Geobios 72-73 (2022) 12–21

Contents lists available at ScienceDirect

Geobios

journal homepage: www.elsevier.com/locate/geobio

# High-accuracy in the classification of butchery cut marks and crocodile tooth marks using machine learning methods and computer vision algorithms ☆

Natalia Abellán [a,d], Enrique Baquedano [a,e], Manuel Domínguez-Rodrigo [a,b,c,*]

[a] Institute of Evolution in Africa (IDEA), University of Alcalá de Henares and Archaeological and Paleontological Museum of the Community of Madrid, Covarrubias 36, 28010 Madrid, Spain
[b] Area of Prehistory, Department of History and Philosophy, University of Alcalá, Alcalá de Henares, Spain
[c] Department of Anthropology, Rice University, Texas, USA
[d] Department of Prehistory and Archaeology, UNED, Paseo Senda del Rey, 7, 28040, Madrid, Spain
[e] Archaeological and Paleontological Museum of the Community of Madrid, Plaza de las Bernardas s/n, 28801 Alcalá de Henares, Madrid, Spain

ARTICLE INFO

*Article history:*
Received 19 April 2022
Available online 8 July 2022

*Keywords:*
Taphonomy
Cut marks
Tooth marks
Machine learning
Deep learning
Convolutional neural networks
Butchery

ABSTRACT

Some researchers using traditional taphonomic criteria (groove shape and presence/absence of microstriations) have cast some doubts about the potential equifinality presented by crocodile tooth marks and stone tool butchery cut marks. Other researchers have argued that multivariate methods can efficiently separate both types of marks. Differentiating both taphonomic agents is crucial for determining the earliest evidence of carcass processing by hominins. Here, we use an updated machine learning approach (discarding artificially bootstrapping the original imbalanced samples) to show that microscopic features shaped as categorical variables, corresponding to intrinsic properties of mark structure, can accurately discriminate both types of bone modifications. We also implement new deep-learning methods that objectively achieve the highest accuracy in differentiating cut marks from crocodile tooth scores (99% of testing sets). The present study shows that there are precise ways of differentiating both taphonomic agents, and this invites taphonomists to apply them to controversial paleontological and archaeological specimens.



## 1. Introduction

Identifying cut marks in the fossil record is essential for the interpretation of early hominin lifestyles. Recently, it has been argued that microstriated tooth marks imparted by crocodile teeth could mimic cut marks to a point in which secure differentiation was compromised (Sahle et al., 2017; McPherron et al., 2021). This was based on metric data from three-dimensional analysis of both types of marks. Nevertheless, an alternative analysis based on a much larger sample of bone surface modifications (BSM), including crocodile tooth marks, cut marks made with simple and retouched flakes and trampling marks, using mostly categorical variables factorized according to microscopic structural mark features, yielded an opposite result through the use of machine learning (ML) algorithms, with 96%–100% of BSM correctly identified (Domínguez-Rodrigo and Baquedano, 2018). Metric and categorical variables should not yield such divergent results, unless BSM can only be differentiated by the expression of microscopic features and not by their dimensions. A subsequent reanalysis of the metric data showed that experimentally-derived crocodile tooth marks and butchery cut marks could be separated in Euclidean space when applying several multivariate methods (hierarchical clustering on factor map, K-means partitioning, and ML random forest; Domínguez-Rodrigo and Baquedano, 2018).

It could be argued that the high accuracy in the classification of these types of BSM by Domínguez-Rodrigo and Baquedano (2018) may result from having artificially expanded the samples through bootstrapping prior to analysis (McPherron et al., 2022). Bootstrapping can generate label-specific large samples that artificially separate classes. This would be a potential bias when dealing with small samples or samples unrepresentative of the population from which they derive (Chernick and LaBudde, 2014). In order to avoid this potential bias, it would be necessary to reassess the efficiency

☆ Corresponding editor: Gildas Merceron.

* Corresponding author.
*E-mail addresses:* manuel.dominguezr@uah.es, mdr@rice.edu (M. Domínguez-Rodrigo).

https://doi.org/10.1016/j.geobios.2022.07.001
0016-6995/

of ML algorithms without the use of such a statistical procedure. For this reason, here we intend to reanalyze the same dataset used by Domínguez-Rodrigo and Baquedano (2018) in order to evaluate the impact of bootstrapping and the classification of BSM without it.

Another second potential bias that we will address is the subjective assessment by the analyst of variable factors when using categorical variables. It has been shown how divergent categorization of variables could be when the same BSM were analyzed by different researchers (Domínguez-Rodrigo et al., 2017, 2019). This underscores the important bias introduced by the researcher and portrays the traditional approach to BSM identification as a subjective endeavor.

Fortunately, Deep Learning (DL) methods, through the use of deep convolutional neural networks (DCNN), has enabled automatic classification of BSM through an objective process. DCNNs operate through different DL architectures to generate a mathematical understanding of the micro-features found on taphonomic images of BSM, which are used to discriminate among labels. This supervised computer vision (CV) method is even more effective than human experts. A pioneer application of this method to a limited set of BSM showed that the machine could classify >50% better than human experts (Byeon et al., 2019). Subsequent applications to various taphonomic problems showed the enormous potential of DL methods for implementing objective approaches to BSM identification in modern experiments and in the archaeological record. Cut marks generated with or without flesh on the bone showed distinctive micro-features that enabled the discrimination of the resulting cut marks in both experimental scenarios with >90% of accuracy (Cifuentes-Alcobendas and Domínguez-Rodrigo, 2019). These involved a higher degree of micro-flaking inside the groove, and disruptions of the trajectory of microstriations. The dynamic morphing of cut marks through abrasive processes could also be identified with high accuracy by CV methods (Pizarro-Monzo and Domínguez-Rodrigo, 2020). Even the extreme similarity (to the human eye) of tooth marks generated by diverse carnivores could be discriminated to specific agent by these methods (Abellán et al., 2021); to the point of even differentiating tooth marks from similar carnivore types, like lions or jaguars (Jiménez-García et al., 2020a, 2020b). The application of DL to the most commonly found BSM in the archaeological record (cut marks, trampling marks and carnivore tooth marks) has led to the successful identification of these BSM in more than 92% of cases (Domínguez-Rodrigo et al., 2020). Given that these CV methods have overcome traditional BSM identification approaches, based on the widely variable experience and subjectivity of human experts, their application to this kind of taphonomic problems is more than warranted.

Here, we will apply several DL architectures also to differentiate crocodile tooth marks from human-imparted butchery marks in the form of cut marks. We will show that the accuracy in the classification of these types of marks by DL exceeds those reported in previous studies, granting more confidence to the visual identification of these two types of BSM in the archaeological record through the analysis of high-quality images. This provides the basis for a confident identification of agency in prehistoric BSM.

## 2. Material and methods

### 2.1. Machine learning analysis

Table 1 shows the set of variables used for the present study. The BSM sample consists of 105 cut marks made with retouched flakes, 246 cut marks made with simple flakes, 224 trampling marks and 58 tooth marks (scores) made by crocodiles. These BSM samples were already published in several studies, and the protocols applied for the performance of the experiments, the cleaning of bones and the identification of marks can be found in the original publications (Domínguez-Rodrigo et al., 2009; Baquedano et al., 2012; Domínguez-Rodrigo and Baquedano, 2018). All the experiments followed the same protocol in the application of these and other variables. A summary of all these experiments can be found in Appendix A. A total of 70% of the original BSM sample was used for generating the training models on the raw data. Testing was carried out on the remaining 30% of the sample. This is a standard procedure in predictive models in order to deal with the bias/variance tradeoff. To minimize the impact of heterocedasticity, data were centered and scaled prior to analysis.

Several ML algorithms were compared for efficiency and accuracy. Model evaluation took place through resampling techniques that estimate performance by selecting subsamples of the original data and fitting them in multiple submodels. The results of these submodels were aggregated and averaged. Several techniques can be used for this subsampling and submodelling: generalized cross-validation, k-fold cross-validation, leave-one-out cross validation or bootstrapping. Here, we selected a 5-fold cross-validation approach.

Once all models are completed, model selection takes place. This is usually done combining indicators of error or accuracy. Cost values of bias-variance were evaluated vis-à-vis accuracy with the caret function "tuneLength" up to 10 (i.e., $2^2 \ldots 2^7$). This makes the system tunes the algorithm automatically, enabling specifying the number of tuning values for each parameter. The parameter selected for measuring model performance was the 'kappa' indicator. For class prediction, these can come in two forms: a discrete category (showing the factor classification) and a probability of membership to any specific category. This latter can be continuous (as in random forests or discriminant analyses, for example) or binary when using sigmoid classifiers (as in logistic regression or classical support vector machines). The Kappa statistic (which evaluates the amount of accuracy generated by chance) range from −1 to 1 (as in correlation). Coheńs kappa value is a more robust measure of prediction and classification than accuracy, because it does not quantify the level of agreement between different datasets, but it represents the degree of similarity of datasets corrected by chance. Additionally, we paid special attention to balanced accuracy, sensitivity and specificity, given the unbalanced nature of the original samples.

The original analysis was previously carried out using the complete set of variables, which included intrinsic as well as extrinsic variables (Domínguez-Rodrigo and Baquedano, 2018). Intrinsic variables are those that define the structural features of the mark, such as shape and microscopic characteristics on the edge and inside the groove. Extrinsic variables are those that refer to the configurational properties of BSM and their surrounding areas on the bone surface (Domínguez-Rodrigo et al., 2010). Making assessments about the efficiency of ML methods using the complete set of variables would be prone to error if considering archaeological BSM. The reason is that the experimental BSM samples were derived in absence of interference of other taphonomic signals; that is, all BSM assemblages occur on clean bone surfaces not impacted by posterior biostratinomic or diagenetic processes. Most archaeological BSM are not preserved as pristinely as these experimental marks. Several authors have emphasized the need to model BSM under more dynamic processes which consider the impact of additional modifications on bone surfaces, such as carnivore modification, or natural processes such as abrasion and diagenesis (Gaudzinski-Windheuser et al., 2010; Pineda et al., 2014, 2019; Pizarro-Monzo and Domínguez-Rodrigo, 2020). For this reason, the most reliable ML results in the interpretation of archaeological BSM will be obtained when using intrinsic (i.e., structural)

**Table 1**
Definition of each of the variables used for the ML analysis in the present study.

1. **Trajectory of the groove**. Marks can show a straight trajectory (1a), a curved one (1b), a sinuous one (1c) and a variable one (1d). The latter involves changes in the trajectory direction more than twice. This categorization applies to most of the outline of the mark, excluding the presence of barbs at the end of the mark. Butchery marks are commonly straight grooves. In some cases, the abrasive marks created by sediment grains show a somewhat sinuous trajectory in part of the groove due to the rolling of the grain and the use of different edges of the grain for abrading the bone surface. Some apparently straight trampling marks, when observed under magnification, show trajectories that are not perfectly straight but are rather somewhat wavy. The movement of bones during the tight grasping by the jaw makes some crocodile tooth scores change direction.
2. **Presence (2a) or absence (2b) of a barb**. In some butchery marks, a barb can be observed at the end of the straight groove, defined as a shallower end of the groove slightly curved to the side in the form of an open hook. Testing how frequent this feature is in cut marks and in trampling marks can be potentially important, since it has also been observed in the latter.
3. **Orientation of the mark**, relative to the axis of the bone. The orientation can be parallel (3a), perpendicular (3b) or oblique (3c) to the axis of the bone. Trampling marks, in theory, should show no preference in orientation, whereas butchery marks should be more frequently oriented obliquely or perpendicularly to the axis of the bone.
4. **Shape of the groove**. The shapes used are: narrow V-shape (4a), wide V-shape (_/) (4b) and U-shaped (4c). The wide V-shaped section is understood as either V- or _/-shaped but either almost as deep as it is wide, or deeper than it is wide; the latter is understood as an open groove with a broader horizontal base and, therefore, substantially wider (by an order of magnitude $>\times 2$) than deeper.
5. **Symmetry of the groove**: the section and both sides of the groove can be symmetrical (5a) or asymmetrical (5b). The tilting of a stone tool during use can create asymmetrical grooves, and so can certain sediment particles during bone abrasion.
6. **Shoulder effect** and associated shallower striae. Here we define the term as the striae occurring in association with the main groove in a distance not farther than 0.2 mm from the edge of the groove. For this type of analysis, a binocular lens with measuring capability is preferred. These striae frequently are shallow striations occurring parallel to or intersecting with the sides of the groove. They can be present (6a) or absent (6b) and have been documented in trampling marks, cut marks and crocodile bite marks
7. **Presence of flaking** on the shoulders of the groove. The presence (over more [7a] or less [7b] than one-third of the trajectory of one or two shoulders of the groove) or absence (7c) of flaking on the shoulders of the groove can be related to the morphology of the abrasive agent: the bigger and less straight the edge of this agent the bigger the chance that such flaking would appear. Flaking here is defined as not random occurrence of a flaking dent such as those produced in isolated Hertzian cones, but as a continuous series of exfoliation of the shoulder edge, which can occur on part of the trajectory of the shoulder or on most of it.
8. **Extent of the flaking** of the shoulder. The extent of the flaking could also be indicative of the abrasive agent. The category of the flaking can be defined as long (8a) when it occurs over a minimum of one-third of the trajectory of the groove, and short (8b) when it is shorter than one-third. Approximate estimates can be made with hand lenses.
9. **Internal microstriations**. Defined as present (9a) or absent (9b) and observable under $\times 40$.
10. **Microstriation trajectory**. Defined as continuous (10a) when it extends along all the trajectory of the groove or discontinuous (10b) when the microstriations are interrupted at more than one instance inside the groove. A tool is more likely to create continuous microstriations given that it creates uniform friction in its contact with bone. A trampling mark is more likely to created discontinuous microstriations if friction forces the sediment particle to move inside the groove. This is also documented in crocodile bite marks given the movements of teeth during the grasping of the bone.
11. **Location of microstriations**. On the walls of the groove (11a), on the bottom (11b) or on both (11c).
12. **Length of the main groove** (in mm).

variables, which reproduce BSM properties under moderate to good preservation, and which can be preserved even after substantial impact of other biostratinomic processes (Pizarro-Monzo and Domínguez-Rodrigo, 2020). For this reason, here we will use only the intrinsic variables of the original dataset (Table 1). We do so knowing that this will lower the accuracy threshold of the tests with respect to previous analyses of the complete set of variables, since extrinsic variables contribute to widening variance, and increasing discrimination (Domínguez-Rodrigo and Baquedano, 2018).

We used the ML algorithms with their default parameter values. After testing the performance of all algorithms over the complete set of variables, we selected the best three models for the second analytical phase. This phase involved reducing the number of variables by selecting the most influential ones; however, each algorithm had its own particular selection of variables, since the most influential ones differed among models. The function "varImp" from the R 'caret' library was selected for this purpose. For classification, the Gini index is used, although it has been recognized that it can lead to false interpretations (Strobl et al., 2007). Conditional forests consider the number of splits per variable through subsampling, instead of the number of features per split (as in random forests). It has been argued that this approach provides better estimates of variable importance, since it is not as biased by cardinality (i.e., the number of factors per variable) as is the Gini index derived from bootstrapping instead of subsampling (Strobl et al., 2007). The higher the cardinality of categorical variables, the artificially higher the Gini index can be. For these reasons, although the cardinality in the dataset is low in general, some exceptions warrant the comparison of variable importance (selection) derived from traditionally-derived Gini indices to those subsampled from conditional forests/trees. No hyper-parameter tuning was applied in order to minimize distance between ML algorithms and more "traditional" tests; however, for some algorithms, tuning was carried out (although not used in the present study) to stress the potential improvement in accuracy when doing so.

The third and final phase consisted of using the same three best-performing algorithms and the linear discriminant analysis (LDA) over a variable set composed of the least influential variables after the removal of the four most relevant variables from Phase 2. The goal of doing this is to show that all variables contain important discriminatory information and that highly-accurate BSM classifications can be made when using them to the exclusion of the most influential ones. This also realistically models dynamic modification of some of the influential variables through biostratinomy or diagenesis. For example, groove shape and size can be modified through diagenetic impact of soil pH (Pineda et al., 2014).

The present work builds upon Domínguez-Rodrigo and Baquedano (2018) initial analysis of the same dataset, and expands the number of ML algorithms used. In order to address any potential skepticism about the preferred use of ML methods over more traditional multivariate methods, we have also added some of the most commonly used traditional tests for comparison. In order to boost their potential, we have implemented these tests with penalization methods. The ML algorithms used include: support vector machines (SVM), K-nearest neighbor (KNN), weighted K-nearest neighbor (wKNN), random forests (RF), decision trees (C5.0), naïve Bayes (NB), partial least square analysis (PLSA), mixture discriminant analysis (MDA), gradient boosting machines (GBM), neural networks (NN), conditional trees (CT), and conditional inference forest (CIT). The traditional multivariate methods used were: linear discriminant analysis (LDA), shrinkage discriminant analysis (SDA),

penalized multinomial regression (PMR), and general linear model with penalized maximum likelihood (pGLM).

### 2.2. Deep learning analysis

The sample used for the DL analysis consists of 488 cut marks and 45 crocodile tooth marks (the same ones used for the ML analysis above) already published by Domínguez-Rodrigo et al. (2020) and Abellán et al. (2020), respectively. For the cut mark experiment, a set of cow long bones (humerus, femur, radius and tibia) was used along with 22 non-retouched flint flakes. Stone tools were used on fully-fleshed elements. Each stone tool was used only 20 times to keep control of edge sharpness, to make sure that edge blunting did not play any significant role in possible cut-mark variability. For additional details of the experiment, see Domínguez-Rodrigo et al. (2020) and Appendix A.

Crocodile tooth marks were obtained from an experiment made in the Faunia zoo in Madrid (Baquedano et al., 2012). This sample is the same as used for the ML analysis (see Appendix A). All the crocodiles used in the experiment were female. They were fed once a week over four complete months with 19 partial carcasses. Carcasses were collected after 15 h 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. The carcass parts were composed of fully fleshed articulated limbs of suids (pig and boar) and bovids (sheep and cow). A total of 198 bone elements were retrieved, counting every end and shaft of unfused bones from juvenile individuals as one. For more details of the sample and experiment, see Baquedano et al. (2012) and Appendix A.

Each BSM was captured with a binocular microscope (Optika) at $\times$30 and images were taken in this magnification using the same light intensity and angle. Then, images were cropped to a point where only the mark and their shoulders, including a minimal surrounding area, were visible, to avoid any bias potentially produced by the cortical surface of the bone. All images were transformed into black and white during image processing in the Keras platform, by using bidimensional matrices for standardization and centering, and they were reshaped to the same dimensions (80 $\times$ 400 pixels). Images were pre-processed using the specific preprocessing functions for each model used.

The DL architectures selected are among the most successful in the Imagenet Large Scale Visual Recognition Challenge (LSVRC), the largest competition of image classification. Given that these models have been trained on millions of images, their feature identification is proficient. For this reason, it has been shown that these models, when imported through transfer learning, can outperform native architectures trained for taphonomic problems from zero (Domínguez-Rodrigo et al., 2020). For this reason, we will use four transfer learning models and only one trained from zero. These pre-trained models were used as standalone feature extractors and classifiers. The layers of the pre-trained models with their weights were integrated within the new model containing a top frozen layer and an output dense layer containing 128 neurons. The DL architectures selected are: VGG16, ResNet50, Densenet 201, EfficientNet B7 (transfer learning) and Jason (a modular version of the VGG architectures). A summary of the description of these architectures can be found in several previous publications where they were used for BSM analysis (Tan and Le, 2019; Domínguez-Rodrigo et al., 2020; Jiménez-García et al., 2020b; Abellán et al., 2021). The DCNN models used here were elaborated using the Keras platform with a Tensorflow backend. Computation was carried out on a GPU HP Z6 Workstation. The DCNN models were processed with the sequential and functional Keras API. All code was made using Python 3.7.

The only model trained from zero was Jason. The architecture represents a variant of the VGG16 block and repeated layer structure. The model consists of a series of three blocks, each of them containing 3 $\times$ 3 kernel double layers of 32, 64, and 128 neurons, respectively. In between each block, there are max-pooling (2 $\times$ 2 kernel) layers. Batch normalization has been applied to all the blocks. Additionally, Dropout has been implemented with increasing proportion (0.2, 0.3, and 0.4). At the end of the network, flattening was performed and a dense layer (128 filters) has been added. This was followed by a 0.5 Dropout layer and topped by a dense layer with 'softmax' activation. Each CNN has been tuned with a 'He uniform' initializer and padding.

For all models used, the activation function for each layer was a rectified linear unit (ReLU). The last fully connected layer of the network used a 'sigmoid' activation for the binary modeling. The loss function selected was 'binary cross-entropy'. The optimizer used was Stochastic Gradient Descend (SGD) with a learning rate of 0.001 and a momentum of 0.9. Accuracy was the metric selected for the training process. F1 score values were also obtained to assess balanced accuracy, given the highly imbalanced nature of the original dataset.

Data Augmentation (DA) is commonly used to artificially increase the sample size of the training dataset and enhance the training process by exposing the algorithm to a higher diversity of positions and features of each independent item in the dataset (Chollet, 2017). It does so by creating hundreds of modified

**Table 2**
Accuracy (including confidence interval), Kappa, sensitivity, specificity and balanced accuracy of the models according to BSM type, including all the intrinsic variable dataset.

| Algorithm | Accuracy | 95%c.i. | Kappa | Sensitivity* | Specificity* | Balanced accuracy* |
|---|---|---|---|---|---|---|
| SVM radial | 94.6 | 0.90–0.97 | 0.92 | (0.82,0.87,1,0.95) | (1,0.98,1,0.94) | (0.91,0.92,1,0.94) |
| SVM linear | 94.6 | 0.90–0.97 | 0.92 | (1,0.87,0.97,0.94) | (0.98,0.99,1,0.95) | (0.99,0.93,0.98,0.94) |
| KNN | 89.3 | 0.84–0.93 | 0.82 | (0.7,0.87,1,0.83) | (80.99,0.96,0.92,0.96) | (0.85,0.91,0.96,0.9) |
| RF | 95.8 | 0.91–0.98 | 0.94 | (0.82,0.93,1,0.95) | (1,0.98,1,0.95) | (0.91,0.95,1,0.95) |
| C5.0 | 96.8 | 0.93–0.98 | 0.95 | (0.82,0.93,1,0.98) | (1,0.99,1,0.95) | (0.91,0.96,1,0.97) |
| NB | 90.9 | 0.85–0.94 | 0.86 | (0.58,0.83,0.97,0.95) | (1,0.98,0.95,0.91) | (0.79,0.91,0.96,0.93) |
| PLS | 87.7 | 82–92 | 0.81 | (0.64,0.87,0.97,0.83) | (0.98,0.99,0.89,0.94) | (0.81,0.93,0.93,0.88) |
| MDA | 92.5 | 88–96 | 0.89 | (0.82,0.83,1,0.92) | (1,0.97,0.96,0.95) | (0.91,0.9,0.98,0.94) |
| LDA | 86.1 | 80–90 | 0.79 | (0.64,0.87,0.97,0.79) | (0.93,0.99,0.93,0.94) | (0.79,0.93,0.95,0.86) |
| wKNN | 95.7 | 91–98 | 0.93 | (0.82,1,1,0.92) | (1,0.96,0.97,1) | (0.91,0.98,0.98,0.96) |
| PMR | 90.9 | 85–94 | 0.86 | (0.76,0.87,0.97,0.89) | (0.97,0.98,0.95,0.95) | (0.87,0.92,0.96,0.92) |
| SDA | 86.1 | 80–90 | 0.79 | (0.64,0.87,0.97,0.79) | (0.93,0.99,0.93,0.94) | (0.79,0.93,0.95,0.86) |
| GBM | 96.8 | 93–98 | 0.94 | (0.82,0.93,1,0.98) | (1,0.99,1,0.95) | (0.91,0.96,1,0.97) |
| pGLM | 87.7 | 0.84–0.93 | 0.82 | (0.76,0.87,0.97,0.86) | (0.96,0.98,0.95,0.95) | (0.86,0.92,0.96,0.9) |
| NN | 95.2 | 0.92–0.98 | 0.93 | (1,0.93,1,0.92) | (0.98,0.98,1,0.98) | (0.99,0.95,1,0.95) |
| cTree | 94.6 | 0.90–0.97 | 0.92 | (0.76,0.96,0.98,0.94) | (1,0.97,0.95,0.99) | (0.88,0.97,0.97,0.96) |

* (croc,rf,sf,tramp). Key: croc, crocodile tooth marks; rf, retouched flakes; sf, simple flakes; tramp, trampling.

versions of the original images. DA does this by changing the orientation of the images (using random angles in a specified rotation range), and by creating new images derived from those in the training dataset. This process is achieved by modifying the ranges in which images are shifted (vertically and horizontally), by zooming in and out and by randomly applying shearing transformations that distort the original images. In the present study, the original images were augmented through random transformations, involving shifts in width and height (20%), in shear and zoom range (20%), and also including horizontal flipping.

Following standard protocols for ML, the architectures of the models used were trained on 70% of the original image set. Models were subsequently tested against the 30% remaining sample, which was not used during the training. Training was performed through mini-batch kernels (size = 64). Testing was made using mini-batch kernels of size = 32. Weight update was made using a backpropagation process for 100 epochs.

## 3. Results

### 3.1. Machine learning analysis

All ML algorithms successfully classified the four BSM types with an accuracy >90% (Table 2). In contrast, the more "traditional" multivariate tests displayed substantially lower accurate classification rates on the testing sets. The most successful ML algorithms were the C5.0, NN, RF and GBM; all of them with accuracy in the classification >95%. There is a difference of more than 10 points between discriminant analyses and the most accurate ML algorithms; all this without having used any hyper-parameter tuning in the later (Table 2).

When looking at the most influential variables (Fig. 1), it must be stressed that there was a wide variability according to the model. The most influential variables were not necessarily so for all models. Likewise, there was wide variability in how models

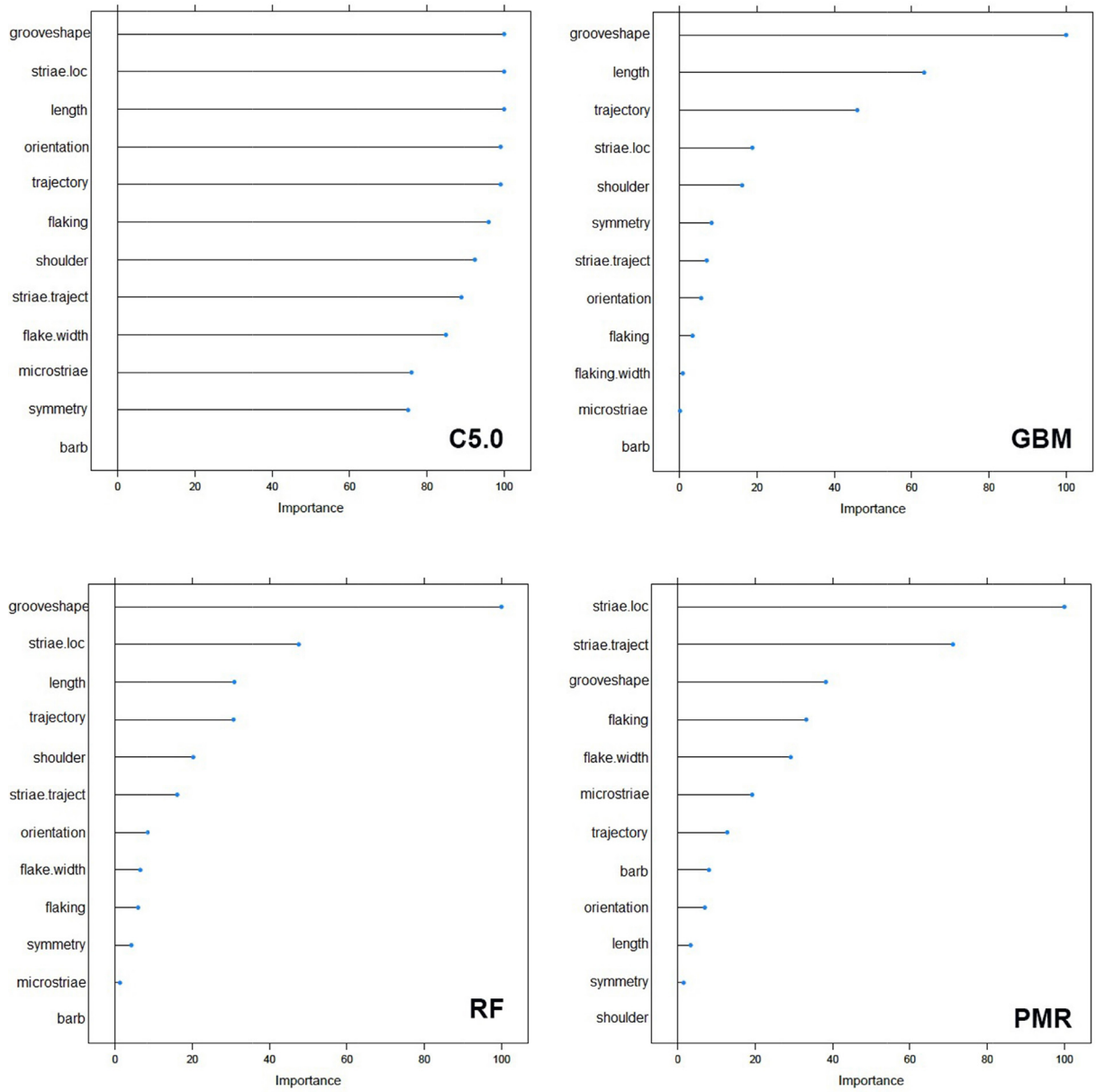



**Fig. 1.** Variable importance according to the Gini index as generated in different models: C5.0 decision tree, gradient boosting machine (GBM), random forest (RF), and penalized multinomial regression (PMR). Note the divergence of the latter with the others and the contrast in the three tree-based models, where some were highly selective regarding number of impacting variables and others were more aggregative.

interpreted variables; for some of them, the most influential variables were substantially so over the others, whereas for other models, most variables had similar importance (Fig. 1). This shows that the use of traditional methods over ML methods, targeting the understanding of the underlying factors of the classification (e.g., McPherron et al., 2022) is at best naïve, given that variable importance will change according to the test used. Therefore, a selection of the four most commonly influential variables was made based on the three most successful models. These involved: groove shape, location of microstriations, length and groove trajectory. These are not the most prominent variables in other models, namely in more traditional tests (Fig. 1). The only variable that was not included in the contrasting conditional inference forest was 'length', probably because of its high range of values (Fig. 2).

The performance of the selected ML algorithms on the reduced variable set was similarly successful, although to a lower degree (Table 3). The C5.0 model yielded an accuracy of 91.4% – when the same tuned model used with the complete intrinsic variable set yielded an accuracy of 98.4%. The GBM and NN models correctly classified 94.6% and 93.1% of the testing set, respectively. Both ML algorithms (when tuning their hyper-parameters) used on the complete set of intrinsic variables also yielded an accuracy >98%. In contrast, as an example, a classical LDA yielded a lower accuracy (84.5%) on the reduced dataset; however, the sharpest contrast was found in the Kappa values, with a difference of up to 12 points of the values produced by the three ML algorithms and the LDA (Table 3). This difference is even bigger when using the complete set of variables (up to 15 points) (Table 2). LDA was used as a contrasting example over PMR and pGLM because these classifiers yielded even lower accuracy and Kappa values. This indicates that traditional discriminant tests are not only substantially less accurate, but they also show poorer performance in real accuracy when sample imbalance is considered (as through the Kappa indicator). This also indicates that reducing the dimensionality of the dataset results in reduced accuracy.

In Phase 3, the use of the least influential variables yielded lower accuracy and Kappa estimates than the influential variable set (Table 4). Despite this, all tests showed an accuracy >72% of correct classification of all BSM. This shows that even low-influential variables can be efficiently used for discrimination when they are handled in a multivariate format. The synergy created by the limited variance of each variable separately expands when using all variables simultaneously. It should be emphasized that whereas LDA showed a difference in accuracy of 13 points between using the complete intrinsic variable ensemble and the reduced least-influential variable set, the ML algorithms showed substantially wider differences (20–25 points).

### 3.2. Deep learning analysis

All the models showed a steady training without large oscillations (Fig. 3). This indicates that micro-feature differentiation of

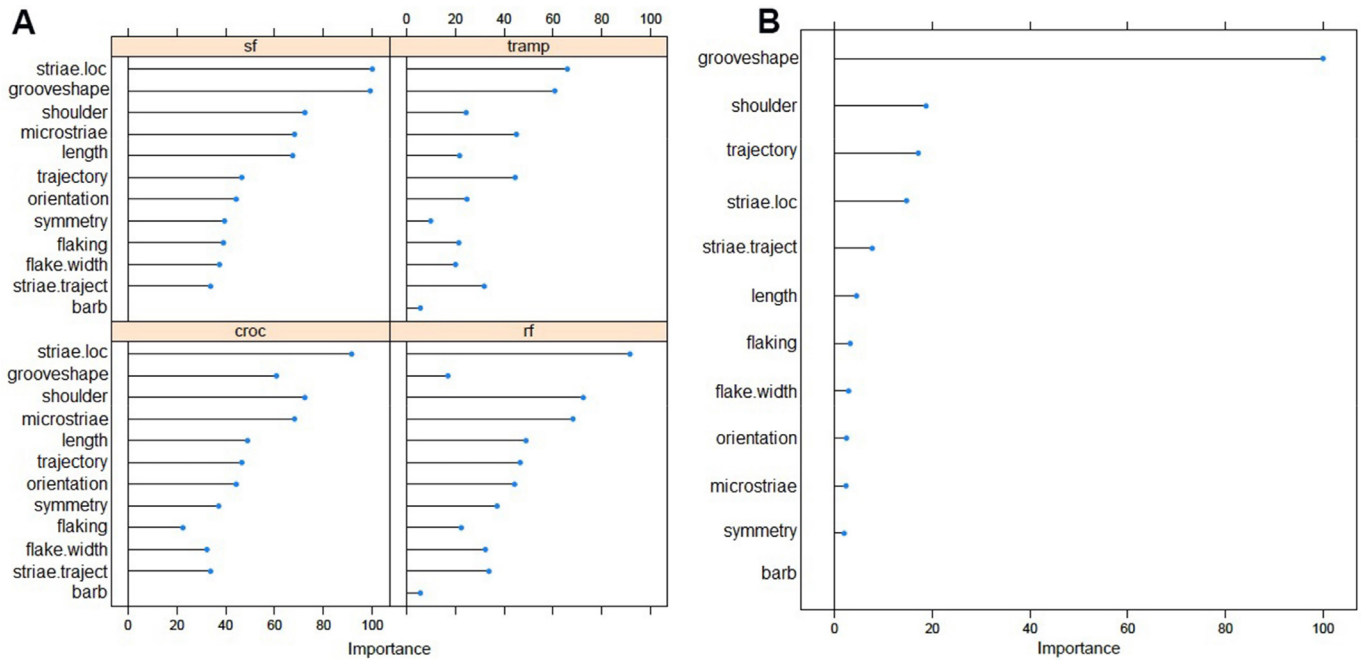



**Fig. 2.** Variable importance according to BSM type as detected through conditional trees (**A**) and of the complete dataset as determined by conditional inference forest (**B**). Both methods were using subsampling instead of bootstrapping. sf, cut marks made with single flakes; tramp, trampling; croc, crocodile tooth marks; rf, cut marks made with retouched flakes.

**Table 3**
Accuracy (including confidence interval), Kappa, sensitivity, specificity and balanced accuracy of the models according to BSM type, based on the reduced (four most influential variables) dataset.

| Algorithm | Accuracy | 95%c.i. | Kappa | Sensitivity* | Specificity* | Balanced accuracy* |
|---|---|---|---|---|---|---|
| C5.0 | 91.49 | 0.86–0.95 | 0.88 | (0.87,0.83,1,0.89) | (1,0.96,0.98,0.93) | (0.88,0.90,0.99,091) |
| GBM | 94.68 | 0.90–0.97 | 0.92 | (0.76, 0.93,0.98,0.95) | (1,0.97,0.99,0.95) | (0.88,0.95,0.98,0.95) |
| NN | 93.1 | 0.88–0.96 | 0.89 | (0.82,0.87,1,0.92) | (1,0.97,0.96,0.96) | (0.91,0.92,0.98,0.94) |
| LDA | 84.5 | 0.78–0.89 | 0.77 | (0.76,0.87,0.97,0.71) | (0.96,0.91,0.93,0.98) | (0.86,0.89,0.95,0.84) |

* (croc,rf,sf,tramp). Key: croc, crocodile tooth marks; rf, retouched flakes; sf, simple flakes; tramp, trampling.

**Table 4**
Accuracy (including confidence interval), Kappa, sensitivity, specificity and balanced accuracy of the models according to BSM type, based on the ensemble of the least influential variables.

| Algorithm | Accuracy | 95%c.i. | Kappa | Sensitivity* | Specificity* | Balanced accuracy* |
|---|---|---|---|---|---|---|
| C5.0 | 72.81 | (0.64–0.78) | 0.60 | (0.41,0.8,0.64,0.83) | (0.96,0.97,0.86,0.77) | (0.68,0.89,0.75,0.8) |
| GBM | 75.53 | (0.68–0.81) | 0.63 | (0.35,0.80,0.73,0.85) | (1,0.98,0.85,0.77) | (0.67,0.89,0.79,0.81) |
| NN | 75.53 | (0.68–0.81) | 0.63 | (0.35,0.80,0.73,0.85) | (0.99,0.99,0.85,0.77) | (0.67,0.9,0.79,0.81) |
| LDA | 73.8 | (0.67–0.80) | 0.62 | (0.76,0.8,0.69,0.74) | (0.94,0.98,0.88,0.79) | (0.85,0.89,0.79,0.76) |

* (croc,rf,sf,tramp). Key: croc, crocodile tooth marks; rf, retouched flakes; sf, simple flakes; tramp, trampling.

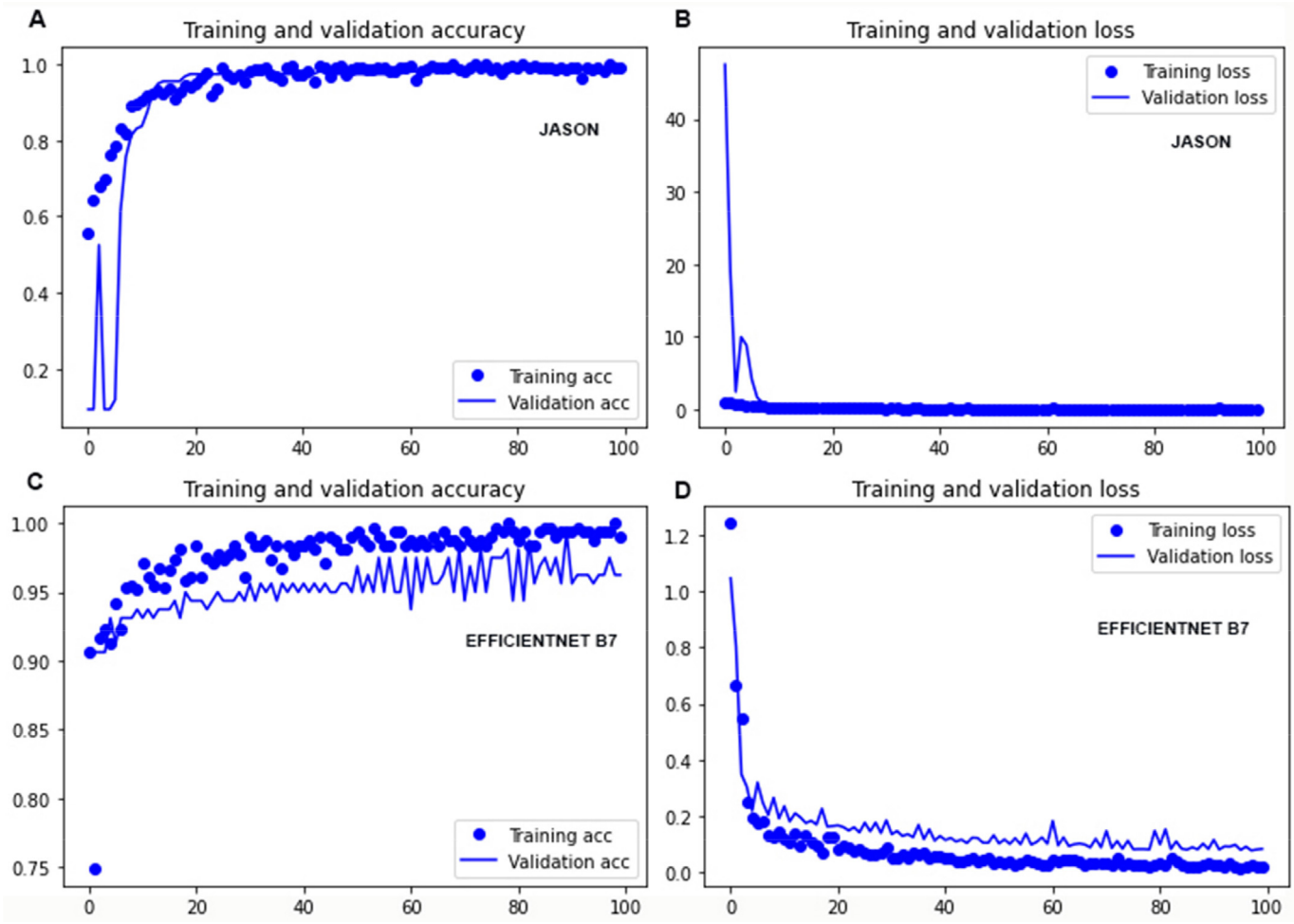



**Fig. 3.** Example of performance in the accuracy and loss during training for the best model (Jason) and comparison with one of the least efficient models (EfficientNet B7). **A, C**. Training and validation accuracy of each model. **B, D**. Training and validation loss of each model.

BSM for both classes is unambiguous. All models coincide in showing a high degree of accuracy in classifying the BSM images from the testing set (96%–99%) (Table 5). The Jason architecture is the most successful by showing the highest accuracy (99%) and lowest loss (0.02). It also displays the highest F1-score (0.96), which shows that despite the highly imbalanced nature of the original samples, the classification rate is balanced between both types of BSM. ResNet50 is very similar in results, differing only minimally from Jason. The other transfer models, although with a lower performance, also show high accuracy (>96%), low loss (<0.15) and high F1-score (>0.86). The DL analysis shows that there is a clear, objective method to differentiate stone-tool imparted cut marks from crocodile tooth marks with a high degree of confidence.

**Table 5**
Accuracy, loss, F1-score and Area under the Curve (AUC) for each DL model.

| Model | Accuracy | Loss | F1 | AUC |
|---|---|---|---|---|
| VGG16 | 0.975 | 0.15 | 0.92 | 0.86 |
| ResNet50 | 0.987 | 0.06 | 0.96 | 0.93 |
| Densenet 201 | 0.962 | 0.109 | 0.86 | 0.81 |
| EfficientNet B7 | 0.962 | 0.08 | 0.86 | 0.81 |
| Jason | 0.99 | 0.02 | 0.96 | 0.96 |

## 4. Discussion

The ML analysis has shown that ML algorithms are more powerful at classification than some robust approaches (i.e., including penalization) to "traditional" multivariate classifiers. This is documented by global accuracy, balanced accuracy and, given the unbalanced nature of the samples used, by the Kappa indicator. Multivariate information increases the chances of correct classification. ML techniques are also better than traditional classifiers

at handling multidimensionality. A clear example is found in the limited variation documented in LDA when using the complete set of variables or the reduced set of the most commonly influential variables (Tables 2, 3). In contrast, the most successful ML models exhibit the highest global/balanced accuracy (and Kappa) when using all the variables instead of the reduced set. Traditional statisticians might feel tempted to minimize dimensionality for classification, but when the number of variables is as limited as the dataset used here, complete use is preferred over partial one, especially if the application of ML algorithms is intended. Exploratory techniques may be suitable to detect non-influential variables and if their variance is negligible, then their discard could be better justified; but as long as variables have any impact on the ML classifiers it is better to include them instead of dropping them. The results shown in Phase 3 clearly supports this assertion.

When considering also the sensitivity and specificity of the models, it can be clearly seen that most marks, especially cut marks, are well identified and differentiated from the other BSM. The high sensitivity of BSM (including crocodile tooth marks) indicate that most are correctly classified by all the models. Therefore, the categorized microscopic features used as variables do have a discriminatory power and question interpretations about the equifinality in the identification of BSM types. The present ML models can classify correctly as many as 96.8% of BSM of the testing set using all variables, or 94.6% of BSM using the reduced variable set. These frequencies can be improved if the ML models are tuned in their hyperparameters. The three models displaying the highest accuracy in the present study, when tuned, showed accuracy rates >98%. This underscores the advantages of ML over other classificatory methods. This also shows that the methodological bias introduced by the use of bootstrapping by Domínguez-Rodrigo and Baquedano (2018) is negligible and has no impact in classification.

The main objection to the use of this approach, based on the categorization of microscopic features and their statistical multivariate treatment, is not its discriminatory capability, but the unavoidable impact of subjective assessment of each variable (Domínguez-Rodrigo et al., 2017, 2019; 2019). For this reason, the use of DL models is a welcome improvement, since it removes the subjective human factor of variable categorization. The most successful DL models used in the present work (on a sample of more than 500 images) show a degree of accurate discrimination of the testing sets of 99% of all the images. In the case of the Jason model, 99% of cut marks and 93% of crocodile tooth marks were correctly classified. An additional advantage of the DL methods is that they provide confidence probability in the identifications. This increases the reliance on interpretations of archaeological and paleontological BSM.

Sahle et al. (2017) argued about the potential crocodile agency in the modification of some Ethiopian Plio-Pleistocene fossils that they published. In order to test this interpretation properly through CV methods, the images should be taken following the experimental protocol of resolution and magnification described here. In the absence of such an image dataset, we can only speculate about agency. We agree with Sahle et al. (2017) that several of the damage traces found on some of the Pliocene fossils resemble tooth marks created by crocodiles. One hominin humeral shaft bears one mark that is virtually identical (Sahle et al., 2017) to the Dikika rib specimen (McPherron et al., 2010). For the 4.2 Ma-old ASI-VP-2/420 hominin humerus, the MAK-VP-1-754 ungulate humerus, the AL339 equid tibia, and maybe the BOU-VP-11-15 bovid tibia, the evidence about crocodile agency seems compelling. We differ from these authors in asserting that such modifications could be mistaken with butchery marks. For those V-shaped crocodile marks, their small size and the absence of several other microscopic features that commonly accompany cut marks (such as microstriations, groove asymmetry, shoulder effect, flaking on mark shoulder) show that they can be differentiated from most cut marks that do indeed show some or all of these features combined. For those crocodile tooth marks that bear microstriations, the groove morphology is broad and not V-shaped when proper magnification is applied (Baquedano et al., 2012; Sahle et al., 2017; Domínguez-Rodrigo and Baquedano, 2018). Despite being unsuitable for analysis, we selected some of the largest magnified images in Sahle et al. (2017) as a preliminary exercise to show the potential of the CV method. Although the conclusions of the results should not be considered serious, because the images used did not follow the same protocols as the experimental mark images, it is worth noting that the two best CV models derived from the experimental dataset seem to potentially interpret the Pliocene fossil traces as crocodile-made (against the multivariate testing of the metric data; Domínguez-Rodrigo and Baquedano, 2018) (Fig. 4).

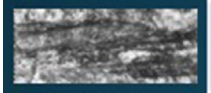
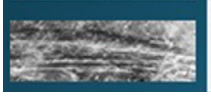
| | | Jason | | ResNet50 | | Classification |
|---|---|---|---|---|---|---|
| | | Crocodile Probability | Cut mark Probability | Crocodile Probability | Cut mark Probability | |
|  | AL339 | 0.99 | 0.99 | 0.99 | 0.99 | Crocodile |
|  | ASI-VP-2/420 | 0.99 | 0.99 | 0.99 | 0.99 | Crocodile |
|  | BOU-VP-11/15 | 0.99 | 0.99 | 0.99 | 0.99 | Crocodile |
|  | BOU-VP-11/14 | 0.99 | 0.99 | 0.99 | 0.99 | Crocodile |
|  | BOU-VP-11/12 | 0.44 | 0.56 | 0.41 | 0.59 | Cut mark |

**Fig. 4.** