Archaeological and Anthropological Sciences (2021) 13: 167
https://doi.org/10.1007/s12520-021-01424-y

ORIGINAL PAPER

More than meets the eye: use of computer vision algorithms to identify stone tool material through the analysis of cut mark micro-morphology

Gabriel Cifuentes-Alcobendas[1,2] · Manuel Domínguez-Rodrigo[1,2]

Received: 1 June 2021 / Accepted: 1 August 2021 / Published online: 20 September 2021

Abstract

Artificial intelligence algorithms have recently been applied to taphonomic questions with great success, outperforming previous methods of bone surface modification (BSM) identification. Following these new developments, here we try different deep learning model architectures, optimizers and activation functions to assess if it is possible to identify a stone tool's raw material simply by looking at the cut marks that it created on bone. The deep learning models correctly discerned between flint, sandstone and quartzite with accuracy rates as high as 78%. Also, single models seem to work better than ensemble ones, and there is no optimal combination of hyperparameters that perform better in every possible scenario. Model fine-tuning is thus advised as a protocol. These results consolidate the potential of deep learning methods to make classifications out of BSM's microscopic features with a higher degree of confidence and more objectively than alternative taphonomic procedures.

Keywords Artificial intelligence · Computer vision · Raw material · Cut marks · Stone tools · Palaeolithic

Introduction

Bone surface modification (BSM) studies have long been at the core of the taphonomic discipline, since such traces represent some of the most direct evidence for interaction between hominins and their environment. Of all BSM, butchering cut marks found on fossilized bones have received the most attention because they provide evidence for meat processing using stone or metal tools. The discovery of early butchering marks on animal bones has been claimed to be the first step in the cognitive evolution of the human lineage (Bello and Soligo 2008). The archaeological record may not always provide all the inferential steps necessary to go from the traces themselves to the understanding of the contextual and behavioural processes that generated these traces. Following Gifford-Gonzalez's (1991) terms, the effector may not always be present, so the traces (butchering marks) are often the only hint that can be used to infer agency, causality, and behavioural and ecological processes. In addition, the Palaeolithic archaeological record is limited in nature and is primarily composed of bones and stones (Domínguez-Rodrigo 2012). When one of these two types of components is missing, interpretations of site formation are even more challenging.

Among BSM, cut marks have been especially targeted by taphonomic research. Several studies have tried to identify the type of stone tools that generated marks through both quantitative and qualitative analyses of mark cross-section (e.g. Bello 2010; Bello et al. 2009; Bonney 2014; Galán and Domínguez-Rodrigo 2014; Merritt 2012) or mark micromorphology (e.g. Courtenay et al. 2017; Greenfield 1999, 2006; Val et al. 2017). In contrast, studies addressing the impact of stone tool's raw material on mark creation have been more limited in number.

The earliest works addressing the impact of tool raw materials on cut mark morphology focused on discerning between stone and metal tools (Walker and Long 1977; Olsen 1988; Von Lettow-Vorbeck 1998; Greenfield 1999). This has a rather limited application to the prehistoric record. Walker and Long's work (1977), however, considered different types of raw materials (i.e. chert and obsidian) in their

✉ Gabriel Cifuentes-Alcobendas
gabriel.cifuentes@uah.es

✉ Manuel Domínguez-Rodrigo

[1] Institute of Evolution in Africa (IDEA), Alcalá University, Covarrubias 36, 28010 Madrid, Spain

[2] Area of Prehistory (Department History and Philosophy), University of Alcalá, 28801 Alcalá de Henares, Spain

study, although their main interpretation was still focused on tool type more than raw material variability. More recent works tried to make these studies applicable to the pre-metallic periods of prehistory by differentiating between different types of rocks through the study of the cut marks alone. Greenfield (2006) tried to differentiate between obsidian, flint and quartzite with little success, being the width of the marks the only observable difference between different materials. Greenfield stated that “it is almost impossible to distinguish raw material purely on the basis of cut marks. There are too many variables” (Greenfield 2006, p. 155). Other studies have tried to approach this question by using geometric morphometrics (GMM). Maté González and his team were the first to develop a standardized methodology to apply GMM to cut marks (Maté González et al. 2015; Maté-González et al. 2016). In their work (Maté-González et al. 2016), they compared three raw materials and got an accuracy of 81.5%, 61.6% and 53% for flint, quartzite and metal respectively in the classification of their experimental cut marks. Courtenay et al. (2017) and Yravedra et al. (2017) expanded on this work by adding basalt to the study and managed to get an average of 60% correct identification for the different raw materials. Courtenay et al. (2017) report that the method works best in pairwise comparisons, where they achieve a 72% and 68% true allocation for basalt and metal, respectively.

When considering other types of raw materials, such as shells or bamboo, the differences are more marked, as would be expected considering the great difference between these two materials and stone. For cutting tools made of reworked shells, Choi and Driwantoro (2007) claimed to have defined qualitative criteria capable of discerning those tools from others made in stone, bamboo, bone and coconut. They used scanning electron microscope (SEM) images to look at the inner features of the marks. Bamboo is another material that has received some attention given its plausible use as a tool by Asian hominins. West and Louys (2007) compared CM imparted by flint and bamboo tools through the use of SEM images. They found that the main difference between stone and bamboo cut marks is that bamboo tends to create an asymmetric V-shaped groove, in which only one of the walls has microstriations (West and Louys 2007). While these results are coherent with the morphology of bamboo tools, stone tools may as well produce asymmetric grooves (Walker and Long 1977), and several taphonomic processes (e.g. abrasion) may delete microstriae on one or both walls of the marks (Behrensmeyer et al. 1986; Pizarro-Monzo and Domínguez-Rodrigo 2020). Therefore, these variables must be carefully considered when used in archaeological scenarios.

Being able to identify the materials of the tools that were used when processing a carcass can be very informative, since it can be efficiently used to link archaeofauna and a portion of the lithic assemblage. This information creates new opportunities to better understand past hominin behaviours. For example, being able to confidently identify raw material from the analysis of cut marks can be used to determine if hominins were using specific tools/materials for specific meat processing activities (i.e. skinning, defleshing, disarticulation). A practical example of this was provided by Yravedra et al. (2017), when they analyzed cut marks from the 1.3 ma site of BK (Olduvai Gorge) and found out that most cut marks were made with quartzite flakes and not basalt implements. Given the documented mix of raw materials in the stone tools of Oldowan and Acheulian hominins, it would be important to determine if butchery was preferentially done with one type of raw material/implement or another. For this reason, here we present a study aimed at applying computer vision algorithms to an experimental reference collection of butchery cut marks made with retouched flakes with the goal of assessing if cut mark microscopic features vary by raw material. The results of this study could open the door to the application of this method to archaeological BSM, with site-specific experimental modelling using the same types of raw materials represented at archaeological sites. The study was made targeting differentiation of raw materials in Palaeolithic contexts, where the stone tools are mostly flaked and detached artefacts. It also focused on slicing cut marks, as opposed to chopping or scraping marks also potentially generated through stone tool butchery. Therefore, when referring to lithic tools in our experiment, we will focus on defleshing detached stone tools, instead of chopping or bashing artefacts.

Materials and methods

The deep learning (DL) approach

Convolutional neural networks (CNN) are a type of DL model that try to imitate the mechanisms of a human brain by passing graphic information (in the form of pixels) through discriminant nodes (neurons) to produce responses in the form of a classificatory output (Schmidhuber 2015; Goodfellow et al. 2016). This is done through trial and error by backpropagating weight adjustment according to loss after each training epoch. It is not our intention to explain the basic functioning of CNNs, which has been abundantly explained in the literature (e.g. Adrian 2017; Ballard 2018; Chollet 2017), but to present how these algorithms can be applied to archaeological questions. DL and computer vision algorithms can be used to analyze graphic data in the form of images. The CNNs analyze the pixel data contained in the images by extracting recurrent features and pixel structures that may be useful to classify an image to any given group. These algorithms benefit from large amounts of data when computing, and thus, large samples as well as several repetitions over the training sample (epochs) are necessary.

When applying the CNNs to different questions, it is mandatory to try different models since it is not common that the same algorithm will be the best option for every case scenario. This is known as the "No Free Lunch Theorem" (Wolpert 1996; Wolpert and Macready 1997), which advocates for the use of multiple algorithms when addressing any given question. CNN algorithms can be fine-tuned to perform as best as possible by tweaking the hyperparameters that rule how the algorithm will be computed. Of all the available hyperparameters (i.e. learning rate, number of epochs, batch size, activation function, optimizer dropouts, etc.), the activation function and the optimizer have the most weight when fine-tuning the model for performance (Domínguez-Rodrigo et al. 2021).

Normally, CNNs require hundreds or thousands of images to prove competent at any given classification test. This is because the feature extraction ability of the algorithm improves when trained with a larger sample of images. Having these vast collections of images is something that is rarely possible in taphonomic research, even when carrying out experiments aiming to produce large amounts of marks. The intrinsic variability of the experimental context and the difficulty to adequately replicate specific types of BSM precludes obtaining these types of samples by normal means. To compensate for this, transfer learning can be used to re-train models that already have some training in feature extraction. Models in the ImageNet competition are trained upon thousands of images belonging to a wide range of categories (animals, objects, landscapes, etc.), and are publicly accessible. These models keep the variables that they use to extract the features of the images and to classify them (weights). Then, they can be re-programmed to keep those weights and use them to identify features for a different classification problem. By doing so, we can use the training they already have in feature recognition, and apply it directly to our classification problem, without the need for the algorithm to be trained to "see" features from scratch, which requires more time and larger samples. In our tests, transfer learning models outperformed models trained from scratch in every scenario because they were exposed to a larger dataset thanks to the weight transfer process mentioned above (Domínguez-Rodrigo et al. 2020; Jiménez-García et al. 2020).

Finally, the last thing to consider when working with small sample sizes (i.e. hundreds of images) is the necessity to avoid overfitting in the model. Training the model with a small sample will likely reduce the variability that the model associates with the different classes. This can impact the model's ability to correctly classify new images into the classes that it has learnt, in case these new images show any variability to which the model has not been subjected before. To account for this variability, augmentation techniques can be used to increase image variation and therefore improve the model's ability to identify the key features in widely varying scenarios. These procedures rely on applying modifications to images randomly (i.e. rotation, crop, skew, horizontal/vertical displacement, horizontal flip, etc.) to widen the instances in which key features can be identified. In our tests, all the typical augmentation procedures were applied except rotation, which proved to be unnecessary to prevent overfitting, and reduced the accuracy results by approximately 10%.

To maintain control on how each model was performing, we used accuracy and loss metrics, and the F-1 scores. Accuracy and loss contrast the algorithm's classification of the images against the expected one determined through the group labels. These two metrics are used by the model to guide weight alteration in the different iterations, and are a good way to live-control model performance. F-1 scores, on the other hand, compare the classification results for each of the classes in a confusion matrix and result from averaging the differential classification performance per group. This is helpful to determine if the model performs evenly on all classes, or if it obtains good results in some while unsuccessful in others. With this method, model consistency and balanced accuracy can be controlled and accounted for.

The models used in this study were trained using greyscale images of the slice marks at 80 × 400 resolution in.bmp format. We also tried using resolutions of 224 × 224 (ResNet50's native resolution), and 64 × 64 (for reduced computation), but the results were lower in accuracy when compared to those obtained using a 80 × 400 resolution. The images were captured using a binocular microscope (Optika) at × 30 magnification, and an external 3 Mpx camera (OptiCam B3). The database was composed of 574 images, with a balanced distribution between raw materials (200, 182 and 192 images for sandstone, quartzite and flint, respectively). The training phase examined 70% of images for each class, while the remaining 30% of images not used for training were used for testing the model's performance. Models were trained on batches of 32 images and updating the weights with backpropagation for 100 epochs each. The coding and training of the models were done in Tensorflow (v. 2.3.0) and Keras (v. 2.4.3), in a conda environment capable of CUDA computing with cuDNN (v. 11.2) (Chetlur et al. 2014).

Model architectures used

The model architectures used in this study are some of the best performing models in the ImageNet competition, trained on more than 1,000,000 images for the 1000-image category ILSVRC competition. This makes the use of transfer learning more efficient since these models have been trained in feature extraction in widely different classes with large amounts of data. The architectures selected for this study are ResNet50, VGG16, InceptionV3 and DenseNet (Fig. 1). These model architectures have been deeply covered elsewhere for its use in similar identification tasks, so for more information on these please refer to Domínguez-Rodrigo et al. (2020) (ResNet50, VGG16 and InceptionV3) and Abellán et al. (2021) (DenseNet). Other models such as NasNet Large and Inception-resnet were tested, but their preliminary results were not as good as with the other so we did not use them any further.

Hyperparameters used

The two main hyperparameters that we altered to fine-tune the models were activation functions and optimizers. Activation functions are parameters that define whether or not a specific neuron will be activated by the input (i.e. pixel value). If the neuron is activated by the input, it will pass the information to the next layer, and if not, that specific information will be discarded and not used for that iteration. Of all the different activation functions available, we selected because of their optimal performance: ReLU, Swish and Mish. ReLU (rectified linear activation) has become the main activation function because of its good overall performance (Chollet 2017; Ballard 2018). Swish is a recent activation function developed by Google's Brain team (Ramachandran et al. Ramachandran et al. 2017a, b) since it is supposed to work better than ReLU on deeper architectures (> 40 layers). Mish is also a recent function that has been claimed to outperform ReLU and Swish while improving overfitting during training (Misra 2019). For further information on the mathematical construction of these functions, refer to their original papers, or to the summary found on Domínguez-Rodrigo et al. (2021).

On the other hand, optimizers are algorithms used to control attributes of the CNN, such as the learning rate, in order to reduce the loss and get to the point of convergence (maximum accuracy) as fast as possible. Here we used two different optimizers (SGD and Adam) that proved to be the best in our preliminary tests. SGD (Stochastic Gradient Descent) is one of the most commonly used optimizers for its versatility and the ability to adjust learning rate and momentum to control how the model will adjust its weight during training. Adam is an optimizer based upon the AdaGrad algorithm that scales the learning rate and is supposed to work better with sparse gradients and large numbers of parameters (Kingma and Ba 2015). Further information on how optimizers work and which are available can be found at Keras Documentation webpage (https://keras.io/api/optimizers/) and in Domínguez-Rodrigo et al. (2021).

Ensemble learning

Ensemble learning is based on combining the weights of different models at the same time to improve classification. The rationale behind this is that different models may identify image features differently, and therefore, if we are able to merge all of them into one single model, this merged model should be the best performing since it results from combining all models´ predictions.

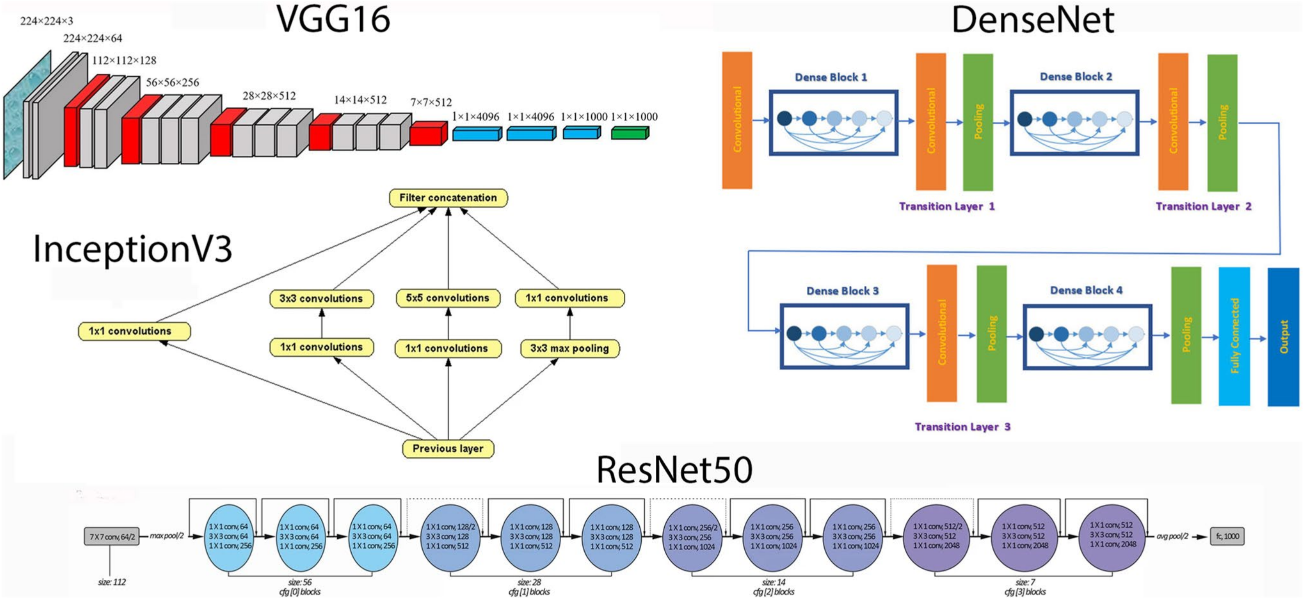

Fig. 1 Architectures of the four transfer learning models used in the study. Image of VGG16 is by Nshafiei and is licensed under CC BY-SA 4.0. Image of DenseNet is by Attallah (2021) licensed under CC BY 4.0

To account for ensemble model performance, we carried out two analyses using models of the four architectures aforementioned (i.e. Resnet50; VGG16; InceptionV3; DenseNet). In the first analysis, we used the best performing model for each architecture, while in the second analysis, we compared only the most similar models in variance (the algorithms used for each test and their results can be found in Tables 3 and 4). To select which models had the most similar variance, a principal component analysis (PCA) was conducted using the weighted accuracy and the F-1 scores per class of all the different models (Fig. 2). The PCA and its plot were done in R (v.3.4.4).

The experimental sample

For this experiment, we used a reference collection of 702 cut marks imparted on modern fresh bone still bearing some meat (Fig. 3). This reference collection was created for a previous experiment, and more information about it can be found in the original publication (Cifuentes-Alcobendas and Domínguez-Rodrigo 2019). To keep variability under control, we only used one type of stone tools to be able to confidently associate cut mark micro-morphology variation only to raw materials. We chose retouched flakes because they expand the diversity of shape and sizes of resulting cut marks. Since we targeted to provide a solid referential framework for categorizing cut marks, we opted for these instead of simple flakes, which produce a more limited range of variation. We standardized the size of each tool and carried out the retouch on one edge following the same protocol in every flake (Fig. 4). The stone raw materials used here were flint, sandstone and quartzite. To ensure that the edge of the tool underwent no attrition that could alter

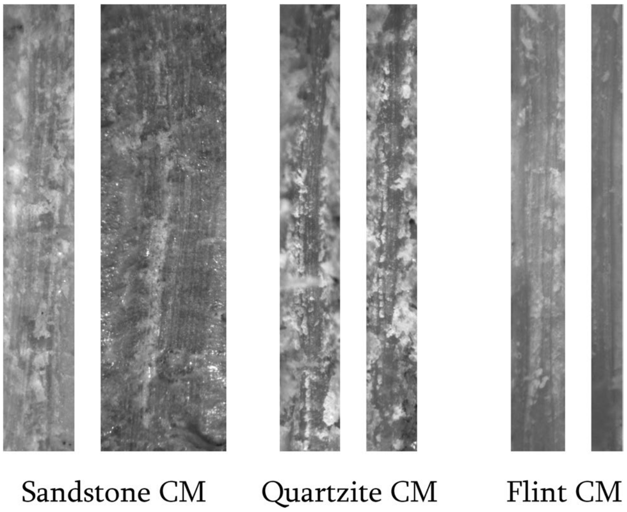

Fig. 3 Examples of the CM images used to train the models

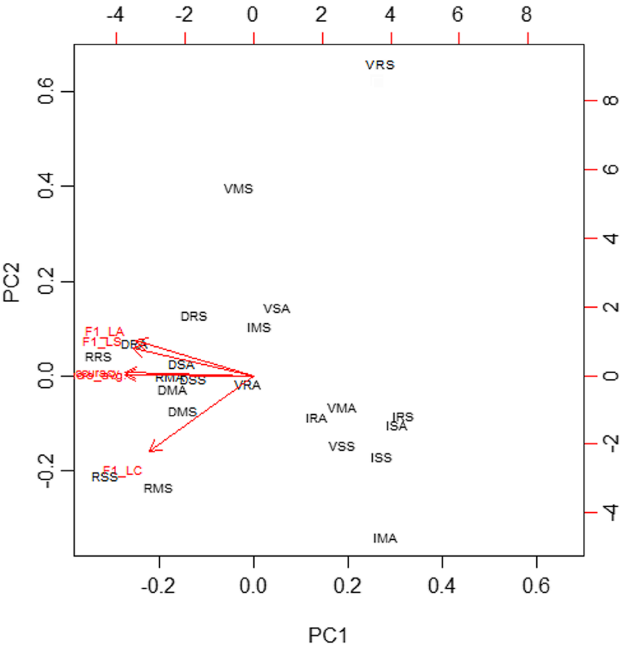

Fig. 2 PCA plot showing the variance in model performance according to weighted accuracy and the F-1 scores per class of all the different models. The initials in the plot correspond to (Model) + (Activation function) + (Optimizer). Therefore, DRS stands for (DenseNet) + (ReLU) + (SGD). PC1 explains 87% of the variance while PC2 just accounts for 9%

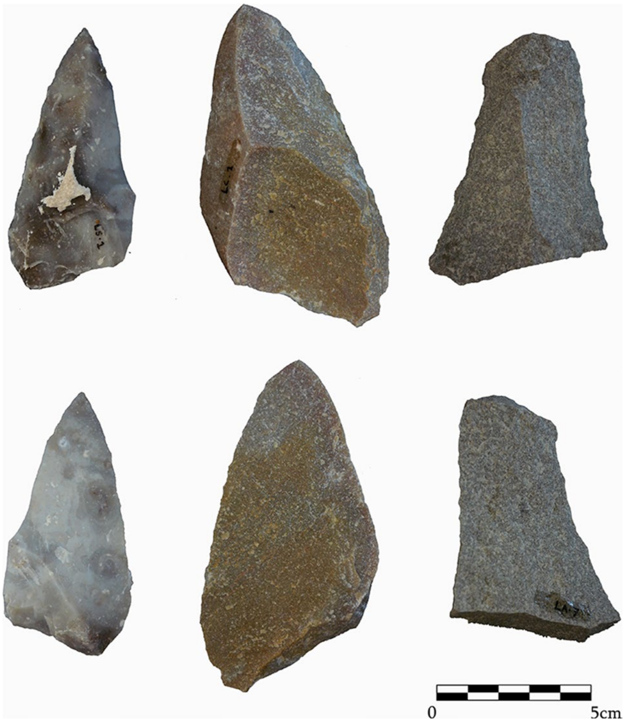

Fig. 4 Examples of the retouched flakes used to create the marks. The upper row corresponds to the dorsal view, while the lower row is the ventral views of the tools

the micro-features of the marks, we limited the number of uses of each tool to 20 strokes. This is further supported by other studies that concluded that tool attrition is not linked to impacts on bone (CM), but to skinning and disarticulation processes instead (Braun et al. 2008). Further information on the stone tool collection can be found in the Supplementary Information.

Results

Table 1 presents the results for every model out of the 24 model iterations run for this study using the three activation functions (i.e. ReLU, Swish and Mish) and both optimizers (SGD and Adam). The best performing model was ResNet50, using SGD as optimizer and ReLU as activation function. The accuracy for this model reached 78%, while the loss was 0.56. ResNet50 with SGD and swish activation function scored second with 78% and 0.58 in accuracy and loss, respectively. After ResNet 50, DenseNet was the second most successful architecture achieving 75% accuracy and 0.40 loss with Adam and ReLU as hyperparameters (Fig. 5). It is noteworthy that all the models using DenseNet's architecture achieved results above 70% of correct classification while maintaining the lowest loss values among all the models (< 0.48).

When the F-1 scores and inter-class classification were examined (Table 2), some patterns became apparent. Marks created with flint flakes display the highest accuracy most of the time (reaching 81%), while marks imparted with sandstone flakes tend to yield the lowest accuracy scores (57% for the lowest scoring model). This means that flint-produced BSM tends to create distinctive patterns in the inner groove of the marks that allow the computer vision algorithms to confidently classify them. Meanwhile, sandstone and quartzite marks tend to have more balanced classification rates, which is coherent with the fact that both raw materials share a similar grain size in their composition (see Cifuentes-Alcobendas and Domínguez-Rodrigo 2019).

