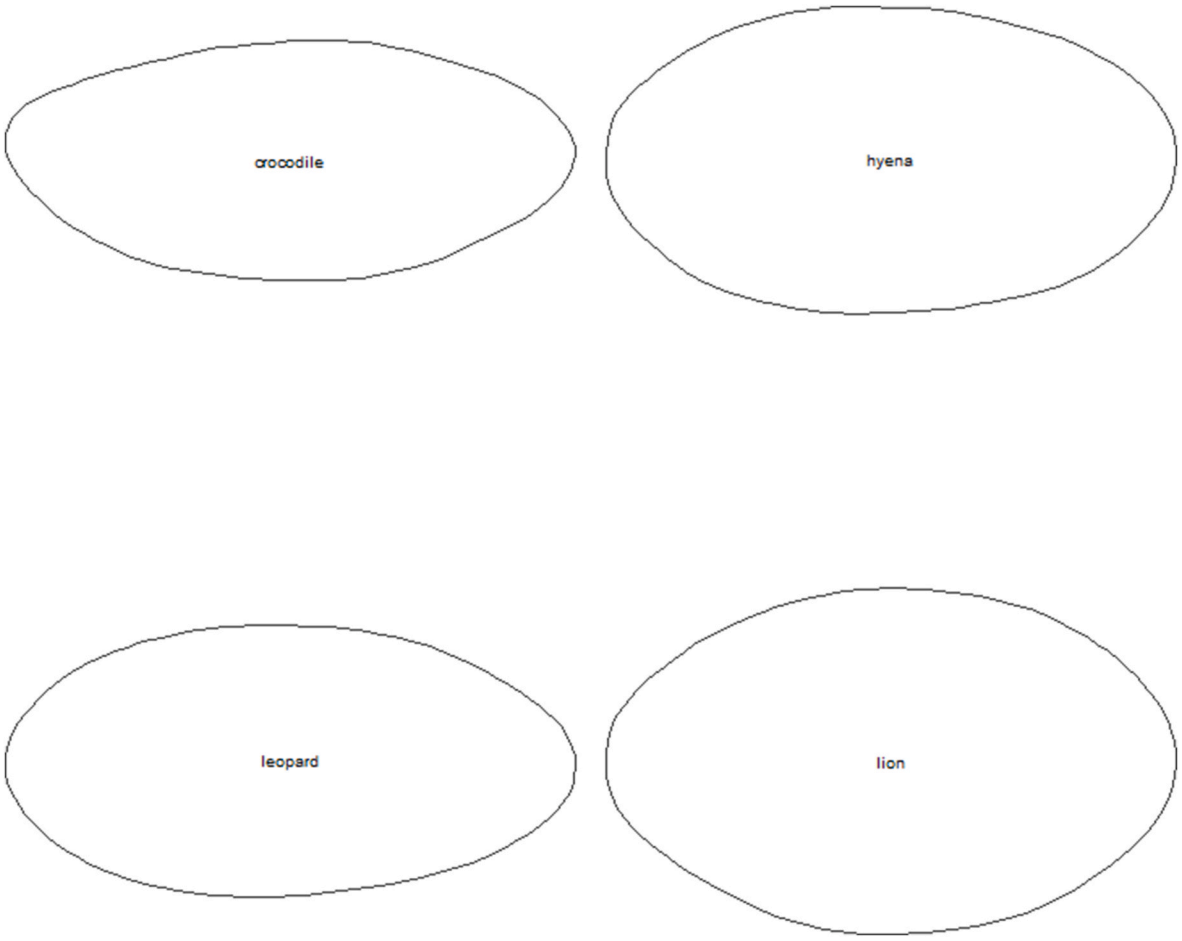

Fig. 17. Mean shapes of the four carnivores analyzed (crocodile, leopard, hyena and lion). The former two display more elongated mean shapes. See Fig. 10 for a thin-plate spline pairwise comparison.

using the three models between the complete non-color augmented dataset and the completely cropped and color-augmented dataset was <1% in favor of the latter. This underscores the rather marginal or null effect that contextual bone texture/color may have on the classification efficiency of the models. This is further supported by the fact that some models seemed to also improve their generalization capabilities when the surrounding cortical area was removed and color was not an issue. This is the example of Densenet 201, whose generalization performance improved from 77.3% to 78.56%, or VGG19, which improved from 73% to 74.23% of accuracy.

3.2.2. FSL analysis

The use of a Resnet50 base model with a 5-shot approach (ten tasks per epoch) yielded an accuracy of 78.31% on the testing set (76.83% on the validation set) (Fig. 19). Early stopping interrupted the 100-epoch original cycle at epoch 28 when no improvement was documented over the preceding 15 epochs (Table 4). The medium-sized model (10-shot, 9 tasks) had early stopping produce a final model at epoch 30, with slightly higher accuracy: 79.52% of the testing set. The high computational model (15-shot, 6-task) did not yield a better performance on the testing set (accuracy = 77.11%). In all the sets, the testing accuracy was higher than the validation accuracy (71.95%–76.83%). It is interesting to note that all the Resnet50 models displayed some degree of overfitting, despite which, accuracy on the testing sets was high (Fig. 19).

The use of Densenet201 as a base model yielded lower accuracy, as was the case for DL (Fig. 20). The low model (five-shot 10 tasks per epoch) yielded an accuracy of 69.88%; the medium-sized model (10-shot, 9 tasks) was slightly lower (68.67%), and the high-computational model (15-shot, 6 tasks) yielded 66.27% of accuracy (Table 4). In all cases, early stopping prevented the 100-epoch cycle to finish, and models were stopped between the 26 and 93 epochs. In the case of the Densenet model, the validation accuracy was slightly higher than the one yielded by the testing set (range = 79.27%–82.93%). It is interesting to note that despite having reached a lower accuracy than Resnet50, the Densenet201 models produced marginal or no overfitting (Fig. 20). This clearly shows that the degree of model overfitting is not related to performance in generalization. With Resnet50, it was the intermediate model (balance between number of shots and tasks) that yielded the highest accuracy; almost 80%. With Densenet201, both the low and intermediate models produced similar results (about 70%). This suggests that meta-learning depends on the projection of a significant number of tasks, more so than on the number of shots.

Although they cannot be compared because the generalization method for both approaches (DL and FSL) differs, it is remarkable to see that the performance of FSL at generalization is broadly comparable to that documented for DL, with only a slightly lower accuracy by the FSL models compared to the DL models (Table 4; Figs. 19–20). This implies that the generalization performance (given the imbalance nature of the dataset) is almost as good in the FSL models as in the DL ones. The F-1 score values for models from both approaches are substantially similar (see Supplementary Information). However, when comparing the performance of Resnet50 in the DL and FSL analyses, the results are very close, being virtually identical both in accuracy and loss. This underscores the great potential of FSL when samples are not large enough, with a performance potentially comparable to DL in specific models and architectures.

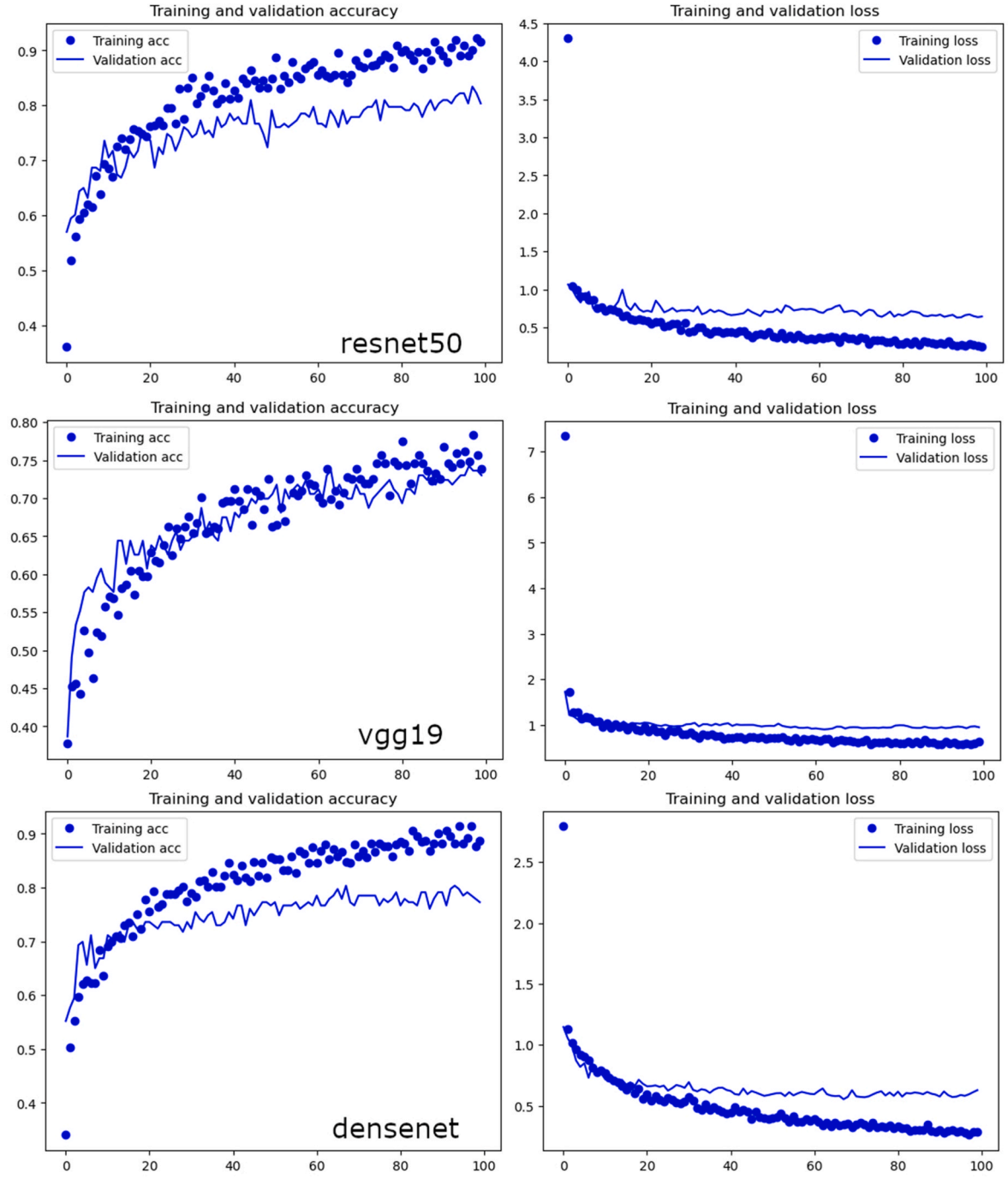

Fig. 18. Accuracy and loss of the DL models during the training and validation process: Resnet50, VGG19 and Densenet201.

3.2.3. *Archaeo-paleontological analysis*

The application of the most accurate DL model (Resnet50) to the quasi-optimal preservation set (Fig. 21) yielded a classification of 3 marks out of 4 as "hyena", with probabilities >90% (Table 5). Densenet201 yielded the same classification on the same marks with probabilities ranging between 82.7% and 91%. VGG19 only classified two marks the same as the previous models, but with a probability >90% in both cases. Only one mark (FLKN+49d + 3) was classified randomly as "lion" by Resnet50 (prob = 57.2%), "crocodile" by Densenet201 (prob = 48.3%), and "hyena" by VGG19 (prob = 96.9%) (Table 5). This makes the classification of this particular mark unreliable. The ensemble analysis classified all four marks as "hyena", but only two marks exhibited probabilities >90%. When using a weighted ensemble, all marks but three comply with the minimum threshold of confidence (prob=>70%). Again, FLKN+49d + 3 is classified as "hyena" but with lower probability (53%) and, therefore, low reliability. A similar result was obtained when using the FSL models (see Table 7 below).

When the original images with the four tooth pits were cropped, in order to minimize the impact of information from the perimetrical area surrounding the pit (Fig. 22), Resnet50 classified all images as "hyena" (three of them with prob>89%), although the previous controversial mark is classified as such with low probability (Table 5). Densenet201 reproduced the same classification as with the uncropped images, but with different probabilities. The only difference is that the classification probabilities are on average substantially higher. VGG19 also increased the probabilities of the classified images, but switched the interpretation

Table 3
Accuracy and loss information of each of the three models used. The combination of optimizer and activation function for each model is the one showing the highest accuracy. Accuracy and loss scores are shown for the original complete image sample, the same sample with color augmentation (upper section), and for the cropped sample and extremely cropped samples (lower section), both with color augmentation.

Model	Optimizer	Activation function	Without color augmentation		With color augmentation	
			accuracy	loss	accuracy	loss
Resnet50	Adagrad	swish	0.81	0.7615	82.21	0.753
VGG19	Adagrad	swish	0.73	0.948	73.62	0.834
Densenet 201	Adagrad	relu	0.773	0.628	78.53	0.597
			Cropped with color augmentation		**Completely cropped with color augmentation**	
Resnet50	Adagrad	swish	80.01	0.613	76.46	0.645
VGG19	Adagrad	swish	71.67	0.685	74.23	0.71
Densenet 201	Adagrad	relu	76.67	0.481	78.56	0.561

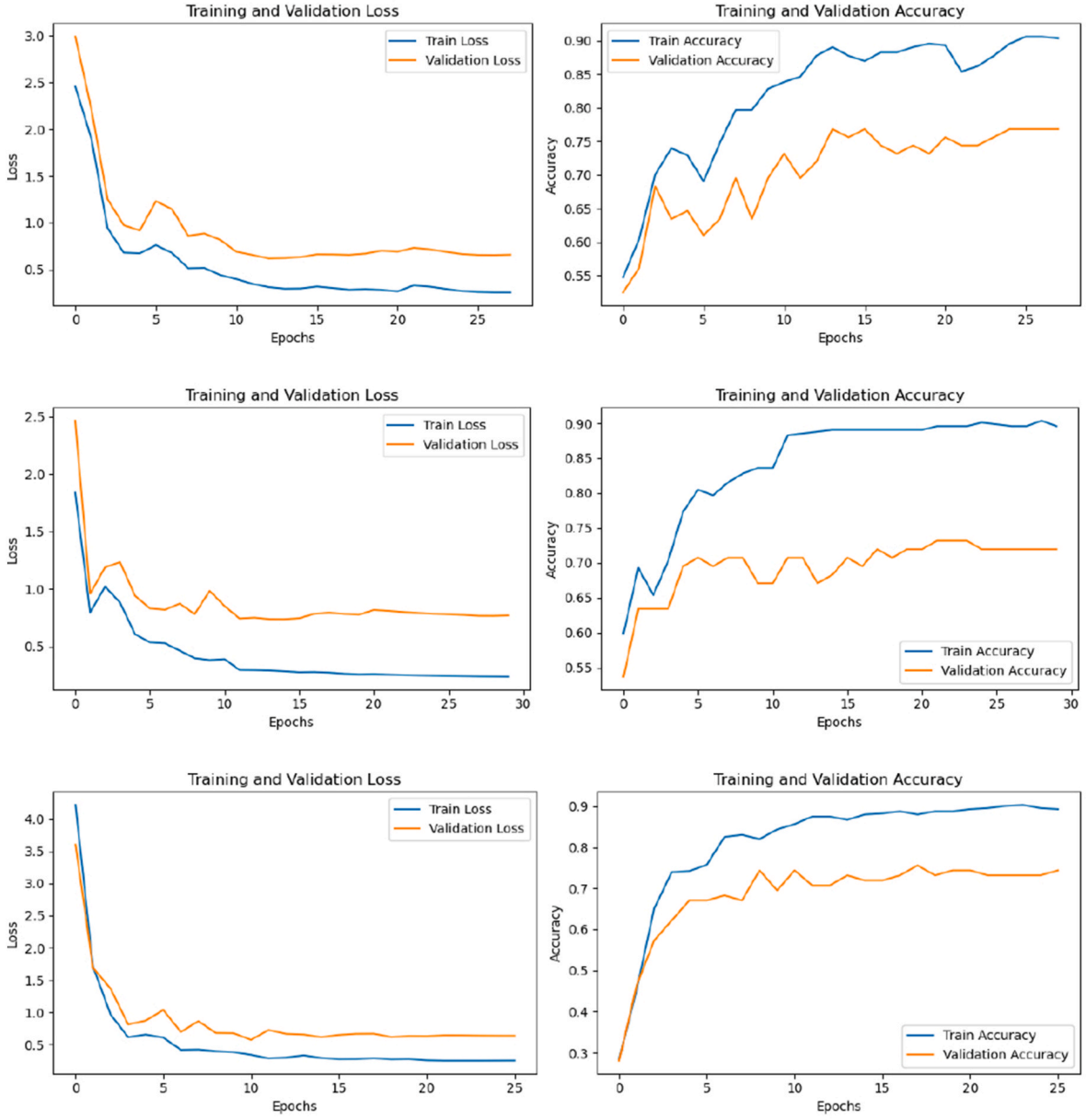

Fig. 19. Training graphs (accuracy and loss) of the FSL Resnet50 model with a combination of 5-shot and 10-tasks (upper), 10-shot and 9-tasks (middle), and 15-shot and 6-tasks (lower).

Table 4
Number of shots and tasks for the Few-Shot learning analysis, including the validation and testing accuracy and loss. The relationship between number of shots and tasks is set to require a smaller number of images than the size of the training data set.

Model	number shots	number tasks	validation accuracy	validation loss	testing accuracy	testing loss
Resnet50	5	10	76.83	0.254	78.31	0.606
	10	9	71.95	0.770	79.52	0.715
	15	6	74.39	0.604	77.11	0.756
Densenet201	5	10	82.93	0.536	69.88	0.768
	10	9	79.27	0.678	68.67	0.820
	15	6	81.71	0.637	66.27	0.936

Table 5
Probability of classification of each well-preserved mark selected from FLK N (1.8 Ma) by each DL model (Resnet50, Densenet201, and VGG19, ensemble analyses). The agency identified is highlighted in bold.

model	mark	Extensive mark				Cropped mark			
		crocodile	hyena	leopard	lion	crocodile	hyena	leopard	lion
Resnet50									
	FLKN+45 + 1	0.0006	**0.979**	0.002	0.017	0.001	**0.978**	0.004	0.012
	FLKN+410 + 3	0.018	**0.921**	0.047	0.0126	0.0007	**0.994**	0.004	0.001
	FLKN+49d + 3	0.049	0.338	0.039	**0.572**	0.201	**0.495**	0.067	0.235
	FLKN+811-13	0.005	**0.972**	0.006	0.015	0.086	**0.893**	0.008	0.011
Densenet201									
	FLKN+45 + 1	0.059	**0.827**	0.047	0.0655	0.016	**0.96**	0.004	0.018
	FLKN+410 + 3	0.005	**0.91**	0.007	0.077	0.084	**0.896**	0.003	0.015
	FLKN+49d + 3	**0.483**	0.314	0.046	0.155	**0.9**	0.038	0.021	0.037
	FLKN+811-13	0.0196	**0.906**	0.072	0.001	0.129	**0.849**	0.015	0.005
VGG19									
	FLKN+45 + 1	<0.001	0.245	**0.749**	0.005	<0.001	**0.979**	0.014	0.002
	FLKN+410 + 3	0.026	**0.964**	0.008	0.0001	0.239	**0.759**	0.0008	0.0002
	FLKN+49d + 3	0.0001	**0.969**	0.007	0.023	0.026	0.315	0.065	**0.597**
	FLKN+811-13	0.003	**0.825**	0.059	0.1119	0.023	**0.962**	0.0007	0.012
Ensemble									
	FLKN+45 + 1	0.019	**0.683**	0.2666	0.0295	0.006	**0.972**	0.007	0.013
	FLKN+410 + 3	0.016	**0.932**	0.021	0.03	0.108	**0.883**	0.003	0.005
	FLKN+49d + 3	0.177	**0.54**	0.031	0.25	**0.37**	0.283	0.049	0.29
	FLKN+811-13	0.009	**0.901**	0.046	0.042	0.079	**0.9**	0.008	0.01
Weighted ensemble									
	FLKN+45 + 1	0.02	**0.71**	0.253	0.029	0.006	**0.972**	0.007	0.013
	FLKN+410 + 3	0.016	**0.931**	0.021	0.03	0.103	**0.887**	0.003	0.005
	FLKN+49d + 3	0.178	**0.53**	0.031	0.259	**0.38**	0.286	0.05	0.283
	FLKN+811-13	0.009	**0.904**	0.045	0.041	0.08	**0.9**	0.008	0.01

of one of them (Table 5). The ensemble analysis classified the FLKN+49d + 3 tooth pit as "crocodile", but with extremely low probability (38%), thus making this classification unreliable. The other marks appear classified as "hyena", with high probabilities (ranging between 88% and 97% in the weighted ensemble analysis) (Table 5). The analysis of the cropped images shows that, in general, the most confident identifications seem to remain unaffected by the additional information contributed by the bone surface area, but the perimetrical areas contain information that may influence decisions made by the algorithms, especially in circumstances of identification bearing low probabilities. For this reason, it is advised to minimize the influence of areas outside the marks when using CV.

When classifying the three diagenetically modified marks (Fig. 23), it is clear that there is no pattern (Table 6). For Resnet50, one mark is "lion", another one is "leopard" and the third one is "crocodile". In the three cases, the probabilities are >85%. For Densenet201, the same marks are classified as "lion" (two pits) and "leopard" (on pit); two marks with probabilities >95%. For VGG19, the marks are classified as "hyena, leopard, lion" (one of each), only one exhibiting a probability >70%. This shows that, despite the high probabilities, the low convergence in classification by individual models renders the final classification unreliable. A weighted ensemble analysis classifies the three marks similarly as the individual Resnet50 model, but with only one mark showing a probability >70%. The lack of pattern, as well as the low probabilities in the ensemble analysis indicates that the classification must not be trusted. This clearly shows how when fossil marks show additional modifications (i.e, abrasion, chemical etching, mineral coating, bioerosion) that have not been experimentally modeled, the final classification results are not reliable.

When using the FSL models on the original uncropped images of the well-preserved, tooth marks Resnet50 identified three marks as made by hyena, but only two with probabilities higher than 70% (Table 7). Only when using the cropped marks did the model classify three marks as hyena with probabilities >70%. Again, tooth mark FLKN+49d + 3 was classified as lion uncropped) or crocodile (cropped); in both cases with lower probabilities. Densenet201 classified all the marks as made by hyenas with probabilities >90% (only FLKN+49d + 3 displayed a probability <70%). When using the cropped images of the same marks, again (as was the case for the DL models), three marks are classified as hyena (with high probabilities), whereas FLKN+49d + 3 is classified as crocodile. The ensemble analysis yielded a classification of all marks as hyena-made if using the uncropped images, and only three as hyena if using the cropped one. In both the DL and FSL models, the convergence in the classification of three of the marks make them reliable, and the variable discordance of the classification of FLKN+49d + 3 makes it unreliable.

When using the diagenetically-modified tooth marks (Table 8), Resnet50 classifies all the three marks as lion-made (in contrast with the DL models, which only classified one as lion). Densenet201 classified one mark as leopard and the other two as lion, similarly as it did with the DL models, but the ensemble model show discordances with the DL models. Again, The clear convergence in classification (and associated

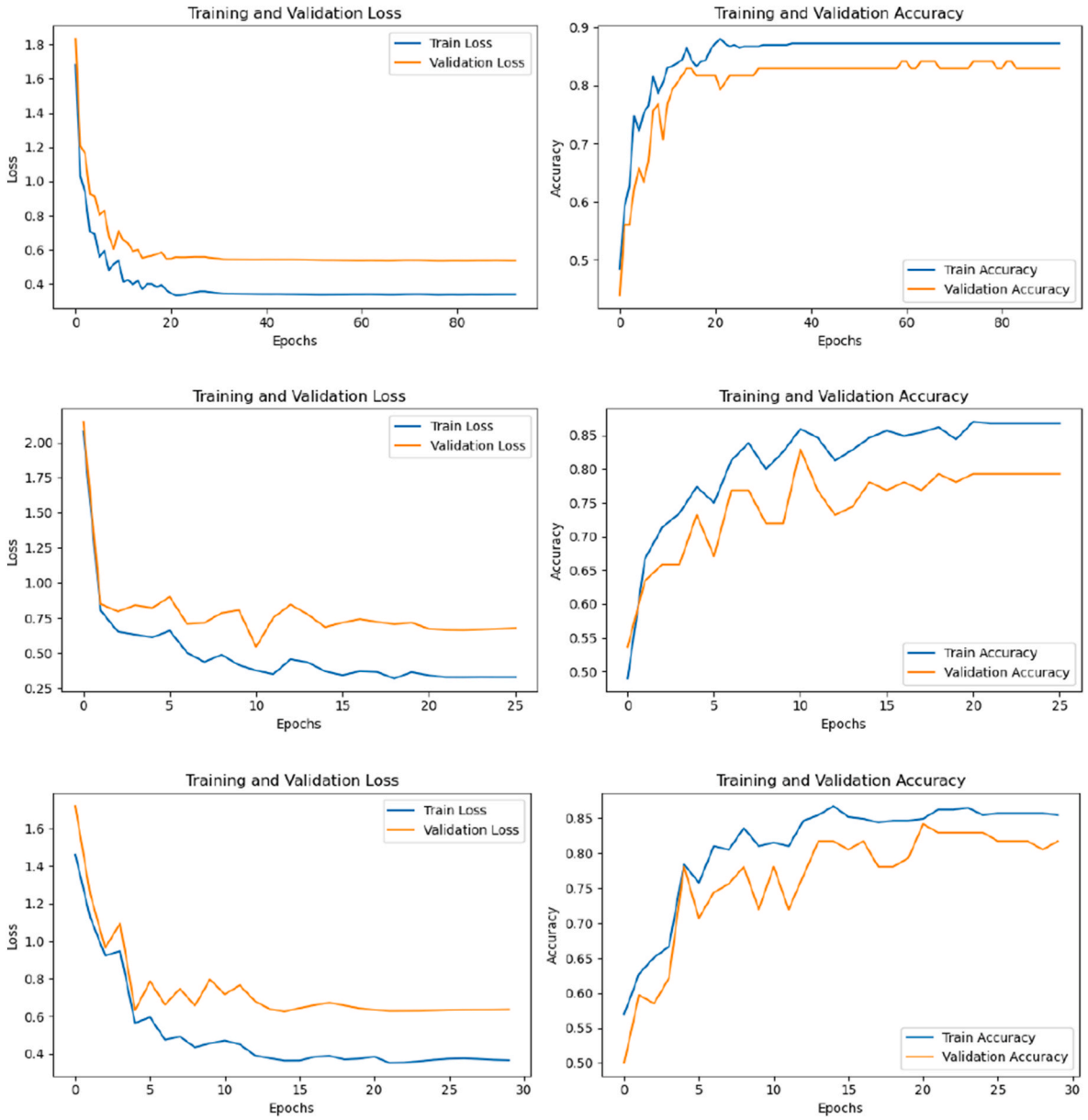

Fig. 20. Training graphs (accuracy and loss) of the FSL Densenet 201 model with a combination of 5-shot and 10-tasks (upper), 10-shot and 9-tasks (middle), and 15-shot and 6-tasks (lower).

high probabilities) of the well-preserved tooth marks when using DL and FSL models indicates a reliable classification, whereas the divergence in classification and probabilities of the three diagenetically-modified marks when using DL, FSL and ensemble methods suggests those attributions are unreliable.

4. Discussion

4.1. GMM analysis

Previous analyses of tooth pits using three-dimensional GMM yielded extremely high accuracy in carnivore agent identification (Yravedra et al., 2019; Courtenay et al. 2019, 2021). This was even the case when using functionally and morphologically very similar carnivores (e.g., different types of canids) (Yravedra et al., 2019). This was feasible even when the odds were initially against it. First, most studies suffered from low power; namely, sample sizes ranged from small to extremely small (Table 9). Such small samples were sometimes additionally reduced by removing variance-increasing outlying data (Courtenay et al., 2021) or by splitting datasets into training/testing sets, when machine learning approaches were used (Courtenay et al., 2019). Secondly, none of the 3D GMM studies used complete topographies of the marks, as suggested by Otárola-Castillo et al. (2018), which would have objectively represented the complete marks (Otárola-Castillo et al., 2018). Instead, only 17 semi-landmarks (defined by some as homologous, but most of which were obviously Type III semi-landmarks) were used (Arriaza et al., 2019; Aramendi et al., 2017). A total of 12 of those semi-landmarks represented the contour of the mark shoulder, and 5 were inserted inside the pit. In only one study, 30 semi-landmarks were used, 16 of them used for outline reconstruction, and 14 presumably for the inner pit (although only 9 are visible in the reconstructed mean landmark configuration) (Courtenay et al., 2021).

Even though the placement of these semi-landmarks was intended to represent homologous sections of the mark, the lack of metrically-based methods for the insertion of each semi-landmark made their location subjective (and, therefore, non-homologous). As a matter of fact,

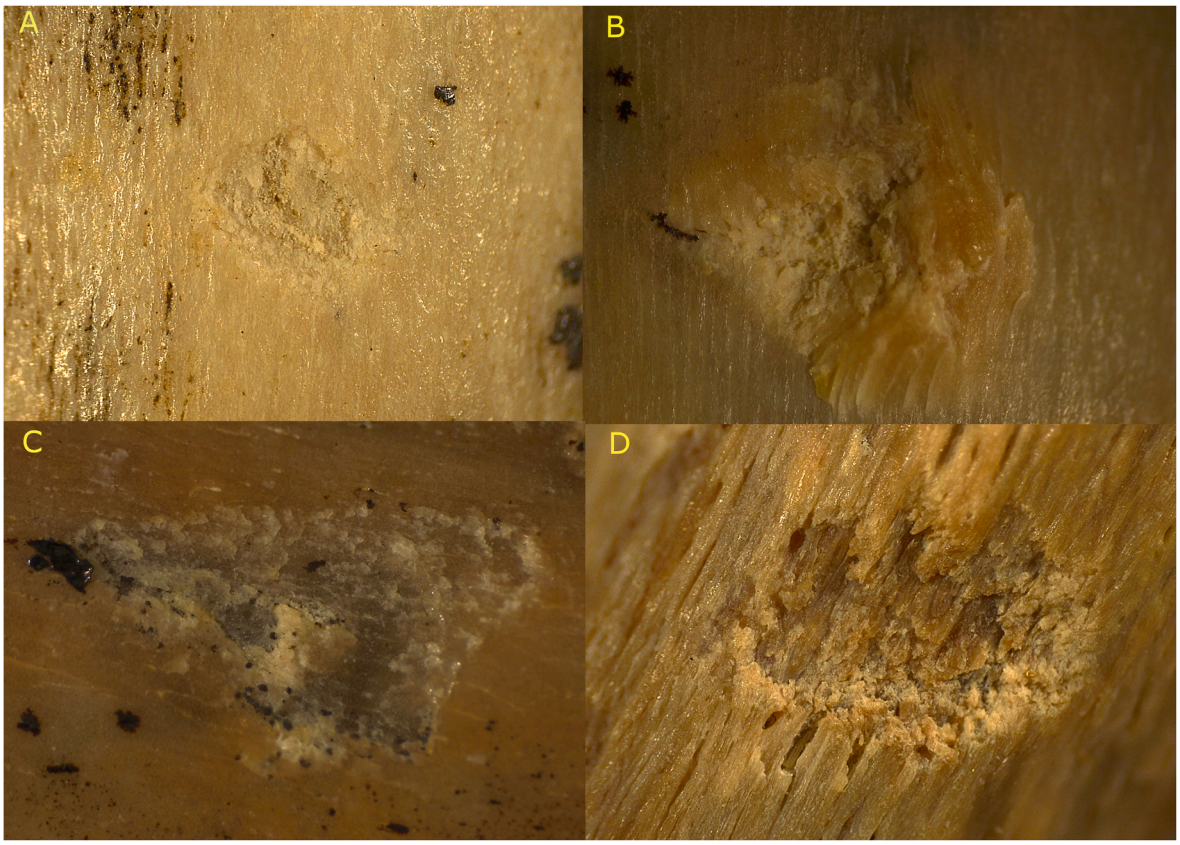

Fig. 21. Selected well-preserved tooth pits from the FLK N site, dating to 1.8 Ma. These are: FLKN+45 + 1 (A), FLKN+49d + 3 (B), FLKN+811 + 13 (C), and FLKN+410 + 3 (D). Notice the well-preserved cortical surface, being intact and only with a few manganese speckles (A–C) or rugose original bone texture (D).

Table 6
Probability of classification of the three modified marks selected from FLK N (1.8 Ma) by each DL model (Resnet50, Densenet201, and VGG19, weighted ensemble analysis), and by the best performing FSL model (Densenet201, 10-shot, 20-task). Agency identified is highlighted in bold.

model	mark	crocodile	hyena	leopard	lion
Resnet50					
	1272_1 × 30	0.027	0.021	0.001	**0.949**
	58_1 × 30	0.063	0.018	**0.891**	0.026
	8632_2 × 30	**0.853**	0.023	0.019	0.102
Densenet201					
	1272_1 × 30	0.011	0.012	0.019	**0.956**
	58_1 × 30	0.004	0.003	**0.97**	0.021
	8632_2 × 30	0.31	0.011	0.29	**0.386**
VGG19					
	1272_1 × 30	0.126	**0.866**	0.003	0.004
	58_1 × 30	0.13	0.084	**0.654**	0.131
	8632_2 × 30	0.084	0.26	0.076	**0.578**
Weighted ensemble					
	1272_1 × 30	0.055	0.3	0.008	**0.653**
	58_1 × 30	0.064	0.034	**0.842**	0.058
	8632_2 × 30	**0.429**	0.094	0.127	0.347

repositioning of semi-landmarks by different analysts on the same tooth pits yielded divergent locations (Courtenay et al., 2020). Thirdly, a somewhat biased selection (or reconstruction) of oval pits was carried out, so that the semi-landmark approach could be justified (especially for those landmarks referred to as Type II). This is an underrepresentation of the wide diversity of tooth pit morphologies in each agent, as documented in the present work (Figs. 11–14). Without the inclusion of all tooth pit types, a comprehensive analysis cannot be made to eventually sustain that reliable differentiation among carnivore taxa can be carried out. In the 3D GMM approach, circular pits could, in theory, not be used because the major axis that determines the insertion of the first two semi-landmarks cannot be established.

The fact that with so few tooth marks per carnivore agent, GMM analyses could yield 100% of accuracy [but see contradicting results in Arriaza et al. (2021); Aramendi et al. (2017)] can only be accounted for by two options: a) the 17 (or 30) semi-landmark approach is capturing some essential part of the within-sample (co)variance that the present bidimensional outline analysis is not seizing,[1] or b) the accuracy obtained is an artifact of method. For the former option, if the 17–30 semi-landmarks truly represent the mark morphology, the results should be additionally confirmed with 3D analyses that would include the complete mark (i.e., complete topography or grid). The logic behind this is that if a small set of semi-landmarks capture a small portion of the mark morphology, the complete morphology should be even more informative, and therefore, more discriminatory. The opposite would be indicative of artificial classification through methodological bias. As Gunz et al. (2005: 86) emphasize, "the more semi-landmarks, the better as far as representation of a geometric form is concerned". If this testing step is confirmed, it is our suspicion that the reason for success of the published 3D GMM studies must be the peculiar features captured by the five inner semi-landmarks representing partially the internal pit morphology. We base this assumption on three premises: one is that the present study, including a wider array of morphologies, shows that the shoulder-contour outlines are not resolutive; this should be even more so when most of the marks used in 3D GMM studies are in essence symmetrically oval. If 50 evenly-spaced semi-landmarks tuned to the tooth pit outline (this study) fail to detect agency, 12 semi-landmarks placed

[1] There is data confirming that "landmark displacements across the z-axis" display the greatest differences, and account for most inter-agency shape variance (Courtenay et al., 2021).

Table 7
Probability of classification of each well-preserved mark selected from FLK N (1.8 Ma) by the best performing FSL models (Resnet50, 10 shot-9 task; Densenet201, 5-shot, 10-task). Agency identified is highlighted in bold.

model	mark	Extensive mark				Cropped mark			
		crocodile	hyena	leopard	lion	crocodile	hyena	leopard	lion
Resnet50									
	FLKN+45 + 1	0.0049	**0.854**	0.0291	0.1116	0.0009	**0.9826**	0.0071	0.0094
	FLKN+410 + 3	0.0006	**0.9985**	0.0002	0.0008	0.0027	**0.9972**	0.0001	0
	FLKN+49d + 3	0.233	0.1932	0.1225	**0.4513**	**0.7636**	0.0682	0.0824	0.0859
	FLKN+811-13	0.231	**0.559**	0.1531	0.0569	0.1664	**0.7512**	0.0116	0.0708
Densenet201									
	FLKN+45 + 1	0.0026	**0.9745**	0.0072	0.0157	0.0014	**0.981**	0.0066	0.0109
	FLKN+410 + 3	0.022	**0.9932**	0.0002	0.0044	0.003	**0.9939**	0.0005	0.0027
	FLKN+49d + 3	0.235	**0.6335**	0.0122	0.1193	**0.914**	0.0169	0.0059	0.0632
	FLKN+811-13	0.0608	**0.9336**	0.0048	0.0007	0.2646	**0.6937**	0.0062	0.0355
Ensemble									
	FLKN+45 + 1	0.003	**0.914**	0.018	0.063	0.001	**0.981**	0.006	0.01
	FLKN+410 + 3	0.001	**0.995**	0.0001	0.002	0.002	**0.995**	0.0002	0.001
	FLKN+49d + 3	0.234	**0.413**	0.067	0.285	**0.838**	0.042	0.044	0.074
	FLKN+811-13	0.31	**0.581**	0.078	0.0288	0.43	**0.507**	0.008	0.053

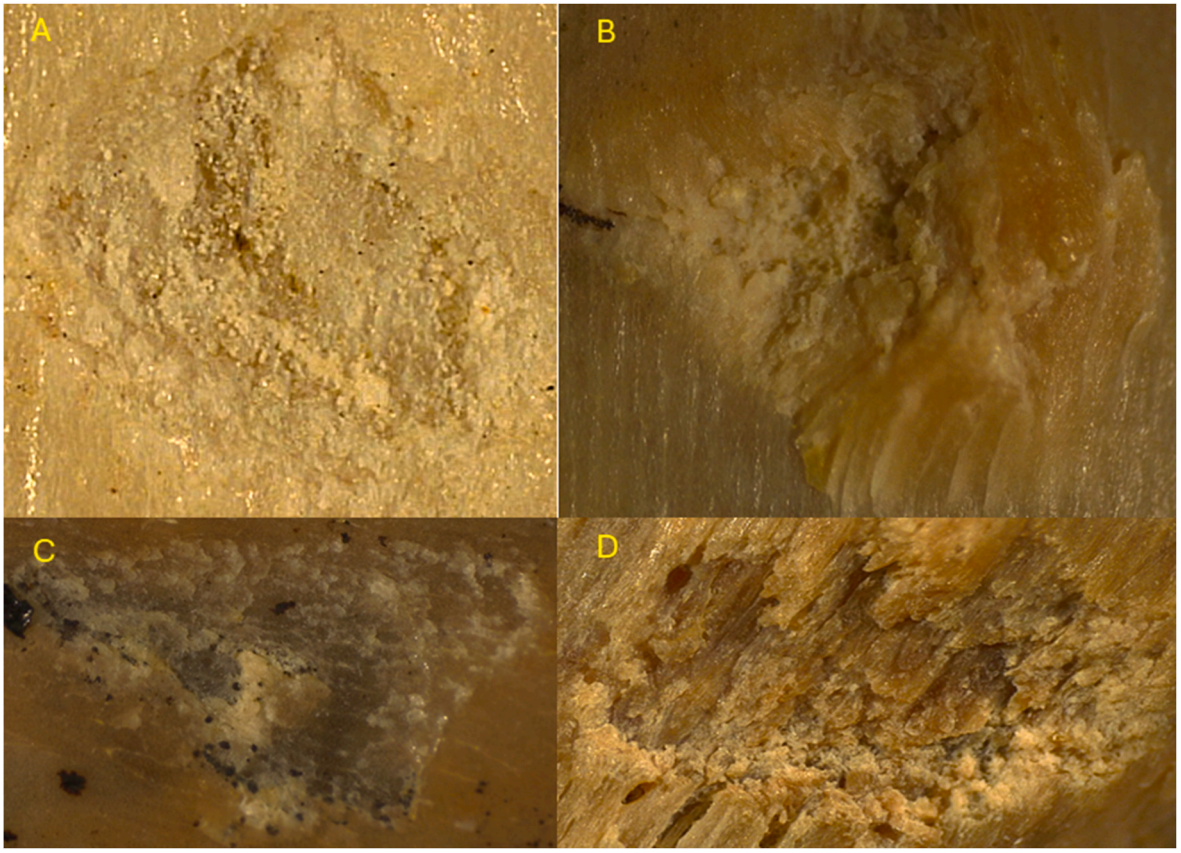

Fig. 22. Selected well-preserved tooth pits from the FLK N site, dating to 1.8 Ma, having their perimetrical surface cropped. These are: FLKN+45 + 1 (A), FLKN+49d + 3 (B), FLKN+811 + 13 (C), and FLKN+410 + 3 (D). The tooth pit shown in B is the most insecurely classified, probably because of its shallow modification and its angular geometric form attached to a portion of uplifted uppermost cortical bone.

by hand should be even more equifinal when targeting shape. The second reason is that mean shapes for tooth pits spanning different carnivore agents documented via 3D GMM are systematically displayed showing almost identical tear-shaped or oval symmetrical outline, regardless of carnivore (Courtenay et al., 2021; Yravedra et al., 2019) (Fig. 24). Such a tear shape is unreported in the present study, despite the wide array of shapes documented for some of the same types of carnivores. The third reason is that such an oval-dominated sample could also result from forcing that shape on the tooth pit. In several examples on some published 3D GMM studies of how the semi-landmarks were inserted on the tooth pit rim, the tridimensional rendering of the bone surface makes delimiting the actual tooth pit outline difficult, and alternative placements could equally have been selected, as proved by the non-overlying location of semi-landmarks placed by different analysts on the same marks (Courtenay et al., 2020). This also affects the inner semi-landmarks. Although by definition it is said that these inner semi-landmarks represent inflection points on the breadth and longitudinal axes, these are inserted at unequal distances (the selection of location for such inflective points is not described) and even outside the main axes (e.g. Fig. 3 in Courtenay et al.,

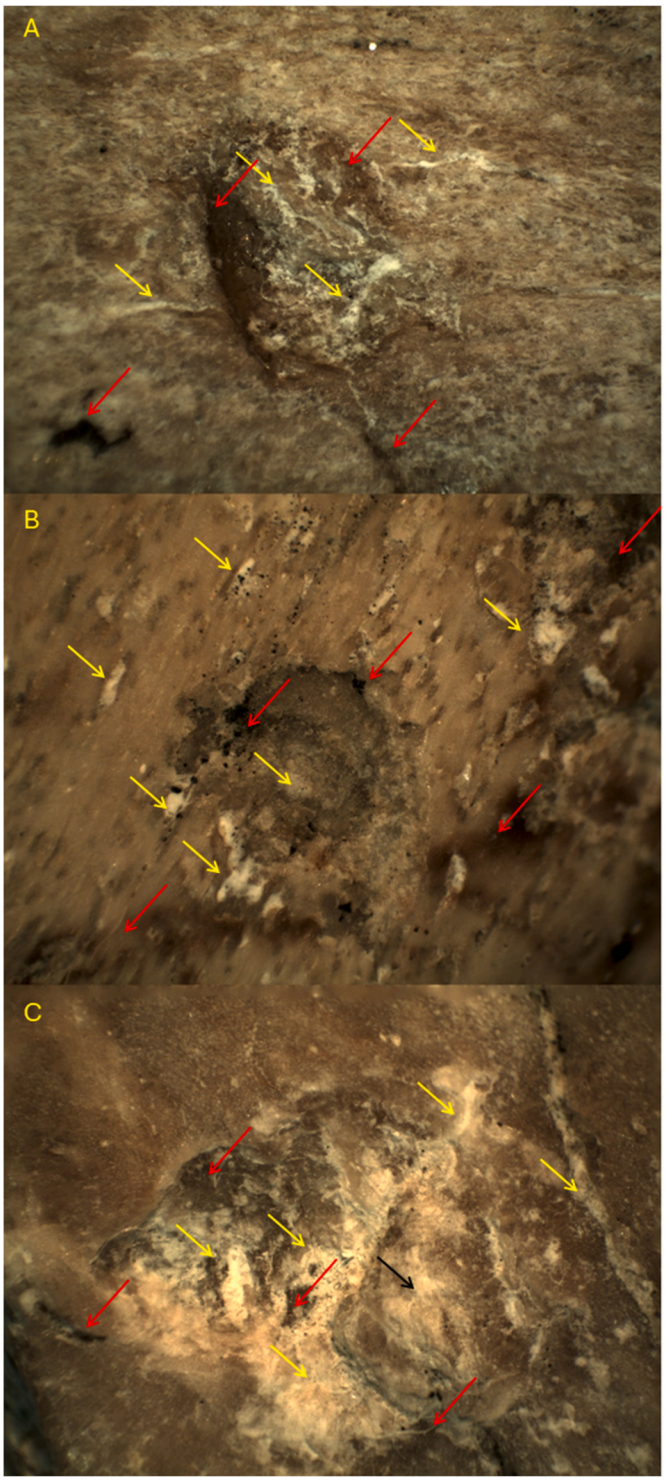

Fig. 23. Moderately-modified tooth pits from the FLK N site, impacted by biochemical processes during diagenesis. A, 8632_2 × 30; B, 1272_1 × 30; C, 58_1 × 30. Notice a selection of diagenetic modifications in the form of biochemical marks created by plant roots in light color (yellow arrows) and dark color, including manganese spotting (red arrows).

2019). How this affects the published accuracy rates remains to be tested. In essence, this approach involves a highly subjective interpretation of the mark contour and interior. As a matter of fact, the actual contour of marks used in some of these examples is not as oval as reconstructed (e.g. Fig. 3 in Courtenay et al., 2019).

It is very relevant that in several of these studies (e.g. Aramendi et al., 2017; Courtenay et al., 2021), it is argued that most of the variation is allometrically determined, clearly suggesting that it is form and not shape that drives more of the differences among carnivore tooth pits. Therefore, tooth pit size is a relevant discriminatory feature; thus influencing potential biases when single-agent subsamples do not include all size ranges (see below).

The alternative option to the resulting high discriminatory resolution in 3D GMM analyses is that the high accuracy on such small samples could reflect a methodological artifact. A known biasing factor in GMM is related to bending energy resulting from the use of equidistant semi-landmarks. In the published morphometric analyses of tooth marks, homology is justified by the placement of equidistant semi-landmarks, especially around the tooth mark contour (Yravedra et al., 2019; Courtenay et al. 2019, 2021). Equidistant semi-landmarks placed along a geometric surface tend to produce a false deformation grid, creating local shape differences that do not exist (Gunz et al., 2005). When using such an approach, it is essential to apply measures to minimize bending energy preventing equivocal deformation of individual marks and classes. Without this, "the true spacing of geometric homologues along biologically homologous curves or surfaces" is unknown and its uncritical application is certainly morphologically biasing (Gunz et al., 2005, p. 80). Procrustes methods have been presented as a way to keep bending energy invariant, but they are highly dependent on the assumption that semi-landmarks are truly homologous. This cannot be when different analysts place semi-landmarks on different locations.

In addition, it is important to stress that traditional geometric morphometric analyses that have been made following the Procrustes transformation are no longer recreating the original shapes, and therefore, are not properly estimating variation in the natural morphological space. This is because there are nuisance parameters (rotation,

Table 8
Probability of classification of the three modified marks selected from FLK N (1.8 Ma) by each FSL model (Resnet50, 10 shot-9 task; Densenet201, 5-shot, 10-task) and using an ensemble approach. Agency identified is highlighted in bold.

model	mark	crocodile	hyena	leopard	lion
Resnet50					
	1272_1 × 30	0.0012	0.0011	0	**0.997**
	58_1 × 30	0.1637	0.208	0.0325	**0.783**
	8632_2 × 30	0.0024	0	0	**0.997**
Densenet201					
	1272_1 × 30	0.031	0.0609	0.0113	**0.8927**
	58_1 × 30	0.0027	0.0066	**0.9544**	0.0363
	8632_2 × 30	0.0299	0.0175	0.0511	**0.9015**
Ensemble					
	1272_1 × 30	0.0832	0.0137	**0.4934**	0.409
	58_1 × 30	0.018	0.031	0.005	**0.9452**
	8632_2 × 30	0.0161	0.008	0.025	**0.949**

Table 9
Sample size per carnivore (unspecified as c1-c8), and accuracy in classification in each published 3D GMM study of tooth pits.

c1	c2	c3	c4	c5	c6	c7	c8	accuracy	ref.
21	24							–	Arriaza et al., 2019a
21	24	34	24					100%	Courtenay et al. (2019)
69	86	80	89	53	77	84	82	>90%[a]	Courtenay et al. (2021)
21	24	41						60.30%	Arriaza et al., 2019b
29	24	30						100%	Yravedra et al. (2019)
20	20	20	20	9				50%	Aramendi et al. (2017)

[a] Between 3 and 10 pits were subsequently removed for each carnivore taxon, reducing the sample.

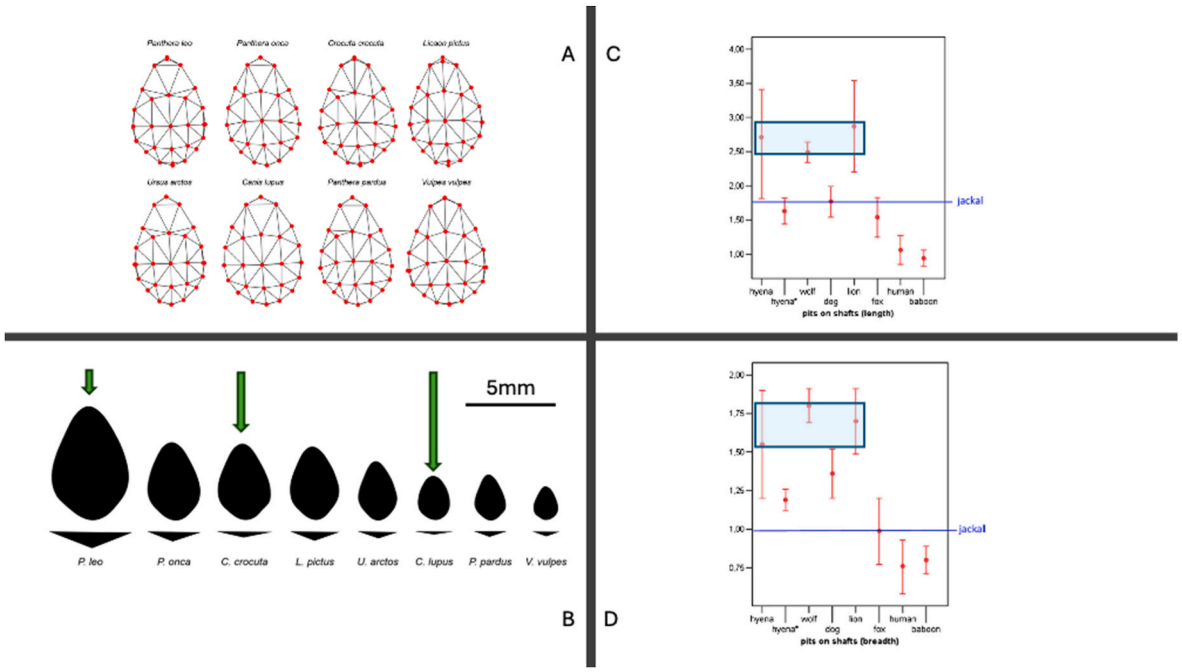

Fig. 24. A, Mean shape of the tooth pits as reconstructed with semi-landmarks for eight different carnivore taxa in Courtenay et al., (2021) work. B, Mean-shape resulting from the use of GMM methods on the semi-landmark framework showing form differences (Courtenay et al., 2021). C and D, mean and 95% confidence intervals of tooth pit sizes according to length (C) and breadth (D) on long bone shafts.

translation, reflection) that impact the way variation is captured and how mean shapes are estimated, since their effect cannot be removed from the individual coordinate system of each shape (Lele and McCulloch, 2002). Likewise, the Procrustes method has some inherent assumptions that are systematically ignored, such as: a) homogeneity in the variance of all landmarks, and b) independence of variation among landmarks (Richtsmeier et al., 2005). Both assumptions are commonly violated when analyzing shape in biological entities, since morphological integration and modularity require landmark co-variation and inter-dependence. For all these reasons, biased sampling of shape and Procrustes transformation could generate remodeled shape assemblages that can display artificial classification accuracy.

The artificial methodological discrimination of tooth marks in published GMM tooth mark works using Procrustes-transformed geometric morphometrics would make sense for three reasons: a) in a 3D model, the mark shoulder (i.e., contour) should only be a discriminant factor if it was different among agents, but b) our study shows that the outline diversity of the tooth pit contours cannot be effectively be used to identify agency, and c) all outline semi-landmarks occupy a horizontal surface with no relevant 3D information. The inner semi-landmarks maintain a spatial interrelation among themselves and in relation to the contour semi-landmarks. The actual location of such points is not trivial. It establishes a coordinate network impacted by the single topography of each mark and allometric factors. These semi-landmarks act as a labeled barcode that is individual for each mark, and subsequently, for each carnivore class. If this relationship is impacted by allometry, then the method may be actually capturing more variance due to size than to shape. This seems to be partly the case, since our colleagues recognize that "considering the relatively small sample size … (and the) important margins of error product of landmark quality … all samples are described by notable allometric patterns, indicating tooth pit size to be an important conditioning factor in morphological variation" (Courtenay et al., 2021). By "morphological variation", it must be understood that there are different spatial networks of an uncertain number of semi-landmarks, since the mean tooth pit shape of all carnivoran taxa in these 3D GMM studies displays virtually the same contour (Fig. 24a and b). This underscores that if the sample selected is not random, and if not all single-agent subsamples are unbiased in form, statistically-identified differences may be spurious.

In this regard, a remarkable mismatch is documented between variations in tooth pit size across several taxa (and their respective sizes) (Courtenay et al., 2021), and tooth pit sizes documented for some of these taxa by a much more extensive sample that included conspicuous and inconspicuous (i.e., small and microscopically-identified) tooth marks (Andrés et al., 2012) (Fig. 24). The largest multi-taxa 3D GMM study to date includes an average of 77 pits per agent (prior to removal of variance-impacting marks) (Courtenay et al., 2021). For example, regarding large carnivores, hyenas are represented by 86 tooth pits, lions by 82 pits and wolves by 80 pits. In Andrés et al. (2012) study, the sample sizes for the same taxa and only documented on long bone shafts are substantially bigger: hyenas (n = 779), lions (n = 206) and wolves (n = 365). The higher variability in the latter study may explain the differences in tooth pit sizes when comparing these (and other) taxa in both types of studies (Fig. 24). For example, in the 3D GMM published work, the mean size for tooth pits made by lions is > 5 mm (>3 mm for breadth) (Courtenay et al., 2021) (twice as much as documented in Andres et al.'s much bigger sample) (Andrés et al., 2012). For hyenas, it is > 4 mm (substantially bigger than in Andres et al.'s sample). Proportionally, lion tooth pits in the 3D GMM study are twice as big as those made by hyenas, and wolves are about six times smaller (Courtenay et al., 2021). In contrast, in the much more comprehensive tooth mark sample by Andrés et al. (2012), the mean size of tooth pits from hyenas, lions and wolves is very similar (with <0.5 mm of a difference among them), and their corresponding confidence intervals show intense overlap and no statistical difference (Andrés et al., 2012) (Fig. 24c). This is suggestive of a biased selection (or representation) of tooth pits in different agents in the published 3D GMM studies, with a clear selection of bigger pits for lions and differential selection of size-determined pits for several other agents. If success in classification is allometrically determined by size, then the samples selected do not represent the complete range of marks generated by each agent and, thus, the reliability of classification of agency can be questioned. The consequences of this is that if the smaller 3D GMM sample is capturing a smaller spectrum of single-agent variance, the statistical separation of agency

reported might not be real, especially with such a contrasting difference in the tooth mark sizes selected for each agent.

A hint that some methodological biases are conditioning the artificial accuracy of some of these published tooth mark morphometric works is that using exactly the same marks from our own tooth mark dataset, Fourier contours (including the complete outline) are less accurate than Procrustes methods using a subsample of contour semi-landmarks on a distorted mark morphospace (see Results and Supplementary Information), We do not know if any of the objections raised here actually demonstrates that the high-accuracy yielded by 3D GMM studies is spurious. The method might actually be extremely efficient at determining agency. It is our view, however, that for the method to be validated two conditions should be complied first: each carnivoran taxon must be represented by a random and much larger sample involving the whole size-range of tooth pits to avoid allometric biases, and such sample must be tested against the use of the complete tridimensional grid of each mark -not a selection of 17 non-homologous semi-landmarks- to remove suspicions about methodological biases. Until that is done, we believe that data from these studies should be used with extreme caution. The semi-landmark method may have artificially generated a misrepresentation of the actual tooth marks, with artificially-derived differential variance that results in over-confident classification. This is especially the case, since when using the same methods on bidimensional images of tooth pits, the results are non-resolutive (See also Supplementary Information).

4.2. CV analysis

The CV study reported here, in contrast, is not subjected to allometry. Tooth pit size, therefore, is not an impacting factor. CV focuses on mark morphology and microscopic features. The DL models that we developed reached an accuracy in the classification of untrained marks ranging from 73% to 81% (Table 4). The FSL methods, usually created to deal with limited datasets, were very competitive in their performance and reached up to 79.52% of correct classification of the testing set. The efficiency in the classification was almost identical to the DL models. Resnet50 was the best model for DL, and was also the best performing model for FSL (Table 5). The Densenet 201 model was underperforming in comparison with it, and also with its DL counterpart. This is one of the first attempts to compare the performance of DL and FSL techniques, showing that although DL models are generally good at generalization, FSL models can be equally effective and very competitive. These results combined indicate that CV methods are the best option to identify taphonomic agency in carnivoran bone surface modification when using bidimensional information.

This recognition should not overshadow the caution of the application of CV methods to paleontological assemblages. The advantage of CV over other methods (i.e., CV seems to be efficient at detecting microscopic modifications) is also its biggest drawback. Bunge's caution in how analogy is built and used applies here (Bunge, 1981). Confidence in agency identification in the fossil record can only be held if the substantial and structural parts in the use of analogy are matched between both ends of the interpretive process; otherwise said, between explanandum and explanans. In this particular case, this refers to the interpretation of fossil marks that have the same preservational properties as the referential marks used for modeling. Marks appearing on fresh (or unmodified) bone, unaltered by other mechanical or chemical processes would be adequate to interpret through the reference libraries used here. Marks that have undergone additional modifications can certainly be classified with high probabilities using these models, but the underlying confidence in the classification varies in proportion to how much those marks have been impacted by additional biotic and abiotic processes during biostratinomy and diagenesis (Domínguez-Rodrigo et al., 2024). We infer this is not the case of GMM methods, which should remain unaffected by these processes, since overall shape is the target; however, this also remains to be tested.

The present study has shown in this regard that the four well-preserved tooth pits are classified in a patterned manner by individual models (i.e., most show similar classifications). Otherwise said, there is more convergence than divergence. Both when using individual models or all models together in ensemble analysis, there is a consistent high probability of classification in three out of the four marks. The combination of both facts (convergence and high probability) grants reliability in the interpretation of three of the tooth pits. In contrast, when using the three moderately-modified tooth pits, the individual models lack convergence and the ensemble model yields predominantly low probabilities. Both elements of consideration show signs of unreliable classification. According to the CV methods, up to four different agents are represented in such a small sample size of fossil marks (7 tooth pits); two of these agents have not been previously documented or inferred using traditional multivariate taphonomic analyses on the complete assemblages, involving thousands of bones (Domínguez-Rodrigo et al., 2007). This alone should also be taken as a cautionary note.

This underscores that the limitation of the current CV methods lies in their lack of modeling of the palimpsestic processes that impact bone surface modifications on fossil bones. Unless future experimental analyses include these factors or only semi-pristine fossil marks are used with current models, the heuristics of the resulting interpretations can be questioned.

One additional potential limitation of these models is their ability to discern agency when dealing with extinct carnivore taxa. When providing unknown tooth marks to the models, these will always fit them within the carnivores that they were trained with, with a potential for false positives. To control for this bias, it would be advisable to incorporate artificially-made tooth marks from extinct carnivores to the training of the models. This can be done in multiple ways. One would be generating tooth marks on molding material which would enable to capture the morphology and size of every single type of tooth. This was done successfully with the extinct saber-tooth felid *Xenosmilus hodsonae* (Domínguez-Rodrigo et al., 2022). This would also be enough to capture outline shapes for GMM analyses. Another, more sophisticated way, would be to make durable casts of fossil carnivore teeth and create tooth marks on real bone, by applying controlled force while using such proxy with the aid of a universal testing machine (UTM) or tensile testing machine (TTM). Both methods will leave 3D replicas or real tooth marks on bone which can also be analyzed using 3D CV and GMM methods, which remain under-explored in the analysis of tooth marks and taphonomy in general.

The present study also shows that interpretations of fossil marks derived from single models are insufficient to warrant proper heuristics and confidence, since these models may be substantially variable. An ensemble analysis is always advised (with as many models exhibiting high accuracy as possible).

This attempt to compare GMM (via EFT) and CV (via DL CNN) is not unique. Almost simultaneously in time, Bonhomme et al. (2023) have made a similar study on wild and domestic types of seeds and fruit stones, with CV systematically outperforming GMM in all tests. The convergence of the methodological results, using similar methods and different datasets, reinforces the better performance of CNNs over GMM in shape analysis, and its recommendation as the baseline method.

5. Conclusions

Bidimensional information of tooth marks (and, presumably, of other types of bone surface modifications [BSM]) presents some limitations, as exposed in the present study. The most relevant conclusion drawn therefrom is that for any interpretations of BSM to contain what epistemologists refer to as heuristics, analysts must be extremely careful in not breaking the substantial-structural-environmental forms of analogy (Bunge, 1981). Substantial identity in the content of two different phenomena or systems (here, could be using the same types of BSM) is the first step for correct comparison and interpretation (material

analogy). Here, the attributes of our referent and the object or process compared to it must be the same. Then, structural identity between both items or processes that are being compared is also an inferentially-required step. The structural resemblance is essential for the inference to be correct (formal analogy). In the present case, that means that if our referent is composed of single-agency (i.e., a tooth effector imprinting a tooth pit), the object/process to be compared (i.e., tooth pitting and the resulting BSM) must not include additional modifications (modifiers). These two forms of analogy are essential for inferentially-correct interpretations of past BSM. An additional form, which evaluates the impact of context, is also relevant. This is referred to as the environmental or contextual component of analogy; a more generic part of functional analogy. This analyzes how behaviors, prompted by their adaptation to tasks, create variability in the material outcome of the processes or objects to be compared. This can be summarized in all interpretations stemming from the application of reference frameworks must be as tightly linked to the properties of the referents as possible. If our experimental datasets are created with only tooth marks imprinted on fresh bone, our inferences demand that the prehistoric marks that we seek to interpret must have the same or similar agency-related and preservation properties. The further we move from this, the less reliable our interpretations become.

This is the biggest limitation of CV methods to the fossil record. Most BSM are subjected to dynamic processes of transformation across months, years, centuries or millennia. It is also true that the most impacting processes in morphing the original properties of BSM occur at the earliest stages of their taphonomic history: the biostratinomic and very early diagenetic phase, until the bone reaches stabilization with its chemical and lithological context. For many marks that combine the properties of the original actor-effector and those of subsequent processes (i.e., abrasion, exfoliation, weathering, bioerosion, multiple-agent overlap on the same marks), no proper referent exists that can guarantee in any objective way their attribution to agency. Having said that, it must also be remarked that in contexts where assemblages have fairly good preservation and BSM show virtually intact preservation (like some of the 1.8 Ma marks used in the present study), confidence in their interpretation can be high when the CV models applied create convergence and high probability in agent attribution (Cobo-Sánchez et al., 2022).

Likewise, small datasets of experimental tooth marks analyzed using GMM (Table 9) cannot be claimed to have 100% accuracy (or similar) in discriminating agency if they have not randomly sampled all the morphological range of tooth marks created by each agent (which they have not to date), or all the tooth mark sizes per agent, given the strong dependence of these methods on allometry. Therefore, the documented accuracy of these methods to date may be methodologically biased and should be used with caution, especially when applied to prehistoric or fossil marks of unknown origin. Reassessment of their discrimination capabilities need to be carried out, further supported by full 3D inter-agency comparison.

We have seen here that such models exist for bidimensional information. DL and FSL models have achieved accuracy of 81% and 79.52% respectively in classifying testing sets of 4-agent experimental tooth pits. We have also seen that most of the GMM potential probably lies in the 3D realm. The Fourier analysis applied to the four carnivore agents documenting shape variability of the tooth pit outlines created by each agent shows that statistical differences are detected (through MANOVA), but limited discriminant power is reached (<40%). For this reason, the next step is using complete 3D topographical information of tooth marks (instead of a limited number of semi-landmarks) both in more complex GMM studies and in CV analyses. It is our initial hypothesis that many of the problems in the interpretation of fossils BSM outlined here will become less so when the information becomes completely tridimensional.

This study intended to curb enthusiasm for the non-critical use of these new methods. Despite this, and the detailed list of cautions that must be taken when using these new approaches to BSM, taphonomists must welcome them because they provide for the first time an objective way of classifying BSM to taxon-specific agency, with indicators of confidence and reliability that did not exist before. The options that these new approaches open for the interpretation of the past are enormous. More research must build upon these new methods to make them even better, and to increase the reliability of our interpretations of the prehistoric past.

CRediT authorship contribution statement

Manuel Domínguez-Rodrigo: Writing – original draft, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Marina Vegara-Riquelme:** Writing – review & editing, Investigation, Data curation. **Juan Palomeque-González:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation. **Blanca Jiménez-García:** Writing – review & editing, Data curation. **Gabriel Cifuentes-Alcobendas:** Writing – review & editing, Investigation, Data curation. **Marcos Pizarro-Monzo:** Data curation. **Elia Organista:** Writing – review & editing, Supervision. **Enrique Baquedano:** Writing – review & editing, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Tanzanian Commission for Science and Technology (COSTECH), the Department of Antiquities, the Ngorongoro Conservation Area Authority (NCAA), and the Ministry of Natural Resources and Tourism for permission to conduct research at Olduvai Gorge. We also thank funding provided by the Spanish Ministry of Science and Innovation (grants: PID2020-115452GB-C21, PID2023-146260NB-C2), and the Spanish Ministry of Culture (program: Archaeological Research Abroad) for supporting our research. M. V. R. was funded by the Spanish Ministry of Universities with an FPU predoctoral grant (FPU18/05632). The authors are thankful to Edgard Camarós for having allowed one of us (M. V. R.) access to the Altamira Zoo crocodile collection. We thank the producers of the R "Momocs" library, and more specifically V. Bonhomme and J. Claude. We are indebted to the constructive comments made by Arati Deo and two anonymous reviewers.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.qsa.2025.100268.

Data availability

Data will be made available on request.

References

Abellán, N., Baquedano, E., Domínguez-Rodrigo, M., 2022. High-accuracy in the classification of butchery cut marks and crocodile tooth marks using machine learning methods and computer vision algorithms. Geobios - Mem. Spec. 72–73, 12–21.

Abellán, N., Jiménez-García, B., Aznarte, J., Baquedano, E., Domínguez-Rodrigo, M., 2021. Deep learning classification of tooth scores made by different carnivores: achieving high accuracy when comparing African carnivore taxa and testing the hominin shift in the balance of power. Archaeol. Anthropol. Sci. 13, 31.

Andrés, M., Gidna, A.O., Yravedra, J., Domínguez-Rodrigo, M., 2012. A study of dimensional differences of tooth marks (pits and scores) on bones modified by small and large carnivores. Archaeol. Anthropol. Sci. 4, 209–219.

Aramendi, J., Maté-González, M.A., Yravedra, J., Ortega, M.C., Arriaza, M.C., González-Aguilera, D., Baquedano, E., Domínguez-Rodrigo, M., 2017. Discerning carnivore

agency through the three-dimensional study of tooth pits: revisiting crocodile feeding behaviour at FLK- Zinj and FLK NN3 (Olduvai Gorge, Tanzania). Palaeogeogr. Palaeoclimatol. Palaeoecol. 488, 93–102.
Arriaza, M.C., Aramendi, J., Maté-González, M.Á., Yravedra, J., Baquedano, E., González-Aguilera, D., Domínguez-Rodrigo, M., 2019. Geometric-morphometric analysis of tooth pits and the identification of felid and hyenid agency in bone modification. Quat. Int. 517, 79–87.
Arriaza, M.C., Aramendi, J., Maté-González, M., Yravedra, J., Stratford, D., 2021. Characterising leopard as taphonomic agent through the use of micro-photogrammetric reconstruction of tooth marks and pit to score ratio. Hist. Biol. 33, 176–185.
Binford, L.R., 2014. Bones: Ancient Men and Modern Myths. Academic Press.
Bonhomme, V., Claude, J., 2020. Morphometrics Using R [R Package Momocs Version 1.3.2].
Brugal, J.-P., Fourvel, J.-B., 2024. Puncture game: let's play with the canines of carnivores. Quatern. Sci. Advan. 13, 100129.
Bonhomme, V., Bouvy, L., Claude, J., Dham, C., Gros-Balthazard, M., Ivorra, S., Heanty, A., Pagnoux, C., Pastor, T., Terral, J.F., Evin, A., 2023. Deep learning versus geometric morphometrics for archaeobotanical domestication study and subspecies identification. bioRxiv preprint. https://doi.org/10.1101/2023.09.15.557939.
Bunge, M., 1981. Analogy between systems. Int. J. Gen. Syst. 7, 221–223.
Cifuentes-Alcobendas, G., 2025. New Methodologies Applied to the Study of Bone Surface Modifications of Olduvai Gorge's Pleistocene Fossil Record: a Union between Taphonomy and Experimental Archaeology. Alcala University, Madrid, Spain. Ph. Dissertation.
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, 18933.
Claude, J., 2008. Morphometrics with R. Springer Science & Business Media.
Cobo-Sánchez, L., Pizarro-Monzo, M., Cifuentes-Alcobendas, G., Jiménez García, B., Abellán Beltrán, N., Courtenay, L.A., Mabulla, A., Baquedano, E., Domínguez-Rodrigo, M., 2022. Computer vision supports primary access to meat by early Homo 1.84 million years ago. PeerJ 10, e14148.
Courtenay, L.A., Herranz-Rodrigo, D., González-Aguilera, D., Yravedra, J., 2021. Developments in data science solutions for carnivore tooth pit classification. Sci. Rep. 11, 10209.
Courtenay, L.A., Herranz-Rodrigo, D., Huguet, R., Maté-González, M.Á., González-Aguilera, D., Yravedra, J., 2020. Obtaining new resolutions in carnivore tooth pit morphological analyses: a methodological update for digital taphonomy. PLoS One 15, e0240328.
Courtenay, L.A., Yravedra, J., Huguet, R., Aramendi, J., Maté-González, M.Á., González-Aguilera, D., Arriaza, M.C., 2019. Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks. Palaeogeogr. Palaeoclimatol. Palaeoecol. 522, 28–39.
Domínguez-Rodrigo, M., Barba, R., Egeland, C.P., 2007. Deconstructing Olduvai: A Taphonomic Study of the Bed I Sites. Springer Science & Business Media.
Domínguez-Rodrigo, M., Cifuentes-Alcobendas, G., Jiménez-García, B., Abellán, N., Pizarro-Monzo, M., Organista, E., Baquedano, E., 2021a. Author Correction: artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. Sci. Rep. 11, 3708.
Domínguez-Rodrigo, M., Egeland, C.P., Cobo, L., Baquedano, E., 2022. Sabertooth carcass consumption behavior and the dynamics of Pleistocene large carnivoran guilds. Scientici Rep. 12, 6045.
Domínguez-Rodrigo, M., Mabulla, A.Z.P., Bunn, H.T., Diez-Martin, F., Baquedano, E., Barboni, D., Barba, R., Domínguez-Solera, S., Sánchez, P., Ashley, G.M., Yravedra, J., 2010. Disentangling hominin and carnivore activities near a spring at FLK North (Olduvai Gorge, Tanzania). Quat. Res. 74, 363–375.
Domínguez-Rodrigo, M., Pizarro-Monzo, M., Cifuentes-Alcobendas, G., Vegara-Riquelme, M., Jiménez-García, B., Baquedano, E., 2024. Computer vision enables taxon-specific identification of African carnivore tooth marks on bone. Sci. Rep. 14, 6881.
Domínguez-Rodrigo, M., Saladié, P., Cáceres, I., Huguet, R., Yravedra, J., Rodríguez-Hidalgo, A., Martín, P., Pineda, A., Marín, J., Gené, C., Aramendi, J., Cobo-Sánchez, L., 2017. Use and abuse of cut mark analyses: the Rorschach effect. J. Archaeol. Sci. 86, 14–23.
Domínguez-Rodrigo, M., Courtenay, L., Cobo, L., Baquedano, E., Mabulla, A., 2021b. A case of scavenging 1.84 million years ago from Olduvai Gorge (Tanzania). Ann. N. Y. Acad. Sci.
Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, pp. 1126–1135.
Flood, Sung, Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M., 2017. Learning to compare: relation Network for few-shot learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1199–1208.
Gaudzinski-Windheuser, S., Kindler, L., Rabinovich, R., Goren-Inbar, N., 2010. Testing heterogeneity in faunal assemblages from archaeological sites. Tumbling and trampling experiments at the early-Middle Pleistocene site of Gesher Benot Ya'aqov (Israel). J. Archaeol. Sci. 37, 3170–3190.
Gunz, P., Mitteroecker, P., Bookstein, F.L., 2005. Semilandmarks in three dimensions. In: Slice, D.E. (Ed.), Modern Morphometrics in Physical Anthropology. Developments in Primatology: Progress and Prospects. Springer, Boston, MA. https://doi.org/10.1007/0-387-27614-9_3.
Jadon, S., Garg, A., 2020. Hands-On One-Shot Learning with Python: Learn to Implement Fast and Accurate Deep Learning Models with Fewer Training Samples Using PyTorch. Packt Publishing Ltd.
Jiménez-García, B., Abellán, N., Baquedano, E., Cifuentes-Alcobendas, G., Domínguez-Rodrigo, M., 2020a. Corrigendum to "Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars.". J. R. Soc. Interface 17, 20200782.
Jiménez-García, B., Aznarte, J., Abellán, N., Baquedano, E., Domínguez-Rodrigo, M., 2020b. Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars. J. R. Soc. Interface 17, 20200446.
Kuhn, M., Johnson, K., n.d. Applied Predictive Modeling. Springer New York.
Lele, S.R., McCulloch, C.E., 2002. Invariance, identifiability, and morphometrics. J. Am. Stat. Assoc. 97, 796–806.
Liu, Q., Tian, Y., Zhou, T., Lyu, K., Xin, R., Shang, Y., Liu, Y., Ren, J., Li, J., 2024. A few-shot disease diagnosis decision making model based on meta-learning for general practice. Artif. Intell. Med. 147, 102718.
Maté-González, M.A., Courtenay, L., Aramendi, J., Yravedra, J., Mora, R., Domínguez-Rodrigo, M., 2019. Application of geometric morphometrics to the análisis of cut mark morphology on different bones of differently-sized animals. Does size really matter? Quat. Int. 517, 33–44.
Otárola-Castillo, E., Torquato, M.G., Hawkins, H.C., James, E., Harris, J.A., Marean, C. W., McPherron, S.P., Thompson, J.C., 2018. Differentiating between cutting actions on bone using 3D geometric morphometrics and Bayesian analyses with implications to human evolution. J. Archaeol. Sci. 89, 56–67.
Picq, S., Gaucherel, C., Claude, J., 2014. Momocs: outline analysis using R. J. Statistic.
Pizarro-Monzo, M., Domínguez-Rodrigo, M., 2020. Dynamic modification of cut marks by trampling: temporal assessment through the use of mixed-effect regressions and deep learning methods. Archaeol. Anthropol. Sci. 12, 4.
Pizarro-Monzo, M., Organista, E., Cobo-Sánchez, L., Baquedano, E., Domínguez-Rodrigo, M., 2022. Determining the diagenetic paths of archaeofaunal assemblages and their palaeoecology through artificial intelligence: an application to Oldowan sites from Olduvai Gorge (Tanzania). J. Quat. Sci. 37, 543–557.
Ravichandiran, S., 2018. Hands-On Meta Learning with Python: Meta Learning Using One-Shot Learning, MAML, Reptile, and Meta-SGD with TensorFlow. Packt Publishing Ltd.
Richtsmeier, J., Lele, S., Cole, T.M., 2005. Landmark Morphometrics and the Analysis of Variation. Variation 49–69.
Snell, J., Swersky, K., Zemel, R., 2017. Prototypical Networks for few-shot learning. Adv. Neural Inf. Process. Syst. 4077–4087.
Vegara-Riquelme, M., Gidna, A., Uribelarrea del Val, D., Baquedano, E., Domínguez-Rodrigo, M., 2023. Reassessing the role of carnivores in the formation of FLK North 3 (Olduvai Gorge, Tanzania): a pilot taphonomic analysis using Artificial Intelligence tools. J. Archaeol. Sci.: Rep. 47, 103736.
Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D., 2016. Matching networks for one shot learning. Adv. Neural Inf. Process. Syst. 3630–3638.
Wang, Y.-X., Hebert, M., 2016. Learning to learn: model regression networks for easy small sample learning. In: Computer Vision – ECCV 2016. Springer International Publishing, pp. 616–634.
Yravedra, J., Maté-González, M.Á., Courtenay, L.A., González-Aguilera, D., Fernández, M.F., 2019. The use of canid tooth marks on bone for the identification of livestock predation. Sci. Rep. 9, 16301.