IDENTIFIYING DIFFERENT CARNIVORES’ TOOTH SCORES WITH *DEEP LEARNING* ALGORITHMS: Testing the hominin shift in the balance of power

Trabajo de Fin de Máster

Máster en Investigación en Inteligencia Artificial (Escuela Técnica Superior de Ingeniería Informática – UNED)

Autora: **Natalia Abellán Beltrán**

Director: **José Luis Aznarte Mellado**

Madrid, septiembre 2021

Abstract

Were hominins hunted or hunters? The power over the predation hegemony during human evolution is still nowadays controversial. Nevertheless, the truth is they were both at some point, but it is the shift in the balance of power what remains arguable. One of the ways of answering that question is studying carnivore modification on hominin bones. Many hominin remains present bone surface modifications (BSM) made by carnivores which could result from predators feeding on them primarily or from scavengers (post-depositional). In the present work we try using several computer vision models based on convolutional neural networks to compare five different types of carnivores, firstly jointly and then pairwise. This method shows different outcomes according to specific carnivore taxa, evidencing great results when comparing tooth scores made by lions (*Panthera leo*) and spotted hyenas (*Crocuta crocuta*), for instance, with an accuracy of 92% in the testing set. It also proves the huge potential of *deep learning* algorithms for correct classification of BSM and their implications. Furthermore, we apply some of the models to the tooth mark observed on a femoral shaft from a hominin dated around 500 ka. The results suggest that the carnivore modification may have more likely resulted from post-depositional scavenging instead of predation.

Keywords: Taphonomy · Tooth marks · Convolutional Neural Network · Carnivore · Deep Learning

Resumen

¿Cazadores o cazados? El poder sobre la cima de la cadena trófica durante la evolución humana sigue siendo hoy en día un tema controvertido, aunque la realidad es que los homínidos fueron ambas cosas. Sin embargo, es el momento en el que se produce el cambio de poder entre carnívoros y homínidos lo que sigue dando pie a arduos debates dentro de la comunidad científica. Muchos restos de homínidos presentan marcas de carnívoros, por lo que una buena manera de dar respuesta a la pregunta planteada es analizándolas detenidamente, ya que pueden albergar información sobre si el homínido fue fuente de alimento primaria para un depredador o si son producto de un proceso de carroñeo post-deposicional. En este trabajo comparamos marcas de diente hechas por cinco carnívoros distintos mediante diferentes algoritmos basados en redes neuronales convolucionales, primero de manera conjunta y después por parejas. Este método muestra distintos resultados según el taxón de los carnívoros, alcanzando un 92% de acierto en la clasificación de los datos de validación, por ejemplo, cuando comparamos marcas infligidas por leones (*Panthera leo*) y hienas manchadas (*Crocuta crocuta*). Además, se han probado algunos de los modelos para identificar las marcas de carnívoro que aparecen en un resto de diáfisis de fémur de un homínido que data sobre los 500.000 años. Los resultados sugieren que las marcas de diente fueron probablemente infligidas por carroñeo post-deposicional, en lugar que por depredación.

Palabras clave: Tafonomía · Marcas de diente · Redes Neuronales Convolucionales · Carnívoro · Aprendizaje Profundo (Deep Learning)

Acknowledgements

When I was finishing my degree, I was kind of lost in terms of what to do next, since I was well aware of the difficulties to make a living out of archaeology. For guiding and helping me to find something interesting to pursue and for including me in this amazing project I would like to thank Dr. Manuel Domínguez-Rodrigo and all the IDEA team. Without any of them, the project would not be able to keep going and it is always encouraging be around people who shares the same passion as you: archaeology and science. Thank you for your trust.

For trusting me, the project and letting me in, I need to secondly thank Dr. José Luis Aznarte, because he made it possible for me to enter this master. He has been of great help and understanding with the special situation of having an archaeologist studying artificial intelligence.

It has been two tough years for many reasons for me and because of that, for being there, trying to keep me going, always having some kind words for me and all the trust in the world on my work, my knowledge, and my worth as a whole, as a person, I will strongly thank Dr. José Manuel Maíllo. You are one of the best people I could have encounter in this so particular world which is Archaeology, and I am forever grateful for your help and guidance. Thank you for pushing me when I did not have the strength nor the faith to do it myself.

To my partner in crime, my colleague Blanca Jiménez, for always supporting me and for walking by my side in this arduous process. It is reassuring having you in the same boat.

Finally, without a doubt, thanks to my family and friends for bearing with me during my very low moments, for picking me up and help me move forward, for always, unconditionally being there. You are my strength and I am the luckiest for having you. To my grandmothers and grandfathers: I left my hometown when I was barely eighteen and still, after six years, there is not a single day I do not miss you. You are always on my mind.

A mi flor de vainilla y mi flor de azahar.

INDEX

1. Introduction

Archeology as a scientific discipline was consolidated throughout the nineteenth century, based on three major milestones: the recognition of the "Antiquity of Man" (publication in 1841 by Jacques Boucher de Perthes of convincing evidence of a human existence long before the Biblical Flood, with the support of John Evans and Joseph Prestwich, two important British scholars), the development of the concept of "Evolution" by Charles Darwin (in 1859 publishes *The Origin of Species*) and the establishment of the system of the "Three Ages" (In 1836 CJ Thomsen proposed that museum collections be divided between Stone Age, Bronze Age, and Iron Age). However, the decisive peak occurred in the middle of the 20th century, with the discovery of radiocarbon dating by Willard Libby and, also, with the birth of the so-called "New Archeology" in the 1960s. Lewis Binford, along with other archaeologists, affirmed the great potential that archaeological evidence had for the investigation of the social and economic aspects of past societies, not having to limit itself to the description of objects and possible influences from other societies. Any archaeological interpretation had to be supported by a logical argumentation, being the conclusions of all research work capable of being contrasted by means of the scientific method.

When we think about archaeological assemblages, one of the most common discoveries are bones: human, animal or the ones eaten by them. They are one of the main tools for research and studies about how the people from the past used to live, especially if we refer to Paleolithic archaeological sites, where the only traces remaining are lithic tools and bones. In this sense, taphonomy is crucial, since it is the only way we have to recover all the information left after years and years of decomposition and burial.

The term of *taphonomy* was first used by the Russian palaeontologist Ivan Efremov in his article “Taphonomy: a new branch of Paleontology”, as “the study of the transition (in all its details) of animals remains from the biosphere to the lithosphere”, meaning the study of the processes where organisms pass different parts of the biosphere and become part of the lithosphere, after being fossilized (Efremov, 1940: 84). This concept was used afterwards by archaeologists to describe and study the formation and disturbance of the archaeological record (Lyman, 2010). It became as a field of interest due to research of early hominid evolution during the 70s, that tried to elucidate the agency in modified bones: naturally or by hominids (e.g., Behrensmeyer, 1975; Hill, 1976). Nevertheless, the implications of the concept have been widely discussed, adding some other characteristics, making it evolve to what we nowadays mean by *taphonomy* (Fernández-López, 2006; Rogers *et al*, 2007; Domínguez-Rodrigo *et al*, 2011).

In both palaeontological and archaeological assemblages, loss and bias of the information is the focal point of the studies. It is important to keep these two aspects in mind for making any further interpretations. This is the basis of all neotaphonomic investigations, where all the fragmented information remaining (material culture and artifacts, in archaeology) is used to make reconstructions from past behaviour. Taphonomy creates a balance between the loss of biological information and the taphonomic data collected from the archaeological remains. However, the bias introduced by the researcher is always there: in the identification of the marks, the quantification, interpretation, …

With the present work, we try to reduce to minimum this bias by using artificial intelligence and *deep learning* algorithms. This limits the bias only to the selection of each kind of mark by the taphonomist, that the algorithms use as database. All the marks here have been obtained in controlled environments so that, in each assemblage we are 100% sure that those scores were made my one specific and alone carnivore. In this sense, the possible bias being introduced by us is close to zero. The main objective of the work is to use Convolutional Neural Networks to try to discern types of carnivores only from the analysis of the scores, something never done to date in Archaeology.

Due to the success of this method to differentiate cut marks made on bare bone and on fleshed bone (Byeon et al, 2019; Cifuentes-Alcobendas, 2019), we intend to "test" the limits of this new analysis model.

As stated before, in archaeology, personal bias is always present: from what is decided or not to be collected in the field, to the number of tooth marks observable by the expert. This is one of the reasons why, in cases of controversial marks, such as those presented in this project, the opinion of two experts may be completely opposite, having only their "word" and experience as support. Thus, it is intended to create in the future a large database, made up of photos of all different BSM, which would be accessible to the entire scientific sphere (archaeologically speaking) so that in case of doubt it can be consulted, corroborating or refuting the opinion of the researcher, through *machine learning* and *deep learning* algorithms. Making archaeological research more objective is a way of making it a much more credible and a less "fantasy" science than is believed today.

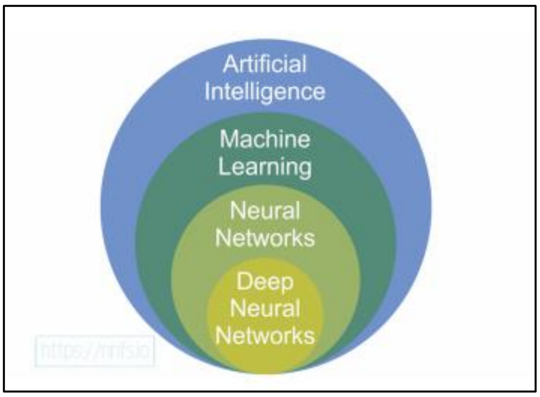

Figure 1: The various fields of artificial intelligence (Kinsley and Kukiela 2020, Fig. 1.01).

The other important part of our project is the use of Machine Learning. Machine learning techniques teaches computers to do what we humans do naturally, that is learn by experience. In our case, artificial intelligence uses algorithms capable of learning directly from the images that are introduced to it, without having prior knowledge about the field of study (that is why its application is so wide, from fields such as medicine to others such as archeology). Normally, in the algorithm learning process, the discriminant characteristics of the objects of interest are first extracted, to be used later in the development of models capable of learning and identifying patterns from the image data. However, initially this feature selection process was done manually, which requires a lot of time and in-depth mathematical knowledge, since feature extraction involves the pre-processing of the images by different operators to discriminate each one. of the parameters that interest us for their classification (Sevillano and Aznarte 2018; Sevillano *et al.* 2020).

Recently, *deep learning* algorithms have been developed that automatically, without human intervention, select and learn these characteristics. This is important because, not only does it reduce the time spent by the researcher on this pre-processing, but it also eliminates any bias that could be introduced by the human hand when making the selection of characteristics. Thus, *Deep Learning* methods are a part of Machine Learning. The term refers to a general principle of "learning multiple levels of composition" (Goodfellow *et al.* 2016).

These types of algorithms are already being used to try to solve problems such as facial recognition, motion detection, autonomous driving for pedestrian detection, automatic parking, etc. Also, in the medical field, they are being used to detect lymph node metastases within breast cancer (Golden 2017); also to perform genetic profiling in order to track diseases and genetic disorders (Chen *et al.* 2016).

Of all the possible image recognition algorithms, Convolutional Neural Networks (CNNs) are mostly used because they have proven to be one of the most powerful *Deep learning* algorithms when classifying images (Krizhevsky *at al.* 2012; Lecun *et al.* 1998). Nevertheless, for a neural network to be a model of *deep learning* they must have two or more hidden layers (most neural networks that are used nowadays have multiple hidden layers, so they all would be a form of *deep learning*) (Kinsley and Kukiela 2020) and for it to be "convolutional" it must contain at least one convolutional layer.

A neural network is an ensemble of interconnected artificial "neurons" that exchange messages with each other (they are inspired by the biological brain) (Goodfellow *et al.* 2016; Kinsley and Kukiela 2020). All connections have numerical weights that the model adjusts during the training process, so that a properly trained network will respond correctly when presented with an image or pattern to recognize (Hijazi *et al.* 2015). Layers are constructed so that the first layer detects a set of primitive patterns at the input, the second one detects patterns of patterns, the third one detects patterns from those last patterns, and so on. Typical CNNs use 5 to 25 different layers of pattern recognition (Kinsley and Kukiela 2020). Their main advantage is that this algorithm eliminates the need for manual feature selection, as we stated before, doing it automatically and extracting the more discriminant features of the set of images (Sevillano and Aznarte 2018). This is one of the reasons why the CNN is the favorite algorithm for artificial vision tasks, like object classification.

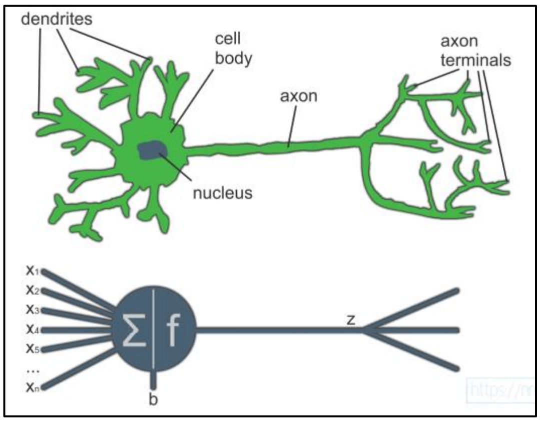

Figure 2: Biological neuron *vs* artificial neuron (Kinsley and Kukiela 2020, Fig. 1.02).

As for the architecture of the Convolutional Neural Network for processing images, it was inspired by the structure of the mammalian visual system (Goodfellow *et al.* 2016). The name of this kind of algorithm indicates that the network uses a mathematical operation called *convolution*, which is a specialized kind of linear operation: “convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers” (Goodfellow *et al.* 2016). This type of NN is a variation of the multilayer perceptron, with the difference that, since its application is performed in two-dimensional matrices, they are more effective for artificial vision tasks, such as image classification and segmentation (Sevillano and Aznarte 2018). The CNNs are specially fitting for processing 2D data. This way, it makes the most of the 2D hidden “convolutional” layers to convolve the features with the input data (Sevillano and Aznarte 2018).

Thus, by combining artificial intelligence algorithms with our sample of images of bone surface modifications we will simply try to answer one question: is it possible for an artificial intelligence to beat the human brain in differentiating these marks?

2. Previous studies

One of the most important questions we must answer when studying any archaeological site is “Who did it?”. This is what taphonomists call agency, and its crucial to make any other interpretations about the site.

Inside of bone surface modifications (BSM), we find the carnivores’ marks. They are very conspicuous and by observing their distribution and frequency on the bones, we can know if the access to the carcass was primary or secondary, namely, if it was preyed on or scavenged. This was decisive in the debate about a hunter or scavenger hominid from the past, along with the study of cut marks (BSM made by hominins when defleshing a carcass). Some authors supported the idea of the hominid as a passive scavenger (Binford, 1981, 1985, 1988; Shipman, 1984; Blumenschine, 1986, 1989, 1991), while others defended combined strategies of scavenging and hunting, that gave hominids primary access to the carcasses (Bunn, 1981, 1982, 1983; Isaac 1983, 1984; Domínguez-Rodrigo 1996, 1997, Domínguez-Rodrigo y De la Torre Sainz, 1999; Domínguez-Rodrigo, 2002; Egeland *et al*, 2007).

Once the main issue of the debate was surpassed and most of the academia recognised the hunting character of early hominids, many thorough studies were carried out about carnivores’ mark themselves, as an attempt to distinguish different carnivore agency or access to the carcasses (primary or secondary), just by looking at their shape and size.

The two more common tooth marks present in bones are tooth pits and scores. They were first described by Blumenschine (1995: 29) as holes with “bowlshaped interiors” (pits, *Fig. 1*) and “U-shaped cross sections” with crushing in the surface, that gives the mark a different patina (*Fig. 2),* whose length is at least three times its width.

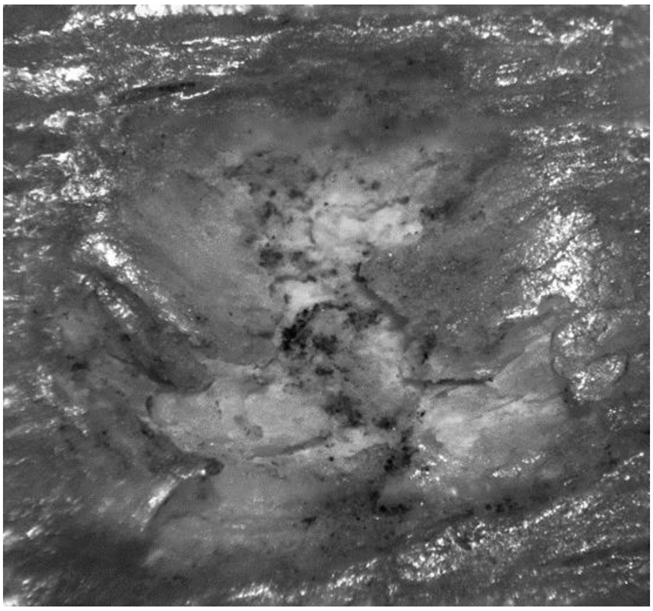

Figure 3: Hyena pit. Photography: self-elaborated.

Since the body size of mammalian carnivores is so different (hyenas, leopards, lions, pumas, lynxes, foxes, wolves, bears), the question about size and shape of tooth marks was raised and studies were developed to try to infer the type of carnivore that made them, by analyzing their main quantitative characteristics (Selvaggio and Wilder, 2001; Domínguez-Rodrigo and Piqueras, 2003; Delaney-Rivera et al, 2009).

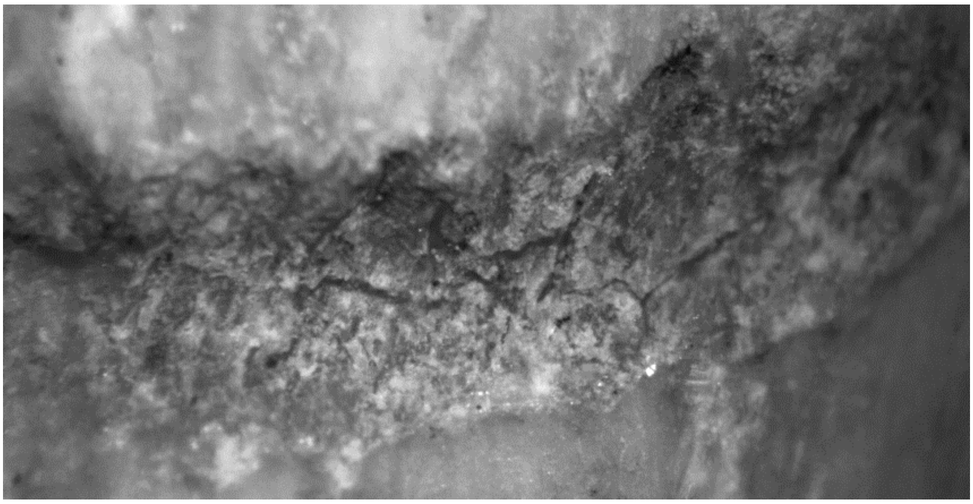

Figure 4: Wolf score. It is easily observable the crushing of the surface and the change of color. Photography: self-elaborated.

Selavaggio and Wilder used the shape of pits to try to recognize the type of carnivore, obtained by the ratio between the major and minor axis of the mark, along with its area in mm. They managed to infer that it was felids who defleshed bones in the FLK Zinjanthropus assemblage and that hyenas had the final access. Nevertheless, the marks in the assemblage showed much more variability in shape than any individual carnivore spices (extant or extinct) present in the comparative sample. Moreover, they could observe that the bone density (cancellous bone, thinning cortical and dense cortical) is directly related to the size of the marks.

In addition, Andrews and Fernández-Jalvo (1997) carried out a study of carnivore punctures to infer more than one type of carnivore in Sima de los Huesos. It showed that the pits size can vary depending on the anatomical element. However, they were only able to divide all carnivore marks in three types: small (<1 mm), mixed (small or large carnivore; 1-4 mm) and larger carnivores (>4 mm). This premise was supported afterwards by Domínguez-Rodrigo and Piqueras (2003), who stated that pits size could indeed be used to differentiate between small and large carnivores, but that they are very ambiguous to stablish specific taxa. The same team of researchers was later able to distinguish human-inflicted tooth marks from those made by other carnivores (Fernánez-Jalvo and Andrews, 2011).

In the same way, Domínguez-Rodrigo and Piqueras (2003) showed skepticism in being able to identify taxa with the scores, since they are subject to much more variability within a single agent than pits. Its length depends on the bite type, portion of the bone, bone size, force of the bite, density of the bone, etc. Nonetheless, the breath could indeed be used as a way to infer the size of the teeth that made it, hence the size of the carnivore.

In 2012, another study about differences of tooth marks was published (Andrés et al, 2012). The analysis includes the largest sample of tooth marks per bone section for some of the most important carnivores, that potentially interacted with the hominins from the Pleistocene, being an active agent in the formation of the archaeological record. They observed that there are two variables statistically detectable and meaningful to differentiate carnivores by tooth marks only: the carnivore size (small versus large) and the carcass size.

On one hand, as previously documented (Delaney-Rivera et al, 2009; Selvaggio and Wilder, 2001; Domínguez-Rodrigo and Piqueras, 2003), small size carnivores could inflict some marks that can overlap with medium size carnivores' ones. The tooth mark dimensions are more related to the size of the carnivore than to the morphology of the tooth: having bears more similar teeth to humans, their tooth marks are closer to the ones of lions and spotted hyenas (Delaney-Rivera et al, 2009; Saladié et al, 2012). Furthermore, the differences in in tooth mark size are more prominent on dense cortical shafts than those in spongy ends. This is logical if we think about the force that impose the dense bone compared to the cancellous bone. The age of the predator is also important, since subadults of the same taxa can generate a very different spectrum of marks (in dimensions), which could be confused by marks of adults from a different size taxon.

On the other hand, the size of the carcass must be considered. Some small carcasses, like sheep or goats, are rather marginal in the predatory range of several large carnivores, in their natural environments. For example, they are not the type of prey consumed by hyenas or lions. However, this changes in game-depleted areas, with high anthropogenic impact in the local ecology, where these predators do consume smaller carcasses. In this sense, it is very important to understand tooth marking of each carnivore within their predatory range and which carcass sizes are represented in it.

Finally, the last noteworthy study of this kind was carried out recently (Aramendi et al 2017, 2019; Yravedra *et al.* 2017), based on geometric morphometrics. It showed that different carnivore types and sizes produce tooth mark samples with their own spectrums They tried to differentiate carnivores' pits (hyenas, jaguars, lions and crocodile) through 3D models and geometric morphometrics, as a way to infer if the fossils OH8 (FLK NN3) and OH35 (FLK Zinj) were preyed by crocodiles or another agent. When considering shape, lions and jaguars presented remarkable differences from the rest of the carnivores, classifying their tooth marks with and accuracy >70%. However, the success of the classification of the other carnivores was below 45%, except for crocodile pits, that were much more successfully recognized from the rest of the carnivores, due to their special and characteristic morphology. Moreover, when using morphometric information based only on form, accuracy was higher, on average. This indicates that the size of the tooth mark is indeed a useful discriminator, even though tooth marks from different carnivores were only successfully classified in approximately 47% of the total sample (range = 37-55%) (Yravedra *et al.* 2017). Furthermore, the figures are probably inflated, since the classification process is done with the same marks that were used for providing the discrimination function (training sample). Nonetheless, the overall analysis shows that lions have the lowest diversity of morphologies among all the carnivores used in this study. They also display considerably small tooth marks regarding their body size (Yravedra *et al.* 2017). Jaguars present wider diversity in every sense, size and shape, compared to lions. On the other hand, durophagous carnivores (hyenas and wolves; they consume not only the flesh but the bone too) show even greater variation in their marks.

In spite of the moderate success in classifying between all carnivores by taxon, this study proves that even though tooth marks of different carnivores are very similar, they still posses features that allow some degree of differentiation, especially between durophagous carnivores (hyenas) and felids (lion), which is already relevant for some important paleoanthropological studies.

Nevertheless, all these taphonomic techniques lack the resolution to discriminate among carnivore types, between taxa. Thus, a lot of the questions involving carnivore agency are still unanswered. The fact that carnivores prayed on hominins during the

Pliocene has been known for many years now (Brain, 1983), but the change in the balance of power remains controversial (Pickering *et al*, 2008; Pickering, 2013; Blasco *et al*, 2010; Cueto *et al*, 2016).

Many hominin fossils present evidence of carnivore damage (tooth marks), but was it peri-mortem or post-mortem, namely, post-depositional? Answering that question is important for establishing agency and if the access was primary (hunted) or secondary (scavenged). In this sense, for instance, it has been debated that the fossil OH7 (holotype specimen of *Homo habilis:* Tobias, 1991) may have been prayed by crocodiles, since it shows presumably crocodile tooth marks. In fact, a new crocodile species (*Crocodylus anthropophagus*) was created under this assumption. This would mean that *Homo habilis* was preyed on since 2 Ma and even later (Brochu *et al*, 2010). However, the empirical evidence from other fossils of hominins, such as OH35, used to support this argument (Njau and Blumenschine, 2012) is questionable (Baquedano *et al*, 2012; Aramendi *et al*, 2017). Furthermore, if these tooth marks could be used to differentiate between carnivore groups (not only by size) the whole passive scavenging behavioral model of hominins from early Pleistocene previously noted, could be reinforced or finally completely rejected.

The same happens in more recent middle Pleistocene hominin accumulations, like Sima de los Huesos or Rising Star, where the lack of resolution in discerning the type of carnivore involvement in their formation and posterior modifications makes their interpretations uncertain. In some taphonomic studies, it is shown that felids may have participated in the formation of Sima de los Huesos assemblage (Andrews and Fernández-Jalvo, 1997), while in a recent revision, other experts suggest that carnivore damage is quite smaller than previously reported, being the result of post-depositional impact by bears (Sala *et al*, 2014; 2015). Something similar happens with the accumulation of *Homo naledi* in Rising Star, interpreted as exempt of carnivore damage (Dirks *et al*, 2015). However, in both cases there is strong evidence to consider that there is a bigger impact of carnivores in the formation of both assemblages (Egeland *et al*, 2018).

Quite a few prehistoric ecosystems sustained a carnivore guild which included, at least, a large and a smaller felid, a large canid, a durophagous carnivore and crocodiles. For instance, Plio-Pleistocene savannas in Africa included lions, leopards, large canids, such as *Megacyon sp.*, *Canis africanus*, *Lycaon;* hyenas and some crocodile taxa. In European Pleistocene ecosystem, we observe *Panthera gombaszoegensis* and *Panthera spelaea* (large and smaller felids), along with wolves and hyenas. The same happens in Quaternary North America, where jaguars (*Panthera onca*), lions (*Panthera atrox*) and crocodiles lived in the same ecosystems at the same time. Large canids also inhabited these environments, some of them with bone crunching capabilities (*Canis dirus*). If taphonomist were able to attribute agency in this kind of habitats, we could gain extensively relevant information from the paleontological record.

The other half of our project begins with understanding what Convolutional Neural Networks are and how they work.

CNNs are *deep learning* algorithms composed of sequences of several layers of artificial neurons, which receive the information (input) and produce an output through a process that mainly involves three elements: weights (weights), biases (bias) and the activation function in each layer. (activation function) (Krizhevsky at al. 2012). The networks receive the images as input and transform them through several hidden layers (the term deep learning refers to the multiple hidden layers). The first convolution layer extracts low-level features like edges, lines; and as the number of layer increases, so does the quality of the selected entities (last layers extract the highest-level features) (Sevillano *et al.* 2020). Each of the hidden layers includes convolutional layers, an activation function, pooling and fully connected layers (Byeon et al, 2019). Weighted inputs are pushed through an activation function that determines the threshold of neural activation and its signal (Cifuentes-Alcobendas and Domínguez-Rodrigo 2019). The signal travels through all the multiple layers and emerges in the last output layer. This output is compared to the expected resolution from a controlled classification and then, the error is backpropagated in the inverse order through each layer, updating the weights in each layer in relation to their contribution to the bias. The more epochs are established, the higher the learning rate, reducing the error sequentially and the bigger the training sample, better results we will obtain.

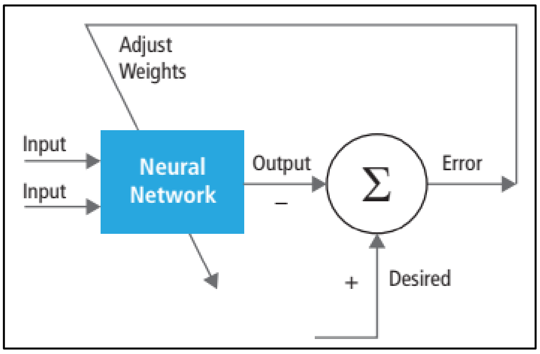

Figure 5: training of neural networks (Hijazi *et al.*, *Figure 2*).

For instance, VGG-19 model is a Convolutional Neural Network that consists of 19 layers. It is a variant of the VGG model (there are some others like VGG11 or VGG16). The algorithm was developed by the Visual Geometry Group (VGG) at Oxford (Simonyan and Zisserman, 2014a, b) and it was first runner up of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2014. ImageNet is an image database organized according to the WordNEt hierarchy: every node of the hierarchy is depicted by hundreds and thousands of images (https://www.image-net.org/). It also hosts the ILSVRC, where researchers from around the world are challenged to find that yields the lowest Top 1 and Top 5 error rates, where Top 5 error rate is composed by the amount of images whose correct label is not one of the model's five most likely labels. In this competition, the models are given up to 1.000 class training set over 1.2 million of images, a 50.000 images for the validation set and 150.000 images for the test (https://www.image-net.org/). This competition is a way to improve the models and push the existing boundaries, as it happened with VGG-19 in 2014, a model that surpassed all previous models results.

ResNet50 was also winner of this competition IN 2015, being trained with approximately 1.2 million of images, as VGG-19. In this kind of deep networks, residual learning is used. In this way, rather than hoping that every stacked layer fits a desired underlying mapping, it expressly let the layers fit a residual mapping (He *et al.* 2016). These networks imitate cortical neurons in biology (human or animal): the apical dendrite skips layers, at the same time, basal dendrites recollect signals from previous layers. In summary, residual networks skip connections, getting a shortcut to fit the input from the layers before on to the next ones without modifying the input (He *et al.* 2016). Moreover, ResNet50 was winner of MS COCO 2015 (COCO is a large-scale object detection, segmentation and captioning (recognition in context) dataset (https://cocodataset.org/#home).

Inception layers work as shortcut branches and few deeper branches for deep networks ("highway networks") (He *et al.* 2016; Srivastava *et al.* 2015). This kind of layers improve the speed and accuracy of the learning process of neural networks. There are several of them, such as Inception v1, v2, v3, v4, Inception ResNet. Every version of this network is an iterative improvement over the one before. InceptionResNet constitutes a hybrid, based on the premise of introducing residual connections that add the outcome of the convolution function inside the inception module to the input

(Szegedy *et al.* 2017). The network Inception, also called GoogLeNet (Google and “LeNet” for professor Yan LeCun’s LeNet) (LeCun *et al.* 1998), was winner of the ILSVRC 2014, competing with the already mentioned VGG-19 and it has been vastly used for image classification problems.

Furthermore, we have NASNet, created by the Google Brain team, which is a team of researchers investigating on artificial intelligence with the goal of “improving people’s live” (https://research.google/teams/brain/). With this network the developers proposed searching for an architectural building block for a small dataset to transfer it afterwards to a much bigger one. Specifically, the idea behind it was to search for the best-performing convolutional layer on the CIFAR-10 dataset (labeled subset of the 80 million-instances Tiny Images dataset, consisting on 60.000 color images in 10 classes; 6.000 images per class: https://www.cs.toronto.edu/~kriz/cifar.html), to apply it afterwards to the ImageNet dataset (Zoph *et al.* 2018).

Lastly but not least, we have DenseNet 201 and Jason networks. After LeNet5 composed by 5 layers, VGG-19 having 19 and Highway Networks and ResNets surpassing 100 layers each, DenseNet 201 became a milestone with its 201 layers (Huang *et al.* 2018). With the appearance of so deep CNNs, a new problem arose: as all the information on the input passed through so many layers, sometimes it disappeared or “washed out” by the time it reached the end or the beginning of the network (Huang *et al.* 2018). This was addressed by the already mentioned Highway Nets and ResNet, by skipping information from one layer to another through identity connections. At the end, what they do is create shortcuts from the first layers to the last ones. In this sense, what DenseNet 201 proposes is a simple connectivity pattern: for ensuring maximum information flow through all the layers in the network, DenseNet connects all layers with matching feature-map sizes directly with each other (Huang *et al.* 2018). In addition, every layer acquire further inputs from all previous layers and passes on its own feature-maps to the next ones, in order to preserve the feed-forward nature of the network. The effectiveness of this kind of network was tested on three different dataset: the already mentioned CIFAR (in this case with the two datasets: CIFAR-10 and CIFAR-100), SVHN dataset (Street View House Numbers; contains 32x32 colred images; 73.257 in the training set and 26.032 in the testing, with 531.131 additional images for training the network) and ImageNet for the ILSVRC 2012 (1.2 million of images for training the net and 50.000 with 1.000 classes for the validation set) (Huang *et al.* 2018). As for Jason2, this comprehensive model was created by Jason Brownlee (Brownlee 2017). It is a sequential model inspired in the structure of VGG-16 and VGG-19 models (Pizarro-Monzó *et al.* 2020). It is a model much simpler that all the ones described above, but it was interesting to introduce since in *deep learning* not always more means better: for instance, overfitting can occur if we expose an inadequate sample to a too complex network, since it learning process is too fast and it becomes "incapable" of improving its results.

In summary, in the present work, we test the power of these pre-trained CNN to classify tooth score marks from five different types of carnivores simultaneously: lions, spotted hyenas, jaguars, wolves and crocodiles. We will also do a pairwise comparison between lions and spotted hyenas and, between both agents and crocodiles, given their relevance to many of the interpretations about hominin-carnivore interactions in the formation of the Africa early Pleistocene archeological record (Abellán *et al.* 2021). Ultimately, we are targeting the discrimination of modifications identified on hominin bones too, in order to assess when the shift in the balance of power took place during human evolution. For this purpose, we will use one preliminary example from the site of Thomas Quarry I (Morocco), where a hominin femur has been modified by carnivores (Daujeard et al. 2016).

3. Method

Sample

For the tooth sample, the marks were obtained from controlled experiments, making sure that only one carnivore intervened. We collected tooth marks from four different carnivore mammals and one reptilian. The crocodile was included with the purpose of checking how diagnostic their ichnological assemblages were compared to those of mammal carnivores (Abellán *et al.* 2021).

For de crocodiles (*Crocodylus niloticus*), eight female crocodiles were used: one small (1.3 m in length from nose to end of tail), two medium-sized (1.8 m) and five large (2.3 m to 10 m). They were feed in an enclosure area of the zoo Faunia (Madrid, Spain), once a week over four complete months with 19 partial carcasses. Carcasses were collected after 15h of exposure to crocodiles, even though most part of the feeding took place during the first hour. The feeding process was monitored for the first 1.5 h, to be able to relate carcass part consumption to individual crocodiles. These carcass parts were composed of articulated limbs of suids (pig and boar) and bovids (sheep and cow). They were prepared by butchers, who removed most feet bones, except in two limbs. A total of 198 elements were retrieved, counting every end and shaft of unfused bones from juvenile individuals as one (Abellán *et al.* 2021).

With the concern of using adequate parameters in analogical reasoning, we used three carnivores (lions -*Panthera leo*-, spotted hyena -*Crocuta Crocuta*- and jaguar -*Panthera onca*-) from Cabárceno reserve (Cantabria, Spain), where animals live in open spaces (areas comprising several thousands of square meters: http://www.parquedecabarceno.com) and do not undergo typic behaviors that carnivore display in cages or small enclosures (Gidna *et al.* 2015). They were fed with the same type of carcasses at regular intervals: equid carcass limbs, being the bones collected after a few days of exposure (when they were completely defleshed and unattended, which usually spanned 1–4 days). In the case of the hyenas, the protocol had to be modified since when bones were exposed for more than 1 day, they tended to be completely consumed. Consequently, the bones in their enclosure were collected earlier only after a few hours of consumption, on the same day most of the times (Abellán *et al.* 2021). This may be due to their durophagous behavior, since hyenas consume the bone as well as the meat, in comparison with felids that tend to avoid touching the bone.

All these predators were fed equid long limb bones to keep the structural and substantial parts of the experimental analogy the same (Abellán *et al.* 2021)—with some variation in the environmental part (Gidna et al. 2013)—conform to Bunge's principles of correct analogy (Bunge 1981)-. The characteristics of the samples and experimental conditions described above over this bone assemblage, was widely explained in the original studies (Gidna et al. 2013; Dominguez-Rodrigo et al. 2015).

The experimental assemblage of bones modified by wolves bones was obtained at Monte Campelo, in the northwest of north Spain. More specifically, the bone collection consists of remains of some colts hunted and consumed by wolves in the summer-autumn of 2009 and of a group of mares scavenged also by wolves during the winter. At the end, 17 carcasses were analyzed. The sample was collected at the kill sites and their surroundings. To be able to confirm and control the modifying agent of the bones, infrared cameras activated by movement were installed in the area, which showed that wolves were the only carnivores that altered the bones. A total of 379 elements were collected and properly identified from the site (Yravedra et al, 2011).

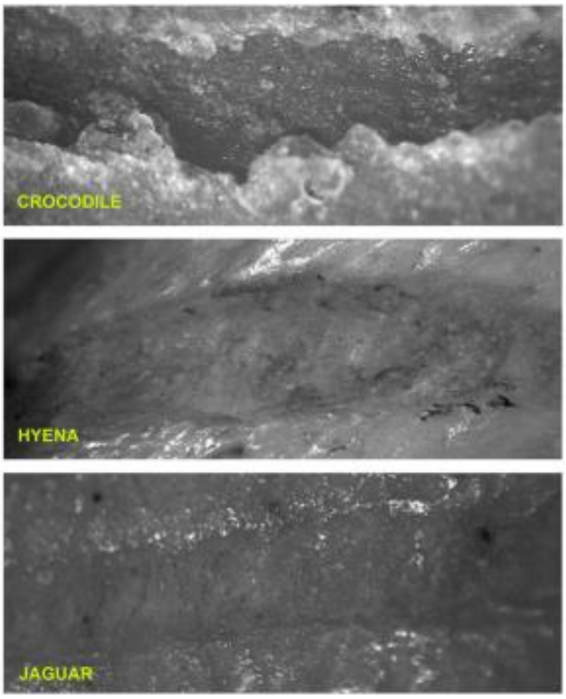

Figure 6: crocodile, hyena and jaguar scores (Abellán *et al.* 2021, Fig. 1).

For all the carnivores involved in the present study, we only used tooth scores, since they are the BSM never included before in this kind of experiment (*Fig. 3* y *4*). Tooth pits were also documented, but they were removed from this experiment, because they produced smaller and unbalanced sample (Abellán *et al.* 2021). The total tooth score sample consists of 591 tooth scores: 207 made by lions, 42 made by jaguars, 207 tooth marks made by wolves, 48 made by crocodiles and 80 by spotted hyenas. Crocodiles leave many anatomical elements complete and without any marks (Njau and Blumenschine 2005, 2012; Baquedano *et al,* 2012; Sahle *et al,* 20), because they usually eat the part completely, sometimes even in anatomic connexion, although the bones that had marks, could have from 1 up to 20 (pits and scores). This makes the recovery of tooth marks more difficult, thus, the smaller number of tooth scores recovered.

The smaller sample from the jaguar- and crocodile-modified bones could be a potential bias, because their small size may preclude any substantial accuracy in CNN methods. Nevertheless, this places the accuracy of the present study on a lower threshold than would probably correspond if their sample size was bigger (Abellán *et al.* 2021). To try to counter this possible effect of the small sample size of those two assemblages, the complete tooth mark sample was artificially increased using image augmentation techniques as is usually recommended (Chollet 2017; Brownlee 2017). All BSM images were collected with a binocular microscope (Optika) at × 30, using the same light intensity and angle, in order to keep the different samples as similar as possible for the later analysis. Successively, they were transformed into black and white shades to standardize them during image processing in the Keras platform, using bidimensional matrices for standardization and centering. They were also reshaped to the same dimensions (80 × 400 pixels). The Keras library was used with the TensorFlow backend (Abellán *et al.* 2021). Finally, this image data bank was used for analysis through the CNN models, explained in detail in the following section.

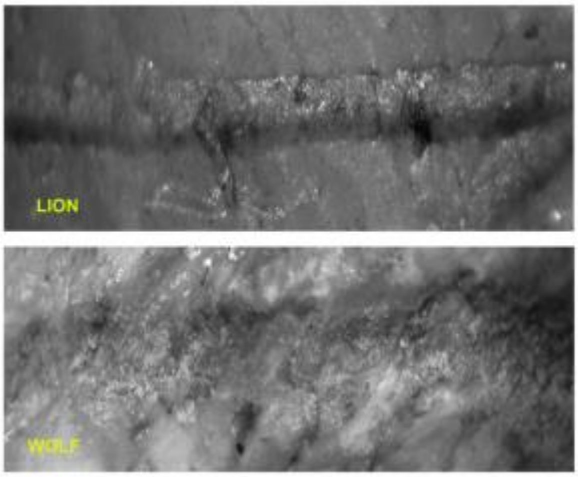

Figure 7: lion and wolf score (Abellán *et al.* 2021, Fig. 1).

As for the analyses, two kind were carried out, the first one involving a simultaneous comparison of all the carnivores; and the second one focusing on marks made from spotted hyenas and lions; lions and crocodiles; and hyenas and crocodiles, due to their potential interaction among themselves and with Plio-Pleistocene African hominins. The pairwise comparisons are important because carnivore-hominin interaction have been modeled after a felid-hominin pattern. If we are able to identify carnivore agency with this model, differentiating between lions and hyenas as potential post-depositional carcass scavengers, we can finally validate or reject completely some carnivore-hominin models. Also, interactions with hominins involves potential predation on hominins (more likely resulting from felids) or post-depositional modification imparted by scavengers, like hyenas. For this second type of analysis, we will use the three most successful models from the previous generic carnivore comparison (Abellán *et al.* 2021).