The results of the ensemble learning models (Tables 3 and 4) demonstrated that this approach is not effective for this specific question. Not only did the models not improve in their results, but the general results were worse both in total accuracy (Table 3B) and inter-class classification rates (Table 4). To test the efficiency of ensemble learning, we used four different algorithms (i.e. Logistic regression, Random Forest, Extra Trees and Gradient Boosting Classifier). These ensemble learning algorithms were used on two sets

Table 1 Accuracy, loss and F-1 scores of all the models run for the study. For accuracy, higher is better while for loss, lower is better. The F-1 score must be similar to the accuracy value to prove that accuracy is balanced between classes

Model architecture	Activation function	Optimizer	Accuracy (%)	Loss	F-1 score
Resnet50	ReLU	SGD	78	0.56	0.77
	Swish	SGD	78	0.58	0.77
	Mish	SGD	74	0.74	0.75
	ReLU	Adam	68	1.02	0.66
	Swish	Adam	72	0.86	0.72
	Mish	Adam	76	0.74	0.74
VGG16	ReLU	SGD	70	0.48	0.71
	Swish	SGD	64	0.53	0.66
	Mish	SGD	63	0.58	0.65
	ReLU	Adam	68	0.56	0.71
	Swish	Adam	67	0.56	0.69
	Mish	Adam	65	0.58	0.66
InceptionV3	ReLU	SGD	65	0.76	0.63
	Swish	SGD	64	0.83	0.64
	Mish	SGD	69	0.79	0.70
	ReLU	Adam	67	0.87	0.68
	Swish	Adam	64	0.88	0.63
	Mish	Adam	65	0.84	0.64
DenseNet	ReLU	SGD	71	0.39	0.73
	Swish	SGD	73	0.40	0.73
	Mish	SGD	72	0.45	0.74
	ReLU	Adam	75	0.40	0.76
	Swish	Adam	74	0.48	0.73
	Mish	Adam	74	0.48	0.74

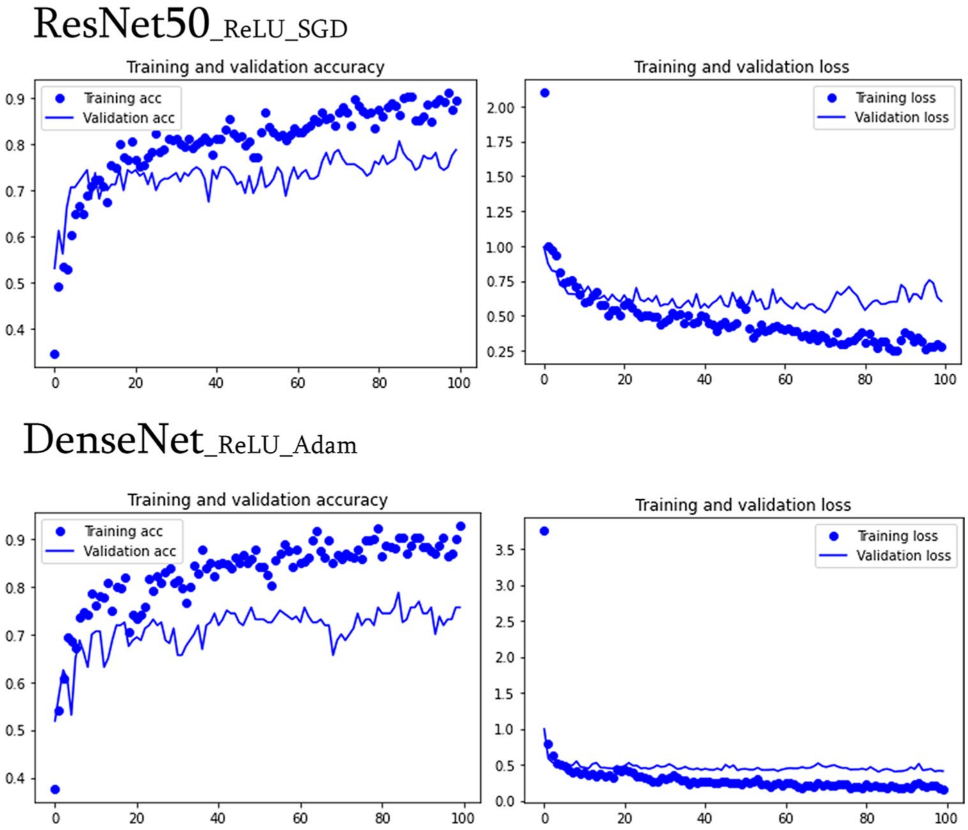

Fig. 5 Accuracy and loss graphics for the best scoring ResNet50 and DenseNet models. Graphics contain both the training and validation accuracies and losses obtained during the 100 epochs of training. These can be used to assess that the models are not overfitting. It is interesting to note that DenseNet tends to generate more overall stable models, even if the best absolute result is obtained with ResNet's architecture

of models. The first one (variance model) was composed of the models that share a similar variance in both the general accuracy and F-1 scores (Fig. 2) for the different architectures, while the second group (accuracy model) corresponds to the best scoring models. Despite this, we were able to use these results to demonstrate that ensemble learning methods work better when used with models that share similar general accuracy and F-1 scores, instead of just using the best scoring ones (see Table 3). This is coherent with the fact that models that share similar results are probably using the same type of features to make classifications, while the other models underperform when merging very different types of classification features, even if they have better overall accuracies. Therefore, a reduced variance among models should be prioritized instead of just using the best performers when performing ensemble learning.

Discussion

The results presented here show that DL algorithms are not only capable of identifying raw material types through their impact on cut mark micro-morphology, but they provide overall better results with a greater degree of confidence than GMM, involving low- and high-magnification approaches. Also, unlike these methods, DL and computer vision algorithms provide an almost fully objective and replicable method since human input stops at image taking. All the processing done after the images are fed to the model is automated and fully replicable if the model's parameters and hyperparameters are kept the same. Furthermore, the algorithm's ability to "learn" which features are the important ones to differentiate between classes, and how to identify those, allows for this "knowledge" (weights) to be saved and later applied to new cases without any alteration. In this sense, the ever present inter-analyst variability and researcher's inherent subjectivity are reduced to a minimum part of an actual BSM study (i.e. taking the images and selecting the hyperparameters). However, caution is still necessary, since subjectivity is still present in other parts of the process and can still affect the results of the study. Thus, researchers still need to be thoughtful about the experimental context when creating the reference collections necessary to train the algorithms. Also, the protocol for taking images must adapt to the specificities of the problem. For example, if the objective is to differentiate between classes based on the inner micro-features of marks, low to medium magnification and an oblique light can help to accentuate these inner micro-morphologies. If the mark's general shape or other features are going to be used for discrimination, the method should vary accordingly. In addition, the process to take images must be applied systematically without variation

Table 2 Inter-class accuracy values for the three raw materials

Model architecture	Activation function	Optimizer	Raw material accuracy (%)		
			Sandstone	Quartzite	Flint
Resnet50	ReLU	SGD	74	77	81
	Swish	SGD	70	80	82
	Mish	SGD	67	78	78
	ReLU	Adam	62	74	79
	Swish	Adam	66	69	62
	Mish	Adam	71	75	76
VGG16	ReLU	SGD	69	67	76
	Swish	SGD	58	70	70
	Mish	SGD	62	58	72
	ReLU	Adam	65	72	74
	Swish	Adam	66	69	72
	Mish	Adam	59	69	70
InceptionV3	ReLU	SGD	57	67	65
	Swish	SGD	58	69	65
	Mish	SGD	66	70	73
	ReLU	Adam	59	70	72
	Swish	Adam	61	68	61
	Mish	Adam	57	71	63
DenseNet	ReLU	SGD	67	72	80
	Swish	SGD	68	74	77
	Mish	SGD	65	75	81
	ReLU	Adam	70	75	82
	Swish	Adam	68	74	79
	Mish	Adam	68	75	79

since doing otherwise could compromise the outcome. Finally, it is still the researcher's responsibility to choose the model's hyperparameters, and this could be considered subjective. Given the variability in the performance of different hyperparameters (especially activation functions and optimizers) when applied to different problems, we suggest adhering to Wolpert's (1996) "No Free Lunch Theorem" and testing out all the possible alternatives when doing hyperparameter optimization. This simple idea has proven effective to achieve the best results possible while keeping subjectivity under control in this study.

The analytical method used in the present work is based on clearly defined model architectures, which are standard if using transfer learning and, therefore, should be equally accessible to anybody with basic training in deep learning. Given that transfer learning enables the use of pre-trained architectures, the final training of the models does not require hard computation and can be carried out with low-cost workstations. Definition of the models also allows replication fairly easily. Creation of BSM libraries, like the one used here, will also contribute to enhancing dissemination of the method and libraries through public repository access.

The results of this study provide useful insights in how a stone tool's raw material can impact CM micro-morphology. In most of the models, flint was the material that produced higher correct classification rates, with 10% better accuracy when compared against quartzite and sandstone. Furthermore, these two later raw materials share similar accuracy scores in the tests. Upon closer inspection of the tools used, we concluded that these differences can be explained by the grain size of the raw materials. Flint has a micro-crystalline structure that creates a smoother surface on the tools when flaked to create an edge. This allows the tool to have thinner and sharper edges that logically create thinner and deeper grooves (Walker and Long 1977; Greenfield 2006; Maté-González et al. 2016; Yravedra et al. 2017), as well as distinctive micro-morphologies (i.e. microstriations) that allow the resulting slicing marks to stand out from the others and to be more confidently classified. On the other hand,

Table 3 Tables containing the models used for both ensemble learning models (A) and the accuracy results of both models using different ensemble learning algorithms (B). In A, the variance model corresponds to the models selected because of their similar variance in the PCA (Fig. 2), and the accuracy model corresponds to the best scoring models. In B, the results (%) correspond to the accuracy of both models (indicated) when using each of the four different ensemble algorithms. In green, the ensemble algorithm used to see inter-class accuracy rates (shown in Table 4)

A		
Variance model		Accuracy model
Models		Models
Resnet50_Mish_Adam		Resnet50_ReLU_SGD
DenseNet_ReLU_SGD		DenseNet_ReLU_Adam
VGG16_Mish_SGD		VGG16_ReLU_Adam
InceptionV3_Mish_SGD		InceptionV3_Mish_SGD
B		
Variance model		Accuracy model
Ensemble algorithm	Results (%)	Ensemble algorithm
Logistic regression	63–52	Logistic regression
Random Forest	65–67	Random Forest
Extra Trees	68–65	Extra Trees
Gradient Boosting Classifier	63–64	Gradient Boosting Classifier

Table 4 Inter-class accuracy rates for each of the three raw materials, and the two ensemble models constructed (including the ensemble algorithm used)

Model	Ensemble algorithm	Raw material accuracy (%)		
		Sandstone	Quartzite	Flint
Variance model	Extra Trees	67	70	68
Accuracy model	Random Forest	58	72	70

our sandstone and quartzite tools have similar grain sizes, which fits well with the more balanced accuracy rates these two materials obtained. Furthermore, similar grain sizes are likely to produce similar micro-morphologies. This helps to explain why these two materials are several points below in classification rates from flint: because these two share similar microstriation patterning due to grain size and lack defining the features of a smoother and thinner blade such as a flint made one. In addition, a closer inspection of the marks themselves also revealed that flaking is more common in coarser grain sizes (i.e. quartzite and sandstone). This, alongside microstriatiae patterning, seems to be driving the algorithm's decisions. These interpretations are not new, since rock hardness and grain size are known to affect CM's morphology (Braun et al. 2016), but this is the first time that these differences have been objectively quantified and proven reliable for sustained raw material identifications.

Another unexpected outcome of this study derives from the ensemble learning techniques. Ensemble learning methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone (Rokach 2010). However, this study demonstrated otherwise, since the ensemble models were not able to outperform the individual models that they learnt from. Since inter-model variability was specifically avoided, this cannot explain why multiple models fail more than simple ones. Given that ensemble learning expect that the different models will add new features to make classification, this could mean there is little to no variation in the features that all four models used. Thus, the averaging of all of them does not enhance the accuracy. Besides, we cannot rule out the possibility that ensemble learning is just not useful for the problem at hand. This is most likely because the sample's variability is not as large as in image classification competitions, where images of vastly different objects are used.