Moreover, to show the great potential of application of these referential analogs, we have selected one hominin femur that supposedly belongs to *Homo heidelbergensis*, from the site of Thomas Quarry I (Morocco), which has marks inflicted by carnivores (Daujeard *et al.* 2016). We chose the best example of tooth score from among all the BSM present in this femur to be analyzed by the CNN model and classified as either felid or hyenid (which are the only two possible agents capable of modifying the specimen to the extent that was reported; thus, it was only necessary a pairwise comparison). The CNN model was applied to the original published photograph of the specimen (Abellán *et al.* 2021). We selected this hominin specimen because it was used to discuss whether Middle Pleistocene hominins were still prey to the larger carnivores or just a post-depositional source of carrion (Daujeard *et al.* 2016), so the results of the analyses, whatever they may be, would be important in trying to put an end to the debate. During the Middle Pleistocene, hominins and other mammal predators competed over the same resources, which ecologically results in predation on competitors (Schaller 2009). Middle Pleistocene North African ecosystems contained large felids, such as *Homotherium* and lions, who lived with hyenas (both *Hyeana* and *Crocuta*) and small canids, like jackals and foxes, but large canids were extremely rare (Abellán *et al.* 2021). Additionally, the modifications described on the ThI94-UA28-7 femur diaphysis (the epiphyses were chewed off) can only be explained by the action of a large felid, such as a lion or a hyena, taphonomically speaking, since the thickness of the shaft would have prevented other carnivores from having modified it to the extent that is documented (Abellán *et al.* 2021). The tooth marks cluster and the “chewing” in the ends of the shafts, also reinforce the hypothesis that it was a large carnivore with a strong durophagous behavior the one who modified the bone. Furthermore, the tooth sizes observed on the specimen fall within the range for only lions, hyenas and bears. However, bears do not damage long bone epiphysis with furrowing, nor they consume them completely (Arilla et al. 2014). Therefore, the two most likely species for having modified the specimen are lions and hyenas. Because of this, we will test de models for the hominin remains using only these two taphonomic agents.

Method: Deep Learning and CNN models

As explained previously, CNNs are a form of *deep learning* since they have several hidden layers and they differentiate from normal neural networks (Fig. 8) on the presence of, at least, one convolutional layer.

Before starting our model, it is important to always partition the data, at least between *training* and *testing*, so that we will put the network to the test by making predictions on a part of the data that has not yet known.

Thus, after training the neural network, it should be tested with a verification sample. This shows the percentage of precision and the predictive classificatory potential. The loss function, the optimizer and the metric selected to monitor the function of the network are the most crucial elements of this part of training. The loss function measures how the network is behaving with the training data. The optimizer updates the network according to the loss function and finally the selected metric monitors the performance of the network in the training and test data sets ("accuracy") (Kinsley and Kukiela 2020).

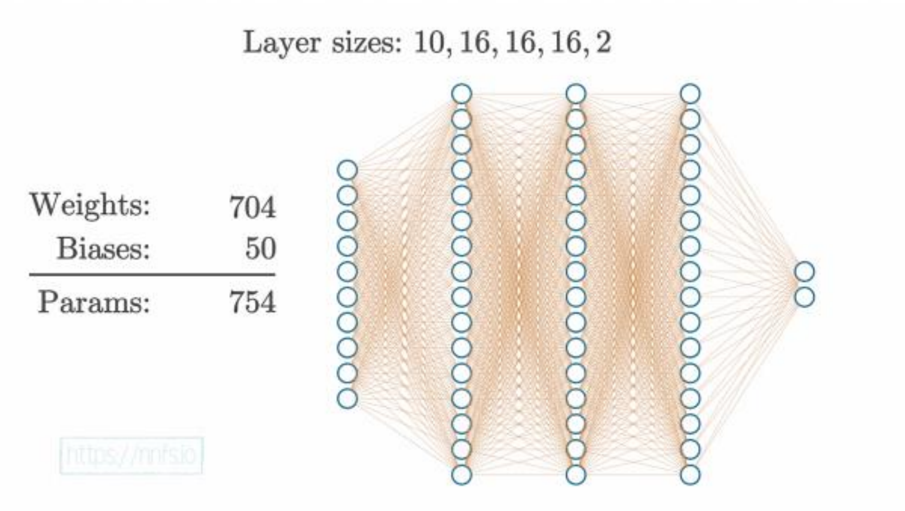

Figure 8: example of a neural network with 3 hidden layers, 16 neurons each (Kinsley and Kukiela 2020).

The pooling/subsampling makes features robust against noise and distortion. It replaces the output of the net at a certain location with a summary statistic of the nearby outputs (Goodfellow *et al.* 2016). There are two basic ways to do pooling: average and max pooling. In both cases, the input is divided into non-overlapping two dimensional spaces (Hijazi et al, 2015).

The activation function are layers that work as the "trigger" function to signal distinct identification of likely features on each hidden layer, computing the hidden layers values. They "describe" the features. CNNs can use a variety of specific functions, such as rectified linear units (ReLU) and continuous trigger (non-linear) functions, to efficiently implement this. The activation function ReLU implements the function y = (max(0,x)) (Goodfellow *et al.* 2016). This increases the nonlinear properties of the decision function and the overall network without affecting the receptive fields of the convolution layer. The advantage of the ReLU is that the network trains many times faster. This activation function is the most recommended one when using feedforward neural networks (Goodfellow *et al.* 2016).

On the other hand, fully connected layers are often used as the final layers of a CNN. They mathematically sum a weighting of the previous layer features, indicating the precise mix of "ingredients" to stablish a specific target output result. In this case, all the elements of all the features of the previous layers get used in the calculation of each element of each output feature.

In this way, the whole neural network forms itself in the following way. Firstly, sequential layers are created with a starting convolutional layer. It applies a convolutional process to the input, that computes the dot products of the weights of the layers and a small region that is connected to the input layer (Kim, 2017). This layer works as a receptive field that has predetermined dimensions and slides of a certain number of small areas, creating a feature map. After this, pooling layers use feature maps to compress information and generalize the features, reducing the overfitting of the training data. This generates a second feature map. It is the alternation between convolutional and pooling layers that forms most part of the neural network, finishing with the fully connected layer. That flat feedforward neural layer uses the non-linear activation function, which depends on the classification problem at hand (for multiple categories: softmax, for example) (Cifuentes-Alcobendas and Domínguez-Rodrigo, 2019).

The most successful CNN models use millions of parameters. Several of these architectures (e.g., Alexnet, ResNet50, VGG16, Inception, GoogleNet) have been winners of the Imagenet Large Scale Visual Recognition Challenge (ILSVRC), the largest competition of image classification (Abellán *et al.* 2021). Some of these models are capable of identifying up to 1000 categories with accuracy rates > 90%.

Nevertheless, training them from the beginning requires powerful computation. Luckily, several of these models, already built and trained on hundreds of objects can be used as pre-trained architectures. This also enables other projects which are further away from computational sciences, like ours, to use this method for their analyses. Transfer learning consists of using a model that has already been trained for a different problem, exposing it to types of images for which the model was not trained in origin (Goodfellow et al. 2016). Training for complex features, such as the 1000 image categories, makes pre-trained models very efficient in detecting minor features that identify different categories. This includes microscopic features of BSM. Here, we use a selection of the most successful pre-trained models to classify tooth marks from the five selected carnivore taxa (Abellán *et al.* 2021)

Previously, in a similar study where we used the same approach, nine of the most successful architectures (some of them winners of the ILSVRC competition) were compared in their accuracy when classifying correctly tooth scores of very similar carnivores like lions and jaguars (Jiménez-García *et al*. 2020a, b). Although most of the models yielded very similar accuracy, the most successful classifiers: VGG19, DenseNet 201, InceptionResNetV2, NASNet Large, ResNet50 (pre-trained models) and a simpler architecture based on VGG16-19 modules (Jason2) (Jiménez-García et al. 2020a). Here, we use these six architectures to test the accuracy in classifying tooth marks from such a diverse set of carnivores. Afterwards, we select the two most successful models to make pairwise comparisons between selected carnivores. The characteristics of every architecture and their parameters were summarized in Jiménez-García et al. (2020a), Domínguez-Rodrigo et al. (2020) and Abellán *et al.* (2021):

- **VGG19**. VGG-16 and VGG-19 architectures were winners of the ILSVRC in 2014 (Simonyan and Zisserman 2014a, b). VGG-16 architecture consists of more than 138 million parameters. It was composed of 16 layers with weights at first, organized in a series of 3 × 3 kernel CNN sequentially so that they followed each other with increasing depth, spanning from 64 filters to 512 filters in duplicated sequences. This way, VGG-19 was born as an extension of VGG-16, with a total of 19 weighted layers. The matrix size was reduced applying max-pooling layers in-between neural layers (Simonyan and Zisserman 2014a, b).

- **ResNet50.** As previously stated, this is a deep residual network, with 50 layers in total. It became winner of the 2015 ILSVRC with an error of only 3.5% in classifying the test set of ImageNet dataset. Its architecture uses residual functions that improve the training of extremely deep networks. These functions are used via a skip connection and allow to pass the input through blocks without having to pass through weight layers. It helps with the problem, usually present in very deep CNNs, of vanishing gradient. This way, through this architecture, it is possible to train a residual CNN made up by more than 100 layers. Specifically, this model extends the VGG repeated layer blocks typical of VGG-16 and VGG-19 architectures. Each block is three layers deep. The first one (initial) has 64 7 x 7 kernel filters, followed by max pooling 2 x 2 kernel layer. The second one is a block of three layers, one contains 64 1 x 1 kernel filters, one 64 3 x 3 and one formed by 256 1 x 1 kernel filters; this block repeats three times. Afterwards, there is another series of CNN, composed of four blocks of two 128 filter units and one of 512 units. They use the same size of filter as the previous block. This is succeeded by one series formed of six blocks of two 256-filter layers and one 1024-filter layer. Finally, the last series has three blocks with three layers each: two with 512 filters and one with 2048. The model is topped with an average pooling layer and fully connected layer (He *et al*. 2016).
- **InceptionResNetV2**. This model has a depth of 164 layers, combining series of different blocks, like the last one explained. The stem block or initial one is composed of a 5x Inception-ResNet-A block, Reduction-A block, 10x Inception-ResNet-B, Reduction-B block, 5x Inception-ResNet-C, average pooling layer and a 0.2 dropout layer. The Inception block A is formed of 1 x 1 and 3 x 3 CNNs; the block B is made up by a combination of 1 x 1 and 1 x 7 CNNs. The third or block C is composed of 1 x 1 and 1 x 3 CNNs. One positive point to highlight of this model is that it is computationally more efficient than some other highly ambitious options, like Inception-V4 (Szegedy et al. 2017).
- **NASNet Large**. “NAS” stands for Neural Architecture Search, since it is the first model not designed directly by analysts: it is the result of using reinforcement learning search methods, through recurrent neural networks (RNN). It uses some CNNs to produce a feature map of the same dimensions

(normal cells) and other CNNs whose feature map is reduced by order of two (reduction cells). The controlling unit is a RNN composed of 100-hidden-unit layer that uses a softmax prediction activator. Then, the RNN processes the joint probability distribution, selecting the most probable classification options. This is operated via parallel computation, because the original model spent 500 GPUs and > 2.000 GPU hours to build the conv cells for the architecture of the model. This way, NASNet Large has become one of the models with highest accuracy existing today (Zoph *et al*. 2018).

- **Jason2**. Its architecture was created in 2017 (Bronwlee 2017), simplifying the VGG architectures. The model consists of three blocks of double layers of 32, 64 and 128 neurons (3 x 3 kernels), separated by max pooling layers (2 x 2). Inside every block, there is a batch normalization layer, finishing with dropout layer that varies in a increasing proportion per block: 0.2, 0.3 and 0.4. Then, there is flattening and a dense layer of 128 filters, followed again by a dropout of 0.5 and lastly a dense layer with softmax activation function. Each CNN has been tuned with a “He uniform” kernel initializer and with padding of the type “same” (Abellán *et al.* 2021).
- **DenseNet 201**. This model constitutes a very deep network with a total of 201 layers. Every sequential layer obtains the feature maps of all the previous layers as inputs, resulting in new feature maps that are passed through subsequently to the following layers. This way, the network becomes thinner and condensed, thus, easier to work with. This structure, combined with the depth of sequential CNN layers, enables the detection of a wider diversity of features in images compared to other alternative architectures (Abellán *et al.* 2021). The network model is built on dense CNN blocks of 1 x 1 and 3 x 3 sequential layers, separated by transition blocks of 1 x 1 CNN and 2 x 2 pooling layers (Abellán *et al.* 2021). Moreover, the sequence of CNN for every dense block is repeated six times for the first block, twelve times for the second one; 24, 32, 48 and 64 times for the third one and finally, 16, 32 and 48 times for the last one. The final transitional layer is a global average pooling layer (7 x 7) (Abellán *et al.* 2021).

For every model, we used the already mentioned ReLU activation function, for the last fully-connected layer a “softmax” activation function was used (joint comparison) and a “sigmoid” one for the pairwise analysis; the loss function chosen was categorical cross-entropy, since it was suitable for comparison with multiple outcomes. Specifically, for the pairwise comparison, we used a binary cross-entropy: it measures distances between probability distribution and predictions (Chollet 2017; Abellán *et al.* 2021). As for the optimizer, we selected Stochastic Gradient Descent (SDG): learning rate 0.001 and momentum 0.9. Finally, for compilating the model we used “accuracy” as metric, to observe the percentage of success in the classifications (Abellán *et al.* 2021).

Furthermore, we tried data augmentation in our sample to try avoiding overfitting by artificially increasing our sample. We considered it necessary since two of the carnivores’ samples were significatively smaller than the rest, due to the factors previously explained. This is usually recommended when dealing with imbalances samples or datasets that are not big enough (Goodfellow *et al.* 2016; Chollet 2017), because it intensifies the heuristics of the NN (Chollet 2017). In the present work, the samples were augmented through random transformation of the original images, shifting width and height (20%) and also shear and zoom range (20%). We included horizontal flipping and a rotation range of 40º too (Abellán *et al.* 2021). In the case of the pairwise comparison, we tested the models with and without augmentation.

The complete image dataset was divided in two parts for implementing the models: one for the training (70%) and other for the testing of the different architectures with unknown data (30%). Both for the test sample and for the training one, we used mini-batch kernels: training size 64 for the multiple comparison and size 32 or 20 for the pairwise analyses; testing size 32 for multiple comparison and size 20 for the pairwise ones. For the weight update we applied a backpropagation process of 100 complete epochs. All the analyses were carried out in a HP workstation (GPU), using Python 3.7 through Jupyter Notebook platform, being the total computation time for the whole study one week (Abellán *et al.* 2021).

As previously stated, for the pairwise comparisons we used only the three most successful model, considering the ones with highest accuracy and lowest loss, always taking into account a moderate or high level in the balance of the classification (Abellán *et al.* 2021). The architecture of the models was not changed for these analyses, excepting the loss function (categorical cross-entropy was replaced by binary cross-entropy since we only had two possible outcomes) and the activation function for the fully-connected layer ("softmax" was changed for "sigmoid" function for the same reason) (Abellán *et al.* 2021).

As for the ensemble learning, we imported the model with all the layers, aside from the top fully connected layer from the output-end of the model (Abellán *et al.* 2021). This way, the training algorithm is able to fine tune the weights for feature-map extracting layers and then use that information to create a new fully connected top layer specifically for our problem, generating a prediction. To achieve this, we replaced the top layer with a flattened one and added the fully-connected layers of the new classifier (Abellán *et al.* 2021).

In addition, we tried stacking ensemble learning (SEL) after all the comparison of the different models was done (Jiménez-García *et al.* 2020b). This method assembles the various classification algorithms into a single classifier (Abellán *et al.* 2021). This way, a baseline set of predictions derived from the original classification algorithms is generated, which are used afterwards by a meta learner to create an aggregate final classification (Abellán *et al.* 2021). The stacking method is recognized for being stronger at classifying than other ensemble methods, in particular, single-trained ones (Wolpert 1992). In the present work, we use four of the most successful models as base learners and we produce other three different SEL architectures with three distinct meta-learners (Abellán *et al.* 2021). Firstly, we generated 100 trees with tuned random forest, without specifying the maximum depth. We did specify the number of features selected through the square root of the feature range. Afterwards, tuned extra-randomized trees was used to produce 100 trees. Lastly, we applied the third meta-learner: gradient boosting tree, up to 500 different models (Abellán *et al.* 2021). This method resulted in different models that were contrasted with a testing dataset, carefully watching the degree of balanced classification. The numeric results are shown in *Table 2* (multiple carnivore analyses), since we observed better results than with the single models (Table 1).

Finally, for the marks present on the femur of the hominin, we applied a gradient visualization technique to detect the possible microscopic features that influenced the

BSM classification, using Grad-CAM (gradient weighted activation mapping algorithm) (Jiménez-García *et al.* 2020a). It uses the weighted activation to generate a heat map overlaying the original image, based on gradients of the predicted class obtained from the last convolutional feature map (Abellán *et al.* 2021). The highlight areas in the mark are the ones that the algorithm considered the most important for the prediction and the classification of the BSM.

4. Results

Multiple-carnivore comparisons: single models

Here I present the results of all the analyses carried out.

Table 1: Accuracy and loss values for every of the model architecture used on the testing set of tooth marks from the five selected carnivores.

	Accuracy	Loss
DenseNet 201	**57.02**	1.58
Jason2	50.72	1.94
VGG19	50.72	1.94
ResNet50	34.06	1.75
InceptionResNetV2	45.65	2.23
NASLarge	37.68	1.45

As shown in *Table 1*, when comparing the single models, the one that yielded the highest accuracy was DenseNet 201, presenting a 57.02% of correct classification of the tooth marks. Moreover, Jason2 and VGG-19 both have shown the exact same accuracy rate: 50.7% (*Table 1; Fig. 9*). In the rest of the models, we observe significant lower scores: ResNetInceptionV2 reaches a 45.6% success rate; being the accuracy of both ResNet50 and NASNet-Large below 35% accuracy. Thus, we observe that the models with similar architecture (Jason2 is a simplified version of VGG-19), succeed DenseNet 201 model as the best ones for classifying BSM with our image dataset (Abellán *et al.* 2021).

In the classification matrix (*Table 3*), we notice that Jason2 and VGG-19 produce the same results, being the ones of the DenseNet 201 also fairly similar, although it shows different misclassification values. *A priori*, the moderate rate of success of these tree models indicates that the tooth marks of the carnivores, when compared mixed together, are broadly similar. However, they are distinctive enough to be well-classified at least half of the times.

Table 2: Classification markers for the multi-carnivore tooth mark sample, Models DenseNet 201, Jason 2, VGG19 and ensemble learning models.

DenseNet 201

	Precision	Recall	F1-score	Support
Crocodiles	0.00	0.00	0.00	13
Hyena	0.86	0.30	0.44	20
Jaguar	0.00	0.00	0.00	10

Wolf	0.47	0.85	0.60	47
Lion	0.66	0.56	0.61	48
Micro avg	0.53	0.53	0.53	138
Macro avg	0.40	0.34	0.33	138
Weighted avg	0.51	0.53	0.48	138
VGG19-Jason2				
	Precision	**Recall**	**F1-score**	**Support**
Crocodiles	0.00	0.00	0.00	13
Hyena	0.73	0.55	0.63	20
Jaguar	0.00	0.00	0.00	10
Wolf	0.55	0.66	0.60	47
Lion	0.54	0.71	0.61	48
Micro avg	0.55	0.55	0.55	138
Macro avg	0.37	0.38	0.37	138
Weighted avg	0.48	0.55	0.51	138
Ensemble learning model (with random forest)				
	Precision	**Recall**	**F1-score**	**Support**
Crocodiles	0.20	0.08	0.11	13
Hyena	0.82	0.70	0.76	20
Jaguar	0.25	0.10	0.14	10
Wolf	0.75	0.70	0.73	47
Lion	0.60	0.85	0.71	48
Micro avg	0.65	0.65	0.65	138
Macro avg	0.53	0.49	0.49	138
Weighted avg	0.62	0.65	0.62	138

Nevertheless, even though we observe this moderate accuracy, that alone cannot be followed. Due to the different sized of the samples, it is important to analyze the precision, recall and F1 score, since they will show us if our classification is balance or unbalance. Looking at *Table 2*, we would have to say that our **classification is unbalanced:** VGG-19 and Jason2 present a global macro-average of 0.37, indicating that there is a high degree of bias when classifying all five carnivores. Surprisingly, DenseNet 201 shows even greater imbalance in the classification (macro-average F1 of 0.33), despite having a higher overall accuracy (*Table* 1). As for the taxa, hyenas, wolves and lions present a high accuracy in classification, with a F1 score above 0.6 in models Jason2 and VGG19. In DenseNet, the hyena's image collection generated more misclassification. In the case of the crocodiles and the jaguars, we have very poor results: they were widely misclassified in all the models (*Table 2).* This is very likely related to the size of their samples, since they were significantly smaller compared with the rest of the carnivores. Therefore, we can conclude that hyenas, wolves and lions are the carnivores best classified. When using the more balance models, BSM of lions were classified in the testing dataset with a 70% success rate, the ones made by wolves with 66% and the tooth scores inflicted by hyenas with 55% accuracy. In contrast with crocodiles and jaguars, this moderate accuracy is a victory.

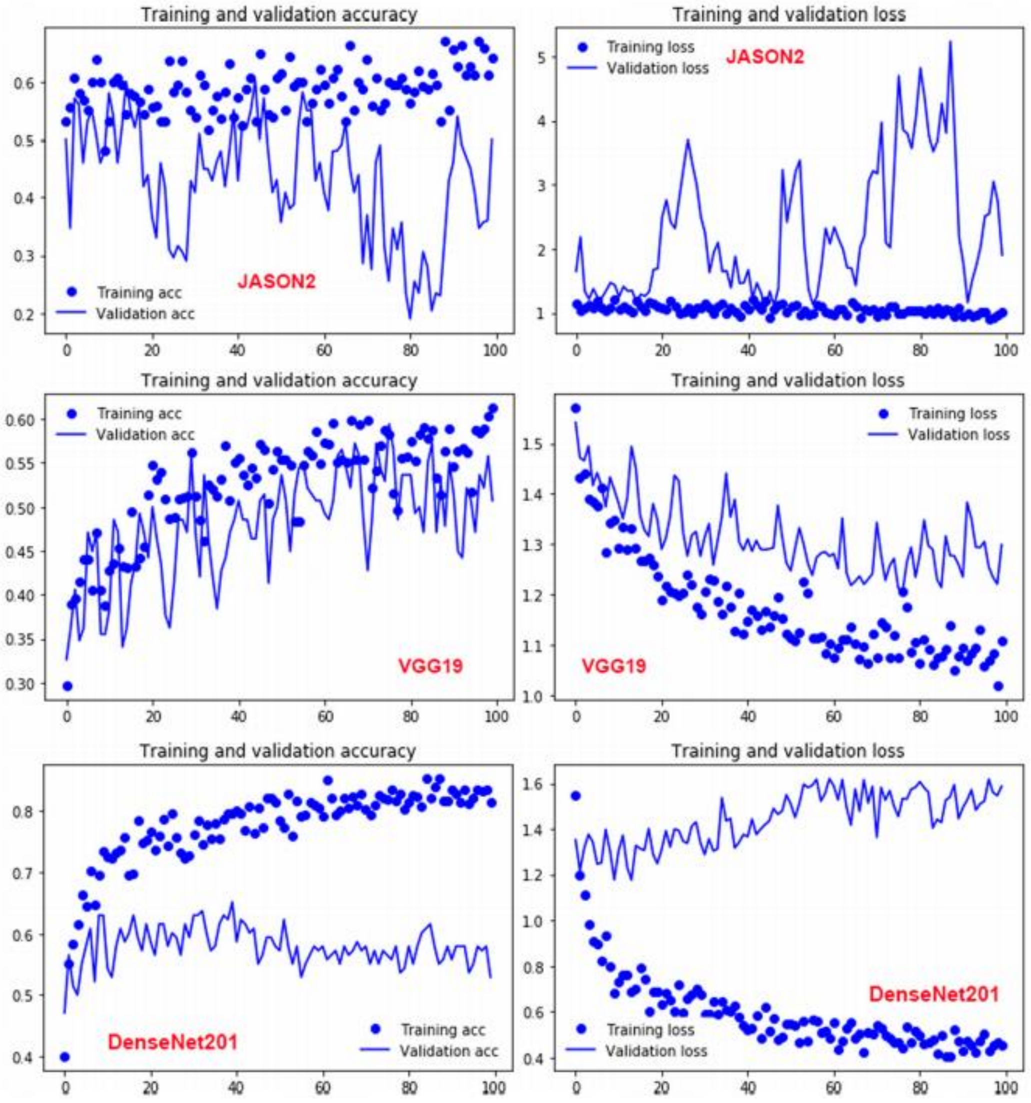

Figure 9: accuracy and loss values for three of the models classifying tooth marks from the five carnivores.

In fact, none of the tooth scores made by lions and wolves, aside from 2% of the wolf marks that were misclassified as hyena's, were classified as any of the other carnivores. In contrast, up to 46% of the crocodile scores were classified as lion marks and around 29% of the lion tooth scores were misclassified as wolf's. Nevertheless, only 5% of the hyena marks were classified as lion tooth scores, but 70% of the jaguar BSM were completely misclassified as lion scores (*Table* 3).

Additionally, DenseNet 201 model showed moderately low loss and the greatest accuracy percentage, but unfortunately, we observed a elevated level of instability and increasing loss values, identified as a clear sign of *overfitting* (*Fig.* 9). On the contrary, VGG-19 presented better stability and increasing accuracy through the training of the network (*Fig. 9)*. In comparison, Jason2, even though it yielded the same accuracy as VGG-19 (a little lower loss), displayed wide fluctuation in accuracy and loss, while VGG-19 presented consistency and limited fluctuation. These three models were also applied to the pairwise comparison among the different African carnivores selected for this study.

Table 3: Confusion matrix for DenseNet 201, Jason 2 and VGG19 models, showing percentages (without decimals) (*n*=138) of correct (diagonal) and incorrect classification of images from the testing set according to carnivore type. Numbers in parentheses are for raw data of the testing sets for each carnivore. Predicted follows horizontal.

DenseNet 201	**Crocodile**	**Hyena**	**Jaguar**	**Wolf**	**Lion**
Crocodile	0 (0)	0 (0)	8 (1)	69 (9)	23 (3)
Hyena	0 (0)	30 (6)	0 (0)	65 (13)	5 (1)
Jaguar	0 (0)	0 (0)	0 (0)	30 (3)	70 (7)
Wolf	0 (0)	2 (1)	6 (3)	85 (40)	6 (3)
Lion	0 (0)	0 (0)	0 (0)	44 (21)	56 (27)
VGG19-Jason2	**Crocodile**	**Hyena**	**Jaguar**	**Wolf**	**Lion**
Crocodile	0 (0)	15 (2)	0 (0)	38 (5)	46 (6)
Hyena	20 (4)	55 (11)	0 (0)	20 (4)	5 (1)
Jaguar	0 (0)	10 (1)	0 (0)	20 (2)	70 (7)
Wolf	0 (0)	2 (1)	0 (0)	66 (31)	32 (15)
Lion	0 (0)	0 (0)	0 (0)	29 (14)	70 (34)

Multiple-carnivore comparisons: stacking ensemble learning models

Examining the results of the single models, the SEL models were generated combining DenseNet, Jason2, VGG-19 and InceptionResNetV2 as base learners (Abellán *et al.* 2021). The results and probabilities of their classification were jointly used for the upper level meta-learner, in three architectures involving a random forest, extra-randomized tree and gradient boosting tree. The *Table 4* presents the results for the three models, while *Table 5* presents the confusion matrix of the most successful model (random forest). As the tables show, every SEL model generated greater accuracy percentage than any of the single models tested before, when comparing the five carnivores all together. Only the extra-randomized tree produced a slightly below accuracy than DenseNet 201 (56.5%): with the other two models we observe a important improvement in the balance score and classification accuracy (*Table 4).* This way, the model that uses a random forest as a meta-learner, reached an accuracy of 65.3% with a F1 score of 0.49. This value in accuracy is more than 3.5 times the expected from random classification. As for the classification of the specific carnivores, we can observe the same as with the rest of the models: the taxa with larger samples have better classification (hyenas, wolves and lions have an accuracy over 70% in the testing dataset (*Table 2* and *Table 5)*; in the case of lions, the tooth scores of the testing dataset are correctly classified in more than 70% of the cases. This gives us even more confidence in the plausible differentiation of tooth marks inflicted by hyenas and lions.

Table 4: SEL análisis with different combinations of base meta-learners and accuracy, loss and balanced (F1-score) classification results.

Base learners	**Meta-Learner**	**Accuracy**	**Loss**	**F1-score**
DenseNet 201	Random Forest	65.3	0.51	0.49
VGG19	Extra-randomized trees	56.5	0.89	0.41
Jason2	Gradient boosting trees	62.5	0.59	0.47
InceptionResNetV2				

Table 5: Confusion matrix (from the SEL model using the random forest as the meta-learner) displaying percentages of correct (diagonal) and incorrect classification of images from the testing set according to carnivore type. Numbers in parentheses are for raw data of the testing sets for each carnivore. Predicted follows horizontal.

SEL (random forest)	**Crocodile**	**Hyena**	**Jaguar**	**Wolf**	**Lion**
Crocodile	7.6 (1)	0 (0)	0 (0)	30.7 (4)	61.5 (8)
Hyena	15 (3)	70 (14)	0 (0)	5 (1)	10 (2)
Jaguar	0 (0)	0 (0)	10 (1)	20 (2)	70 (7)
Wolf	2.1 (1)	2.1 (1)	4.2 (2)	71 (33)	21 (10)
Lion	0 (0)	4 (2)	2 (1)	8 (4)	85.5 (41)

Pairwise comparisons

On the pairwise analyses, the ensemble models showed lower accuracies than the best single models. This is why, we decided not to include them here. These results uphold Domínguez-Rodrigo *et al.* (2020) modeling, that presented some single models performing more efficiently than the ensemble ones when classifying cut, tooth and trampling marks. However, with the pairwise comparisons, model VGG-19 showed more efficiency than DenseNet 201 and Jason2 (*Table* 4). All three models convene in presenting better results with hyenas and lions, reinforcing their easier differentiation, in comparison with crocodiles or jaguars (*Fig. 8)*. In contrast with the efficiency shown by the models when comparing the five carnivores, when performing pairwise comparisons, the VGG19 model exhibited a better performance compared to DenseNet and Jason2 (*Table 4*). The three models converged in showing that hyenas and lions were easier to differentiate than any of them compared to crocodiles.

We also experimented with augmentation to try and reduce the bias introduced by the crocodiles and jaguar samples, as explained before. Nonetheless, with the non-augmented samples we observed better classification percentage than with the augmented ones (*Table 6*).

In the comparisons by pairs in the testing set we noticed the following: Jason2 yielded an accuracy of 91% when differentiating lion inflicted marks from hyenas' marks, with a loss value of 0.43 and a F1 macro-average score of 0.85 (*Table 6; Fig. 10)*; the VGG-19 model becomes the best performing one with an accuracy of 92,5% in the same comparison (loss of 0.004 and F1 macro-average of 0.91). When comparing lion and crocodile tooth scores with VGG-19 we also obtained a moderate accuracy of 77.1% (loss of 0.66 and F1 macro-average of 0.62).

On the other hand, the comparison between crocodile and hyena tooth scores showed very poorly results with all the models tried out. The highest accuracy value was achieved by VGG-19 model with 66.67% of correct classification (loss of 0.27 and F1 macro-average of 0.55). Thus, it is crucial to compare different models and architectures to obtain better resolution in BSM classification. In the models tried with augmented data we observed low F1 score, which indicates unbalances classification: all lions tooth scores were correctly classified over the ones of hyena and crocodile, but there is an important proportion of the hyena and crocodile marks that were misclassified as lion BSM. Specifically in this case, the sample without augmentation shows greater accuracy than the augmented one and also a higher degree of balance (*Table* 6). Nevertheless, most of the pairwise analyses show F1 score values too low in relation to their accuracy, which means imbalance, only with the exception of the lion-hyena comparison, where the score for the degree of balance and the accuracy is quite high, showing more balance classifications (*Table 6*).

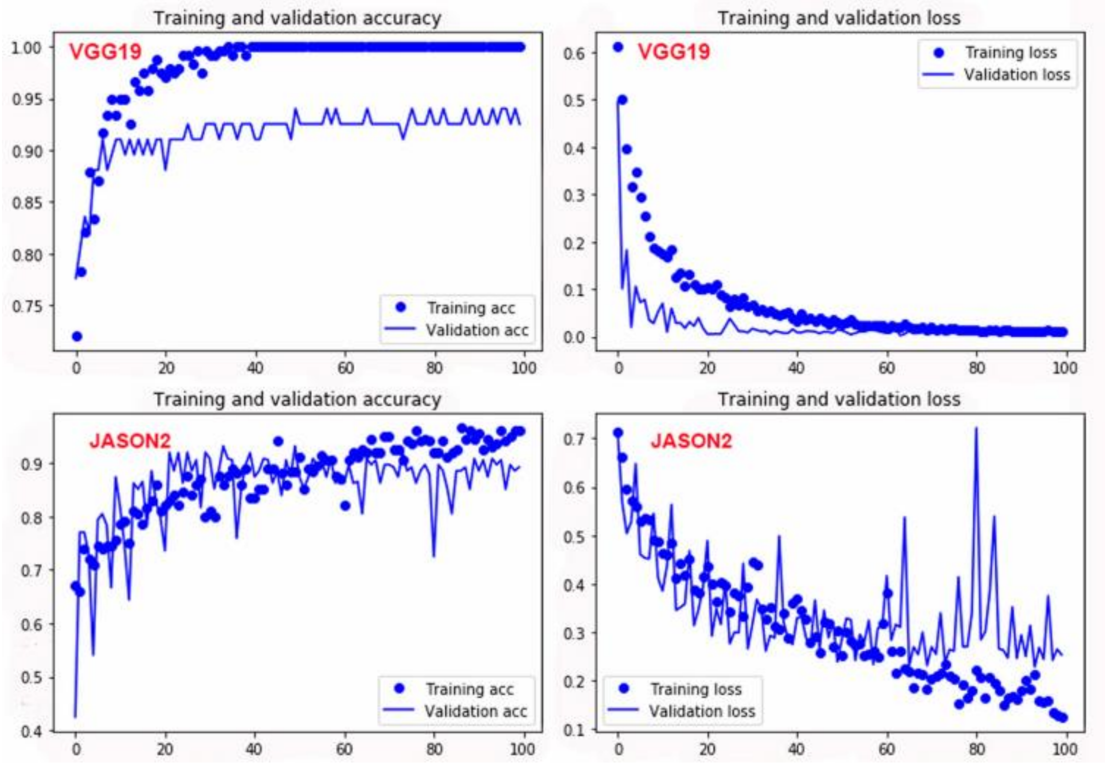

Figure 10: Accuracy and loss values for the VGG19 and Jason2 models comparing lion and spotted hyena tooth scores.

Table 6: Accuracy and los of the VGG19 and Jason2 models when comparing tooth marks made by lions, crocodiles and hyenas. Only models in each pairwise category with the highest accuracy and high F1 values are highlighted. Bold numbers indicate the most successful models.

	Accuracy	Loss	F1 macro avg
Jason2 (lion-hyena) no augmentation	**91.04**	**0.43**	**0.85**
VGG19 (lion-hyena) no augmentation	**92.54**	**0.004**	**0.91**
DenseNet (lion-hyena) no augmentation	77.61	0.002	0.63
Jason2 (lion-hyena) with augmentation	70.15	0.58	0.41
VGG19 (lion-hyena) with augmentation	70.15	0.60	0.41
DenseNet (lion-hyena) with augmentation	70.15	0.56	0.41
Jason2 (lion-crocodile) no augmentation	73.00	0.74	0.43
VGG19 (lion-crocodile) no augmentation	**77.1**	**0.06**	**0.62**
DenseNet (lion-crocodile) no augmentation	74.0	0.003	0.43
Jason2 (lion-crocodile) with augmentation	51.52	1.71	0.43
VGG19 (lion-crocodile) with augmentation	76.67	0.50	0.54
DenseNet (lion-crocodile) with augmentation	74.47	0.41	0.58
Jason2 (hyena-crocodile) no augmentation	56.62	2.17	0.53
VGG19 (hyena-crocodile) no augmentation	**66.67**	**0.27**	**0.55**
DenseNet (hyena-crocodile) no augmentation	60.61	0.014	0.38
Jason2 (hyena-crocodile) with augmentation	51.52	1.7	0.47
VGG19 (hyena-crocodile) with augmentation	45	1.7	0.42
DenseNet (hyena-crocodile) with augmentation	66.67	0.004	0.53

Applaying the models to the ThI94-UA28-7 hominin fossil

As explained in the methodology, we applied the VGG19 and Jason2 models to interpret the carnivore modification on the femur fossil from Thomas Quarry I site (Morocco), attributed to *Homo heidelbergensis* (ThI94-UA28-7). The results showed greater probability that the specimen was modified by hyenas (*Table 7; Fig. 11*).

Table 7: Probability distribution for the attribution of the tooth score form the hominin femur (ThI94-UA28-7) to carnivore agent. Tabulation is made comparing the two most successful models with all the carnivores and pairwise between lions and hyenas. Numbers in bold indicate the probability of the selected agent.

	Crocodile	Hyena	Jaguar	Wolf	Lion
VGG19 all	0.078	**0.601**	0.003	0.071	0.23
Jason2 all	0.039	**0.936**	0.001	0.019	0.002
VGG19 lion-hyena	-	**0.981**	-	-	0.018
Jason2 lion-hyena	-	**0.963**	-	-	0.037

With VGG-19 model, the probability of the mark being inflected by hyena was 98.11% and 1.89% for lion. We observe similar results with Jason2 model: 96.3% for hyena and 3.7% for lion. In both cases, we used the non-augmented samples, since from the results already presented, it was clear that in pairwise comparison, they performed better this way (*Table 7)*. When comparing all the carnivores' samples, the augmented sample showed a 60% probability for hyena with VGG-19 and a 93% probability of hyena with Jason2 model (*Table 7)*. Considering everything we just presented and since the hyena tooth marks tended to be misclassified more than the lion ones, the high probability percentages we obtained for the fossil points to hyena as the perpetrator and reinforces its attribution.

5. Discussion

With the present study we clarified many questions we had. Firstly, we realized that not always the single models with the greatest performance are the best, because some of them can generate quite unbalanced classifications. Consequently, we have considered VGG-19 and Jason2 networks as the best models for our goal of classifying carnivore tooth marks, even though DenseNet 201 showed a slightly better accuracy (57%).

The moderate success in classifying the tooth scores (around 50-57%) indicates that not all the tooth marks inflected by the different carnivores are the same. However, they still present a large overlap among them, supporting the results obtained by the geometric morphometric analyses carried out on the same bone assemblages by Yravedra *et al.* 2017, where they reached an accuracy scarcely below 50%. Nonetheless, we also learned that the ensemble learning models introduce a great improvement over the single ones and the geometric analyses previously addressed, since with those we obtained a 65.3% of correct classification of all five carnivores' tooth scores. Some geometric morphometric methos reached higher accuracy in agent classification (Yravedra *et al.* 2019; Courtney *et al.* 2019), but we observed some problems.

Firstly, the sample. The sample per agent/taxa is around 30 cases each, which we consider insufficient for the analyses to have statistical significance, more so if they had to split it intro training and testing sets. The samples were bootstrapped before the splitting and this generated that some testing samples were already included during the learning process (training), thus the higher accuracy would not be real. It is necessary, for those results to be believable, that the samples are enlarged, being the testing set completely independent from the training one.

As for the comparison among all the carnivores, it could be argued that all of them can never be found in the same ecosystems, but this analysis is still important an relevant since it can be extended to similar carnivores (wolves for wild dogs, for example) in extant biomes.

For the crocodile tooth marks, it has been argued that is one of the very few examples of bone surface modifications that are specific for this agent through their attributes: bisected marks and scores with parallel microstriations (Njau and Blumenchine 2006; Baquedano *et al.* 2012). Even so, these specific marks are only a small proportion of the whole BSM inflected by this carnivore: many of other crocodile tooth scores resemble ones inflected by other mammals predators, as we have proved in this study, observing the very low accuracy yielded by any model on crocodile tooth marks and their misclassification with marks inflected by durophagous carnivores (bone eating carnivores, like hyena).

In any case, as showed in Jiménez-García *et al.* (2020a), there are significant differences between specific carnivores, like lion and jaguar or, as shown in the present study, between lions and spotted hyenas. It may be because lions are flesh-eating carnivores, while hyenas and jaguars (far less than hyenas) are more durophagous in their behavior (they also consume part of the bone) (Domínguez-Rodrigo *et al.* 2015). Thus, their tooth marks are different: ones are inflected with more force on the bones, which generate a bigger range of shape and size, whereas the others are cause of an accident during the defleshing process. This would be why scores made by jaguars, hyenas and wolves showed an intense overlapping in the results presented here. Nonetheless, for the jaguar sample, here the results were surprising, given the great results obtained in Jiménez-García *et al.* 2020a), with jaguar tooth scores correctly classified above 80% of the times when compared with lion ones. Our explanation in the present work is that the values here are due to the already mentioned overlapping between jaguars, wolves and lions: jaguar tooth scores were misclassified as lions' 70% of the times in the testing set. Since jaguar and lions are both feline, they have similar tooth morphology, hence the overlapping, even though their carcass consumption behavior is different.

There may be another reason why the jaguar tooth scores were misclassified. In Jiménez-García *et al.* (2020a), they created a balanced subsample formed by 42 images of marks made by jaguars and 42 marks made by lions, as a way to compensate the unbalance between both. The original image dataset was shuffled and randomly sampled 42 images of each agent. There was a decrease in the accuracy due to the smaller sample, but the classification of both agents' tooth marks was balanced., reaching with VGG-19 model an accuracy of 75.6% with a F1 score of 0.71. More recently, with the same model and a new shuffled and randomly selected sample the model obtained 83% accuracy with a F1 score of 83 (for lions it was 82 and for jaguars 85). However, in Jiménez-García *et al.* (2020a) there was a huge difference between precision and recall, having well classified mostly of the lion marks but systematically misclassifying an important proportion of the jaguar sample. This happened because when using the small subsample of images, the probability of including the minor part of the lion sample is small and, thus, they obtained similar values in accuracy and F1 score. Nevertheless, when they used the larger lion sample, the "jaguar-looking" portion of the lion tooth scores was enough to create a low precision-recall for the testing sample of the jaguar, because the model saw those tooth scores similar to the ones documented in lions. This could mean that the sample of jaguar tooth scores is too small to be statistically meaningful, even though they reached high accuracy in Juménez-García *et al.* (2020a). However, by using this dataset differently, there may be different results, since there might be a problem, both methodological and biological. Ensemble learning was also used in Jiménez-García *et al.* (2020b), managing to surpass the imbalance accuracy and reaching over 82% accuracy in the classification of jaguar marks and more than 92% when classifying lion tooth scores. In consequence, we could affirm that those two tooth marks are potentially differentiable, even though they are both felids and hence, have similar dental morphologies.

The present work also shows that the scores inflected by hyenas and lions can be greatly differentiated, with accuracies over 90% with some models. This is, indeed, very relevant for assessing and interpreting agency in the formation of some African Plio-Pleistocene archaeological and paleontological assemblages.

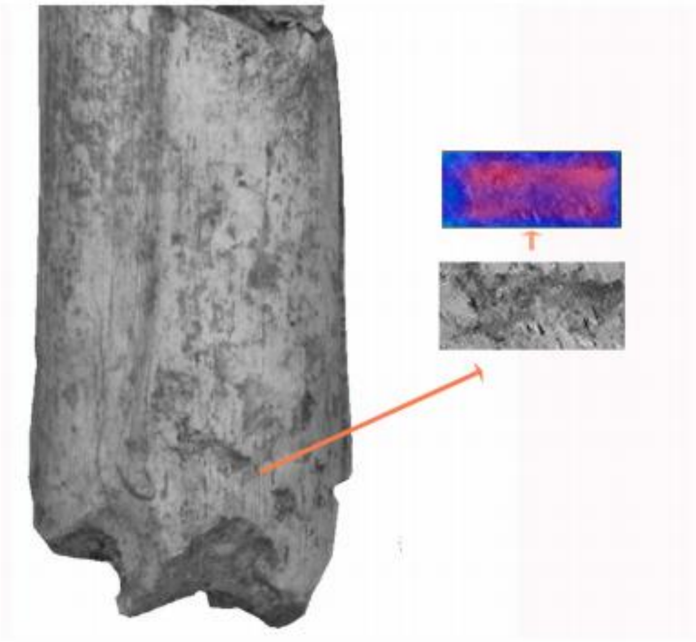

Figure 11: Distal femoral shaft (ThI94-UA28-7) showing carnivore damage and tooth marking (modified from Daujeard *et al.* 2016). The best-preserved tooth score analyzed is shown and highlighted. The heat map over-lying the tooth scores indicates that the CNN model is confident considering the shape and internal features of the groove as indicators of agency. This heat map was made with the Grad-CAM algorithm. For a detailed explanation of its structure and application, see Cifuentes-Alcobendas and Domínguez-Rodrigo (2019) and Jiménez-García *et al.*2020a.

Here we achieved a high level of accuracy, followed by a great degree of balance and therefore, our results with some of the model should be enough for it to be reliable. This way, we could apply it to archaeological BSM to try and she more light on whether the tooth scores on some mid-shaft broken bones are the result of post-depositional hyenas consume (supporting the idea of hominin primary access to the carcass) or lion initially defleshing the carcasses (Abellán *et al.* 2021). This could give us a better glimpse into the lives of the early hominins and how they were taphonomically modified (prayed or not by other carnivores).

6. Conclusion

The present work over tooth marks classifications has shown that micromorphological features of tooth scores are very similar among different carnivores, as previously stated by the morphometric analysis carried out on the same dataset (Yravedra *et al.* 2017). Nevertheless, there are enough differences to distinguish between flesh-eating and more durophagous carnivores, especially when compared pairwise. This fact has important implications for the inference of agency in bone modifications, particularly when considering different agent modifying hominin bones or in sites with equifinality situations (more than one agent can produce similar modifications in bones surfaces, confusing interpretations) (Abellán *et al.* 2021).

On the other hand, the high probability (over 96%) of the tooth score on the hominin bone as result of the action of a hyena, supports the original hypothesis of it being modified by hyenas (Daujeard *et al.* 2016). Furthermore, it proves that the CNN model is extremely reliable when its confidence (probability estimates) is high. The paucity of carnivore-impacted hominin remains could be interpreted as resulting from the control of hominins of the competitive arena with other predators, at least during the middle and upper Pleistocene. Here and thanks to the reassurance over the hyena inflicted mark on the hominin fossil, we also reinforce the interpretation that hominin were on the dominant side of the balance of power during this time, around 500.000 ka., if the pattern documented in this fossils specimen is representative of the general hominin interaction at the time.

Nonetheless, a larger sample of bones modified by carnivores is needed to confirm or reject this preliminary hypothesis. Also, for reaching better results in the multi-carnivore classifications it is undoubtedly essential to extend the dataset.

With the present work, we would like to highlight the urgency to analyze hominin remains from a taphonomic point of view, something that has not been systematically done until present (Abellán *et al.* 2021).

On the side of artificial intelligence, it is also an achievement. In a sphere so close and so interdisciplinary at the same time as Archaeology, it is encouraging that some of these methods are becoming of high relevance, as a way to improve our science and objectivity.

The main issue that has to be addressed is the bias and subjectivity present in any archaeological assemblage. This is introduced mainly by the investigator and the team that studies the remains, those who just by being taphonomist have the authority to say “this is a cut mark” and it has to be believed by the rest, just because.

Of course, knowledge is always necessary and we don’t pretend to turn completely to machines or to let them do all the work for us, but if we can reduce the bias, we surely should try. With the present work, the bias introduced by us was very much reduced, just to the election of the marks that had to be photographed. The remains came from controlled experiments with carnivores (enclosed or semi-free), so that all the marks present in the bones belonged doubtlessly to that specific carnivore.

In this way, there was no much room for error or subjectivity from the investigator’s part and that’s what we wanted the most. In addition, the methodology used and developed is rather simple, as a way for it to expand as further in the academic world as possible. By creating this database of tooth marks, cut marks, trampling marks, etc. and uploading it to a program like R in form of a library, anybody who was studying any archaeological assemblage can have access to it, hence all the interpretations will be achieved through the same method, creating a uniformity inside the academia that is still completely inexistent.

As future projects, the first thing to do is to expand the sample of all carnivores present in this study, especially those of jaguar and crocodile, to create more balance classification. Moreover, it would be also interesting to add some other carnivores to the sample and combine different deep learning algorithms for differentiating the marks, as well as introducing other BSM as percussion marks, biochemical or diagenetic marks.

7. References

ABELLÁN, N.; JIMÉNEZ-GARCÍA, B.; AZNARTE, J.; BAQUEDANO, E. and 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", *Archaeological and Anthropological Sciences,* 13(31). https://doi.org/10.1007/s12520-021-01273-9

ANDRÉS, M.; GIDNA, A. O.; YRAVEDRA, J. and DOMÍNGUEZ-RODRIGO, M. (2012): "A study of dimensional differences of tooth marks (pits and scores) on bones modified by small and large carnivores". *Archaeological and Anthropological Sciences,* 4 (3), pp. 209-219. https://doi.org/10.1007/s12520-012-0093-4

ANDREWS, P. and FERNÁNDEZ-JALVO, J. (1997): "Surface modifications of the Sima de los Huesos fossil humans". *Journal of Human Evolution,* 33, pp. 191-217. https://doi.org/10.1006/jhev.1997.0137

ARAMENDI J, ARRIAZA MC, YRAVEDRA J, MATÉ-GONZÁLEZ MÁ, ORTEGA MC, COURTENAY LA, GONZÁLEZ-AGUILERA D, GIDNA A, MABULLA A, BAQUEDANO E, DOMÍNGUEZ-RODRIGO M (2019): "Who ate OH80 (Olduvai Gorge, Tanzania)? A geometric-morphometric analysis of surface bone modifications of a Paranthropus boisei skeleton", *Quaternary International*, 517, pp. 118–130. https://doi.org/10.1016/j.quaint.2019.05.029

ARAMENDI, J.; MATÉ-GONZÁLEZ, M. A.; YRAVEDRA, J.; ORTEGA, M. C.; ARRIAZA, M. C.; GONZÁLEZ-AGUILERA, D.; BAQUEDANO, E. and 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)". *Paleogeography, Paleoclimatology, Paleoecology,* 488, pp.93-102. https://doi.org/10.1016/j.palaeo.2017.05.021

ARILLA M, ROSELL J, BLASCO R, DOMÍNGUEZ-RODRIGO M, PICKERING TR (2014): "The "bear" essentials: actualistic research on Ursus arctos arctos in the

Spanish Pyrenees and its implications for paleontology and archaeology", *PLoS One*, 9(7). https://doi.org/10.1371/journal.pone.0102457

BAQUEDANO, E.; DOMÍNGUEZ-RODRIGO, M. and MUSIBA, C. (2012): "An experimental study of large mammal bone modification by crocodiles and its bearing on the interpretation of crocodile predation at FLK Zinj and FLK NN3". *Journal of Archaeological Science*, 39, pp. 1728-1737. https://doi.org/10.1016/j.jas.2012.01.010

BEHRENSMEYER, A. K. (1975): "Taphonomy and paleoecology in the hominid fossil record". *Yearbook of Physical Anthropology*, 19, pp. 36-50.

BINFORD, L. R.

1981: *Bones: ancient men, modern myths.* New York: Academic Press.

1985: "Human ancestors: changing viewa of their behaviour". *Journal of Anthropological Archaeology*, 4, pp. 292-327.

1988: "Fact and fiction about the Zinjanthropus Floor: Data, arguments and interpretations". *Current Anthropology*, 29, pp. 123-135.

BLASCO R, ROSELL J, ARSUAGA JL, BERMÚDEZ DE CASTRO JM, CARBONELL E (2010): "The hunted hunter: the capture of a lion (Panthera leo fossilis) at the Gran Dolina site, Sierra de Atapuerca, Spain". *Journal of Archaeological Science*, 37, pp. 2051–2060. https://doi.org/10.1016/j.jas.2010.03.010

BLUMENSCHINE, R. J.

1986: *Early hominid scavenging opportunities. Implications of carcass availability in the Serengeti and Ngorongoro ecosystems.* Oxford: B.A.R. International Series, 283.

1989: "A landscape taphonomic model of the scale of prehistoric scavenging opportunities". *Journal of Human Evolution*, 18, pp. 345-371.

1991: "Hominid carnivory and foraging strategies, and the economic function of early archaeological sites". *Philosophical Transactions of the Royal Society*, London, 334, pp. 211-221.

1995: "Percussion marks, tooth marks and the experimental determinations of the timing of hominin and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania". *Journal of Human Evolution,* 29, pp. 21-51.

BRAIN CK (1983): *The hunters or the hunted?: an introduction to African cave taphonomy*. University of Chicago Press.

BROCHU CA, NJAU J, BLUMENSCHINE RJ, DENSMORE LD (2010): "A new horned crocodile from the Plio-Pleistocene hominid sites at Olduvai Gorge, Tanzania", *PLoS One* 5(2). https://doi.org/10.1371/journal.pone.0009333

BROWNLEE J (2017): *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery.

BUNGE M (1981): "Analogy between systems", *International Journal of General Systems*, 7(4), pp. 221–223

BUNN, H. T.

1981: "Archaeological evidence for meat-eating by Plio-Pleistocene hominids from Koobi For a, Kenya". *Nature,* 291, pp. 574-577.

1982: *Meat-eating and human evolution: studies on the diet and subsistence patterns of plio-pleistocene hominids in East Africa.* Ph. Dissertation, University of California, Berkeley,

1983: "Evidence on the diet and subsistence patterns of Plio-Pleistocene hominids at Koobi Fora, Kenya, and at Olduvai Gorge, Tanzania". In (J. CLUTTON-BROCK, ed.) *Animals and Archaeology 1. Hunters and their Prey*. Oxford: B.A.R. International Series, 163, pp. 21-30.

BYEON, W.; DOMÍNGUEZ-RODRIGO, M.; ARAMPATZIS, G.; BAQUEDANO, E.; YRAVEDRA, J.; MATÉ-GONZÁLEZ, M. A. and KOUMOUTSAKOS, P. (2019): "Automated identification and Deep classification of cut marks on bones and its paleoanthropological implications". *Journal of Computational Science,* 32, pp. 36-43. https://doi.org/10.1016/j.jocs.2019.02.005

CAPALDO SD (1997): "Experimental determinations of carcass processing by Plio-Pleistocene hominids and carnivores at FLK 22 (Zinjanthropus). Olduvai Gorge, Tanzania”, *Journal of Human Evolution,* 33(5), pp. 555–597. https://doi.org/10.1006/jhev.1997.0150

CHEN, Y.; LI, Y.; NARAYAN, R.; SUBRAMANIAN, A. and XIE, X. (2016): “Gene expression inference with deep learning”. *Bioinformatics*, 32(12), pp. 1832–1839. https://doi.org/10.1093/bioinformatics/btw074

CHOLLET, F. (2017): *Deep Learning with Python*. Manning Publications Company.

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”, *Scientific Reports*, 9(18933). https://doi.org/10.1038/s41598-019-55439-6

COURTENAY LA, YRAVEDRA J, HUGUET R, ARAMENDFI J, MATÉ-GONZÁLEZ MA, GONZÁLEZ-AGUILERA D, ARRIAZA MC (2019): “Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks*”, Paleogeo Paleoclim Paleoecol,* 522, pp. 28–39.

CUETO M, CAMARÓS E, CASTAÑOS P, ONTAÑÓN R, ARIAS P (2016): “Under the skin of a lion: unique evidence of upper Paleolithic exploitation and use of cave lion (Panthera spelaea) from the lower gallery of La Garma (Spain)”, *PLoS One*, 11(10). https://doi.org/10.1371/journal.pone.0163591

DAUJEARD C, GERAADS D, GALLOTTI R, LEFÈVRE D, MOHIB A, RAYNAL JP, HUBLIN JJ (2016): “Pleistocene hominins as a resource for carnivores: a c. 500,000-year-old human femur bearing tooth-marks in North Africa (Thomas Quarry I, Morocco)”, *PLoS One*, 11(4). https://doi.org/10.1371/journal.pone.0152284

DIRKS PHGM, BERGER LR, ROBERTS EM, KRAMERS JD, HAWKS J, RANDOLPH-QUINNEY PS, ELLIOTT M, MUSIBA CM, CHURCHILL SE, DE RUITER DJ, SCHMID P, BACKWELL LR, BELYANIN GA, BOSHOFF P, HUNTER KL, FEUERRIEGEL EM, GURTOV A, HARRISON JG, HUNTER R, KRUGER A, MORRIS H, MAKHUBELA TV, PEIXOTTO B, TUCKER S (2015): “Geological and taphonomic context for the new hominin species Homo naledi from the Dinaledi chamber. South Africa*”, Elife*, 4. https://doi.org/10.7554/eLife.09561.001

DOMÍNGUEZ-RODRIGO, M.

1996: “Testing meat-eating in early hominids: an analysis of butchery marks on defleshed carcasses”. *Congreso Internacional de Paleontología,* Orce, Granada.

1997: “Flesh availability and bone modification in carcasses consumed by lions”. *Paleogeography, Paleoclimatology & Paleoecology* (in press).

2002: “Hunting and scavenging by Early Humans: the state of the debate”. *Journal of World Prehistory*, 16 (1), pp. 1-54. https://doi.org/10.1023/A:1014507129795

2015: “Taphonomy in early African archaeological sites: questioning some bone surface modification models for inferring fossil hominin and carnivore feeding interactions”, *Journal of African Earth Sciences*, 108, pp. 42–46. https://doi.org/10.1016/j.jafrearsci.2015.04.011

DOMÍNGUEZ-RODRIGO M, BARBA R, EGELAND CP (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, BAQUEDANO E (2020): “Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications”. *Scientific Reports*, 10(18862). https://doi.org/10.1038/s41598-020-75994-7

DOMÍNGUEZ-RODRIGO M, PICKERING TR, ALMÉCIJA S ET AL (2015): “Earliest modern human-like hand bone from a new >1.84-million-year-old site at Olduvai in Tanzania”. *Nature Communications*, 6(7987). https://doi.org/10.1038/ncomms8987

DOMINGUEZ-RODRIGO M, YRAVEDRA J, ORGANISTA E, GIDNA A, FOURVEL J-B, BAQUEDANOE E (2015): “A new methodological approach to the taphonomic study of paleontological and archaeological faunal assemblages: a preliminary case study from Olduvai Gorge (Tanzania)”, *Journal* *of Archaeological Sciences,* 59, pp. 35–53. https://doi.org/10.1016/j.jas.2015.04.007

DOMÍNGUEZ-RODRIGO, M. and PIQUERAS, A. (2003): "The use of tooth pits to identify carnivore taxa in tooth-marked archacofaunas and their relevance to reconstruct hominid carcass processing behaviours". *Journal Archaeological Science,* 30, pp. 1385-1391. https://doi.org/10.1016/S0305-4403(03)00027-X

DOMÍNGUEZ-RODRIGO, M.; FERNÁNDEZ-LÓPEZ, S. y ALCALÁ L. (2011): "How can taphonomy be defined in the XXI century?". *Journal of Taphonomy,* 9 (1), pp. 1-13.

EFREMOV, I. A. (1940): "Taphonomy: a new branch of paleontology". *Pan American Geologist,* 74, pp. 81-93.

EGELAND, C. P.; DOMÍNGUEZ-RODRIGO, M.; PICKERING, T. R.; MENTER, C. G. and HEATON, J. L. (2017): "Hominin skeletal part abundances and claims of deliberate disposal of corpses in the Middle Pleistocene". *PNAS,* (in press).

EGELAND, C.; DOMÍNGUEZ-RODRIGO, M. and BARBA, R. (2007): "The hunting-versus-scavenging debate" in *Deconstructing Olduvai: A taphonomic study of the Bed I sites* (2), pp. 11-22.

FERNÁNDEZ-JALVO, Y. and ANDREWS, P. (2011): "When humans chew bones". *Journal of Human Evolution,* 60(1), pp. 117-123. https://doi.org/10.1016/j.jhevol.2010.08.003

FERNÁNDEZ-LÓPEZ S. L. (2006): "Taphonomic alteration and evolutionary taphonomy". *Journal of Taphonomy,* 4 (3), pp. 111-142.

GIDNA A, DOMÍNGUEZ-RODRIGO M, PICKERING TR (2015): "Patterns of bovid long limb bone modification created by wild and captive leopards and their relevance to the elaboration of referential frameworks for paleoanthropology", *Journal of Archaeological Sciences: Reports,* 2, pp. 302–309. https://doi.org/10.1016/j.jasrep.2015.03.003

GIDNA A, YRAVEDRA J, DOMÍNGUEZ-RODRIGO M (2013): “A cautionary note on the use of captive carnivores to model wild predator behavior: a comparison of bone modification patterns on long bones by captive and wild lions”, *Journal of Archaeological Sciences,* 40, pp. 1903–1910. https://doi.org/10.1016/j.jas.2012.11.023

GOLDEN, J.G. (2017): “Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen”, *JAMA*, 318(22), pp. 2184–2186. https://jamanetwork.com/journals/jama/article-abstract/2665757

GOODFELLOW I, BENGIO Y, COURVILLE A (2016): *Deep learning*. MIT Press.

HARRIS, J. A.; MAREAN, C. W.; OGLE, K. and THOMPSON, J. (2017): “The trajectory of bone surface modification studies in paleoanthropology and a new Bayesian solution to the identification controversy”. *Journal of Human Evolution*, 110, pp. 69-81. https://doi.org/10.1016/j.jhevol.2017.06.011

HE K, ZHANG X, REN S, SUN J (2016): “Deep residual learning for image recognition”, in: *Proceedings of the IEEE conference on computer vision and pattern.*

HIJAZI, S.; KUMAR, R. and ROWEN, C. (2015): “Using Convolutional Neural Networks for Image Recognition”. *Cadence* (in press).

HILL, A. (1976): “On carnivore and weathering damage to the bone”. *Current Anthropology,* 17, pp. 335-336.

HINTON, G. E. and SALAKHUTDINOV, R. R. (2006): “Reducing the dimensionality of data with neural networks”. *Science,* 313, pp. 504–507.

HUANG, G.; LIU, Z. and VAN DER MAATEN, L. (2018): “ Densely Connected Convolutional Networks”, *arXiv: 1608.06993v5.*

ISAAC, G. L.

1983: “Bones in contention: competing explanations for the juxtaposition of Early Pleistocene artifacts and faunal remains”. In (J. CLUTTON-BROCK and C.

GRIGSON, eds.) *Animals and Archaeology 1. Hunters and their Prey*. Oxford: B.A.R. International Series 163, pp. 3-19.

1984: "The archaeology of human origins: studies of Lower Pleistocene in East Africa, 1971-1981". *World Archaeology,* 3, pp. 1-87.

JIMÉNEZ-GARCÍA B, ABELLÁN N, BAQUEDANO E, CIFUENTES-ALCOBENDAS G, DOMÍNGUEZ-RODRIGO M (2020b): "orrigendum to: Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars", *Journal of Royal Society Interface,* 17(20200782). https://doi.org/10.1098/rsif.2020.0782

JIMÉNEZ-GARCÍA B, AZNARTE J, ABELLÁN N, BAQUEDANO E, DOMÍNGUEZRODRIGO M (2020a): "Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars". *Journal of Royal Society Interface,* 17(20200446). https://doi.org/10.1098/rsif.2020.0446

KIM, P. (2017): "Convolutional Nerual Network". In (P. KIM, ed.) *MATLAB Deep Learning,* pp. 121-147.

KINSLEY, H. and KUKIELA, D. (2020): *Neural Networks from Scracth in Python,* Sentdex. https://nnfs.io

KOUNGOULOS L, FAULKNER P, ASMUSSEN B (2018): Analysis of pit and score tooth-mark sizes from bones modified by Holocene Australian terrestrial fauna in relation to body size", *Journal of Archaeological Sciences: Reports,* 20, pp. 271–283. https://doi.org/10.1016/j.jasrep.2018.05.006

KRIZHEVSKY, A.; SUTSKEVER, I. and HINTON, G. E. (2012): "ImageNet classification with deep convolutional neural works". In (F. PEREIRA, C.J.C. BURGES, L. BOTTOU, K.Q. WEINBERGER, eds.) *Advances in Neural Information Processing Systems,* 25, Curran Associates, pp. 1097-1105.

LEAKEY REF, WALKER AC (1985): "Further hominids from the Plio-Pleistocene of Koobi Fora, Kenya". *American Journal of Physical Anthropology,* 67(2), pp. 135–163. https://doi.org/10.1002/ajpa.1330670209

LECUN, Y.; BOTTOU, L.; BENGIO, Y. and HAFFINER, P. (1998): "Gradient-based learning applied to document recognition". *Proc. IEEE*, pp. 2278-2324. https://ieeexplore.ieee.org/document/726791

LYMAN, R. L. (2010): "What taphonomy is, what it isn't, and why taphonomists should care about the difference". *Journal of Taphonomy*, 8 (1), pp. 1-16.

MOCLÁN, A.; DOMÍNGUEZ-RODRIGO, M. and YRAVEDRA, J. (2019): "Classifying agency in bone breakage: an experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using machine learning (ML) algorithms". *Archaeological and Anthropological Sciences* (in press). NJAU, J. K. and BLUMENSCHINE, R. J.

NJAU, J. K. and BLUMENSCHINE, R. J.

2005: "A diagnosis of crocodile feeding traces on larger mammal bone". *Journal of Human Evolution*, 50, pp. 142-162.

2012: "Crocodylian and mammalian carnivore feeding traces on hominid fossils from FLK 22 and FLK NN3, Plio-Pleistocene, Olduvai Gorge, Tanzania". *Journal of Human Evolution*, 63, pp. 408-417.

PANTE MC, BLUMENSCHINE RJ, CAPALDO SD, SCOTT RS (2012): "Validation of bone surface modification models for inferring fossil hominin and carnivore feeding interactions, with reapplication to FLK 22, Olduvai Gorge, Tanzania". *Journal of Human Evolution*, 63(2), pp.395–407. https://doi.org/10.1016/j.jhevol.2011.09.002

PANTE MC, NJAU JK, HENSLEY-MARSCHAND B, KEEVIL TL, MARTÍN-RAMOS C, PETERS RF, DE LA TORRE I (2018): "The carnivorous feeding behavior of early Homo at HWK EE, Bed II, Olduvai Gorge, Tanzania". *Journal of Human Evolution*, 120, pp. 215–235. https://doi.org/10.1016/j.jhevol.2017.06.005

PICKERING TR (2013): *Rough and tumble: aggression, hunting, and human evolution*. University of California Press

PICKERING TR, EGELAND CP, DOMÍNGUEZ-RODRIGO M, ET AL. (2008): "Testing the "shift in the balance of power" hypothesis at Swartkrans, South Africa: hominid cave use and subsistence behavior in the Early Pleistocene", *Journal of Anthropological Archaeology*, 27(1), pp. 30–45. https://doi.org/10.1016/j.jaa.2007.07.002

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", *Archaeological and Anthropological Sciences*, 12(4). https://doi.org/10.1007/s12520-019-00966-6

ROGERS, R.R., EBERTH, D.A. y FIORILLO, A.R. (eds.) (2007): *Bonebeds: Genesis, Analysis, and Paleobiological Significance*. Chicago: University of Chicago Press.

ROSE MD (1984): "A hominine hip bone, KNM-ER 3228, from East Lake Turkana, Kenya". *American Journal of Physical Anthropology*, 63(4), pp. 371–378. https://doi.org/10.1002/ajpa.1330630404

SAHLE, Y.; EL ZAATARI, S. and WHITE, T. D. (2017): "Hominid butchers and biting crocodiles in the African Plio-Pleistocene". *PNAS Early Edition* (in press).

SALA N, ARSUAGA JL, MARTÍNEZ I, GRACIA-TÉLLEZ A (2014): "Carnivore activity in the Sima de los Huesos (Atapuerca, Spain) hominin sample", *Quaternary Science Review*, 97, pp. 71–83. https://doi.org/10.1016/j.quascirev.2014.05.004

SALA N, ARSUAGA JL, MARTÍNEZ I, GRACIA-TÉLLEZ A (2015): "Breakage patterns in Sima de los Huesos (Atapuerca, Spain) hominin sample", *Journal of Archaeological Science*, 55, pp. 113–121. https://doi.org/10.1016/j.jas.2015.01.002

SALADIÉ, P.; HUGUET, R.; DÍEZ, C.; RODRÍGUEZ-HIDALGO, A. and CARBONELL, E. (2012): "Taphonomic modifications produced by modern Brown bears (*Ursus arctos)*". *International Journal of Osteoarchaeology* (in press).

SCHALLER GB (2009): *The Serengeti lion: a study of predator-prey relations.* University of Chicago Press.

SELVAGGIO, M. M. and WILDER, J. (2001): “Identifying the involvement of multiple carnivore taxon with archaeological bone assemblages”. *Journal Archaeological Science*, 28, pp. 465-470.

SEVILLANO, V. and AZNARTE, J.L. (2018): “Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks”. *PLoS ONE*, 13(9). https://doi.org/10.1371/journal.pone.0201807

SEVILLANO, V., HOLT K. and AZNARTE, J.L. (2020): “Precise automatic classification of 46 different pollen types with convolutional neural networks”. *PLoS ONE*, 15(6). https://doi.org/10.1371/journal.pone.0229751

SIMONYAN K, ZISSERMAN A (2014a): “Very deep convolutional networks for large-scale image recognition”, *arXiv*. arXiv:1409.1556

SIMONYAN K, ZISSERMAN A (2014b): “Two-stream convolutional networks for action recognition in videos”, in GHAHRAMANI Z, WELLING M, CORTES C ET AL (eds): *Advances in neural information processing systems*, 27. Curran Associates, Inc., pp 568–576.

SRIVASTAVA, R.K.; GREFF, K. and SCHMIDHUBER, J. (2015): “Highway networks”, *arXiv: 1505.00387.*

SZEGEDY C, IOFFE S, VANHOUCKE V, ALEMI AA (2017): “Inception-v4, inception-resnet and the impact of residual connections on learning”, in *Thirty-first AAAI conference on artificial intelligence*.

TOBIAS PV (1991): *Olduvai Gorge*, volume 4: “The skulls, endocasts and teeth of. Homo habilis 4:ç”.

WOLPERT DH (1992): “Stacked generalization”, *Neural Networks*, 5(2), pp. 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

YRAVEDRA J, GARCÍA-VARGAS E, MATÉ-GONZÁLEZ MÁ, ARAMENDI J, PALOMEQUE-GONZÁLEZ JF, VALLÉS-IRISO J, MATESANZ-VICENTE J, GONZÁLEZ-AGUILERA D, DOMÍNGUEZ-RODRIGO M (2017): "The use of micro-photogrammetry and geometric morphometrics for identifying carnivore agency in bone assemblages", *Journal of Archaeological Sciences: Reports,* 14, pp. 106–115. https://doi.org/10.1016/j.jasrep.2017.05.043

YRAVEDRA J, MATÉ-GONZÁLEZ MA, COURTENAY LA, GONZÑÁLEZ-AGUYILERA D, FERNÁNDEZ-FERNÁNDEZ M (2019): "The use of canid tooth marks on bone for the identification of livestock predation", *Scientific Reports,* 9(16301). https://doi.org/10.1038/s41598-019-52807-0

YRAVEDRA, J.; LAGOS, L. and BÁRCENA F. (2011): "A taphonomic study of wild Wolf (*Canis lupus)* modification of horse bones in Northwestern Spain". *Journal of Taphonomy,* 9 (1), pp. 37-65.

ZOPH B, VASUDEVAN V, SHLENS J, LE QV (2018): "Learning transferable architectures for scalable image recognition", in *Proceedings of the IEEE conference on computer vision and pattern.*