Yravedra et al. (2017) used a geometric morphometric approach to the use of raw materials at the early Pleistocene site of BK (Olduvai Gorge, Tanzania), based on the analysis of a sample of cut marks on bones. They succeeded in attributing most butchery to the use of quartzite, instead of basalt. This is supported by the overwhelming predominance of quartzite lithic artefacts at the site compared to basalt. In archaeological assemblages where there is a diversity of raw materials represented, the experimental combination of raw material-tool type should produce an adequate referential framework within which differences in cut mark micro-morphology could be approached with the goal of detecting which tools/raw materials were used for butchery activities.

When compared to previous works, this method shows significant advantages both in simplicity and raw accuracy. First of all, since most of the computing for the classification is done automatically, a lot of the subjective decisions that had to be done in previous methods are avoided (i.e. choosing where to put a landmark in GMM, selecting and quantifying which micro-morphological features should be used to make classifications, etc.). This allows this method to be used by different researchers without inter-analyst bias playing a major role. This is also intimately related to replicability, since reducing the subjectivity bias to a minimum has a direct impact in the method being replicable, and thus scientific. Finally, even if training the algorithms require a basic knowledge on how CNNs work, once these models are trained the workflow becomes much simpler. To classify any given mark (in the form of an image), it is only necessary to load the image in the model, and run two lines of code to make a prediction and obtain the classification label for that mark. This two-button workflow is much simpler and faster than using any of the GMM, qualitative or quantitative methods previously described. Because of this, we believe that this method has the potential to be used, even by researchers who are not familiar with DL.

The results of this study show great potential to study BSM through deep learning algorithms. However, caution should still be advised when applying these methods to the archaeological record. To this moment, the experimental reference collections used to train these models are pristine in nature, that is, they have been analyzed without having suffered from any post-depositional processes. The objective of this work is to showcase the ability of these methods to accurately discern between different materials in stone tools, and this has been demonstrated so far. However, the archaeological application of these algorithms, as they are right now, is still limited to bones with good preservation where no post-depositional processes have altered the cortical surface of the bones. For this method to be fully applicable to the archaeological record, a reference collection of BSM affected by biostratinomic and diagenetic processes should be added to the training.

Conclusions

DL and computer vision algorithms outperform previous methods of identification of chipped stone tool raw material through the analysis of the micro-morphology of the resulting cut marks imprinted on bone surfaces during butchery. Unlike previous methods, a DL approach allows classification to be backed with a degree of confidence that can also be used to further support any classification conducted by the algorithm. Furthermore, since all the computing and training takes place with little human involvement, the results of the models can easily be replicated. In fact, the capability of these models to be used in different contexts without inter-analyst variance playing a large role makes them the first systematically objective method to be used in taphonomic analysis of butchery slicing marks. Simple models seem to work better than complex ensemble models for this specific question. However, further experimentation is needed to assess if this remains valid when larger and more variable samples are taken into account. This work is merely a first step, and one that needs to be followed up by more experimental work that expands the control dataset. In this sense, more raw materials need to be considered (e.g. bamboo and shells), as well as different tool types (e.g. simple flakes and handaxes), and different biostratinomic and diagenetic processes. This work further strengthens the role that artificial intelligence can play, not only in taphonomic and prehistoric studies, but in general archaeological research. The ability to accurately identify the traces of different processes through computer vision could be used to address different manufacturing processes (i.e. ceramics, pigments, bone and antler crafting, etc.) and other traces that are present in the archaeological record from the Palaeolithic to our times. It is a method capable of approaching long-debated questions with greater certainty, as well as opening new areas of study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12520-021-01424-y.

Acknowledgements We want to thank the two reviewers of this manuscript for their useful insights in improving it to its final form. GCA also wants to thank Aimée Little for the help provided reviewing and imporving an earlier version of this paper.

Author contribution GCA created the experimental sample; GCA and MDR developed the python environment for CNNs; GCA and MDR carried out the analysis; GCA and MDR wrote and reviewed the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This study was funded by the Spanish Ministry of Economy and Competitiveness through the project (HAR2017-82463-C4-1-P) and the Ministry of Culture through their Archaeology Abroad program. Financial support has also been obtained from the Palarq Foundation and ESIN2.

Data availability The images used to train the CNN models can be openly accessible at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FYHKWMR.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abellán N, Jiménez-García B, Aznarte J et al (2021) Deep learning classification of tooth scores made by different carnivores: achieving high accuracy when comparing African carnivore taxa and testing the hominin shift in the balance of power. Archaeol Anthropol Sci 13. https://doi.org/10.1007/s12520-021-01273-9

Adrian R (2017) Deep learning for computer vision with python - starter bundle. PyImageSearch. https://www.pyimagesearch.com/deep-learning-computer-vision-python-book/. Accessed 16 Apr 2021

Attallah O (2021) MB-AI-His: histopathological diagnosis of pediatric medulloblastoma and its subtypes via AI. Diagnostics 11:359. https://doi.org/10.3390/diagnostics11020359

Ballard W (2018) Hands-on deep learning for images with TensorFlow: build intelligent computer vision applications using TensorFlow and Keras. Packt, Mumbai

Behrensmeyer AK, Gordon KD, Yanagi GT (1986) Trampling as a cause of bone surface damage and pseudo-cutmarks. Nature 319:768–771

Bello SM (2010) New results from the examination of cut-marks using three-dimensional imaging. In: Ashton NM, Lewis SG, Stringer CB (eds) The ancient human occupation of Britain. Elsevier B.V, London, pp 249–262

Bello SM, Soligo C (2008) A new method for the quantitative analysis of cutmark micromorphology. J Archaeol Sci 35:1542–1552. https://doi.org/10.1016/j.jas.2007.10.018

Bello SM, Parfitt SA, Stringer C (2009) Quantitative micromorphological analyses of cut marks produced by ancient and modern handaxes. J Archaeol Sci 36:1869–1880. https://doi.org/10.1016/j.jas.2009.04.014

Bonney H (2014) An investigation of the use of discriminant analysis for the classification of blade edge type from cut marks made by metal and bamboo blades. Am J Phys Anthropol 154:575–584. https://doi.org/10.1002/ajpa.22558

Braun DR, Pobiner BL, Thompson JC (2008) An experimental investigation of cut mark production and stone tool attrition. J Archaeol Sci 35:1216–1223. https://doi.org/10.1016/j.jas.2007.08.015

Braun DR, Pante M, Archer W (2016) Cut marks on bone surfaces: influences on variation in the form of traces of ancient behaviour. Interface Focus 6.https://doi.org/10.1098/rsfs.2016.0006

Chetlur S, Woolley C, Vandermersch P et al (2014) cuDNN: efficient primitives for deep learning. arXiv 1–9

Choi K, Driwantoro D (2007) Shell tool use by early members of Homo erectus in Sangiran, central Java, Indonesia: cut mark evidence. J Archaeol Sci 34:48–58. https://doi.org/10.1016/j.jas.2006.03.013

Chollet F (2017) Deep learning with Python. Manning Publications Co., Shelter Island

Cifuentes-Alcobendas G, Domínguez-Rodrigo M (2019) Deep learning and taphonomy: high accuracy in the classification of cut marks made on fleshed and defleshed bones using convolutional neural networks. Sci Rep 9:1–12. https://doi.org/10.1038/s41598-019-55439-6

Courtenay LA, Yravedra J, Mate-González MÁ et al (2017) 3D analysis of cut marks using a new geometric morphometric methodological approach. Archaeol Anthropol Sci 11:651–665. https://doi.org/10.1007/s12520-017-0554-x

Domínguez-Rodrigo M (2012) Stone tools and fossil bones. Cambridge University Press, Cambridge

Domínguez-Rodrigo M, Cifuentes-Alcobendas G, Jiménez-García B et al (2020) Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. Sci Rep 10:1–12. https://doi.org/10.1038/s41598-020-75994-7

Domínguez-Rodrigo M, Fernández-Jaúregui A, Cifuentes-Alcobendas G, Baquedano E (2021) Use of generative adversarial networks (Gan) for taphonomic image augmentation and model protocol for the deep learning analysis of bone surface modifications. Appl Sci 11(11). https://doi.org/10.3390/app11115237

Galán AB, Domínguez-Rodrigo M (2014) Testing the efficiency of simple flakes, retouched flakes and small handaxes during butchery. Archaeometry 56:1054–1074. https://doi.org/10.1111/arcm.12064

Gifford-Gonzalez D (1991) Bones are not enough: analogues, knowledge, and interpretive strategies in zooarchaeology. J Anthropol Archaeol 10:215–254. https://doi.org/10.1016/0278-4165(91)90014-O

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Massachussets

Greenfield HJ (1999) The origins of metallurgy: distinguishing stone from metal cut-marks on bones from archaeological sites. J Archaeol Sci 26:797–808

Greenfield HJ (2006) Slicing cut marks on animal bones: diagnostics for identifying stone tool type and raw material. J F Archaeol 31:147–163

Jiménez-García B, Aznarte J, Abellán N et al (2020) Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars. J R Soc Interface 17.https://doi.org/10.1098/rsif.2020.0446rsif20200446

Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc 1–15

Maté González MÁ, Yravedra J, González-Aguilera D, Palomeque-González JF, Domínguez-Rodrigo M (2015) Micro-photogrammetric characterization of cut marks on bones. J Archaeol Sci 62:128–142. https://doi.org/10.1016/j.jas.2015.08.006

Maté-González MÁ, Palomeque-González JF, Yravedra J, González-Aguilera D, Domínguez-Rodrigo M (2016) Micro-photogrammetric and morphometric differentiation of cut marks on bones using metal knives, quartzite, and flint flakes. Archaeol Anthropol Sci 10:805–816. https://doi.org/10.1007/s12520-016-0401-5

Merritt SR (2012) Factors affecting Early Stone Age cut mark cross-sectional size: Implications from actualistic butchery trials. J Archaeol Sci 39:2984–2994. https://doi.org/10.1016/j.jas.2012.04.036

Misra D (2019) Mish: a self regularized non-monotonic neural activation function. arXiv

Olsen SL (1988) The identification of stone and metal toolmarks on bone artifacts. In: Olsen SL (ed) Scanning electron microscopy in archaeology. BAR International Series, London, pp 337–360

Pizarro-Monzo M, Domínguez-Rodrigo M (2020) Dynamic modification of cut marks by trampling: temporal assessment through the use of mixed-effect regressions and deep learning methods. Archaeol Anthropol Sci 12.https://doi.org/10.1007/s12520-019-00966-6

Ramachandran P, Zoph B, Le QV (2017a) Searching for activation functions. arXiv 1–13

Ramachandran P, Zoph B, Le QV (2017b) SWISH: a self-gated activation function. arXiv 1–12

Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33:1–39. https://doi.org/10.1007/s10462-009-9124-7

Schmidhuber J (2015) Deep Learning in neural networks: an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Val A, Costamagno S, Discamps E et al (2017) Testing the influence of stone tool type on microscopic morphology of cut-marks: experimental approach and application to the archaeological record with a case study from the Middle Palaeolithic site of Noisetier Cave (Fréchet-Aure, Hautes-Pyrénées, Franc. J Archaeol Sci Rep 11:17–28. https://doi.org/10.1016/j.jasrep.2016.11.028

Von Lettow-Vorbeck CL (1998) El Soto de Medinilla: Faunas de mamíferos de la Edad del Hierro enel Valle del Duero (Valladolid, España). Archaeofauna 7:11–210

Walker PL, Long JC (1977) An experimental study of the morphological characteristics of tool marks. Am Antiq 42:605–616

West JA, Louys J (2007) Differentiating bamboo from stone tool cut marks in the zooarchaeological record, with a discussion on the use of bamboo knives. J Archaeol Sci 34:512–518. https://doi.org/10.1016/j.jas.2006.06.007

Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural Comput 8:1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. Trans Evol Comput 1:67–82. https://doi.org/10.1007/978-3-662-62007-6_12

Yravedra J, Maté-González MÁ, Palomeque-González JF et al (2017) A new approach to raw material use in the exploitation of animal carcasses at BK (Upper Bed II, Olduvai Gorge, Tanzania): a micro-photogrammetric and geometric morphometric analysis of fossil cut marks. Boreas 46:860–873. https://doi.org/10.1111/bor.12224

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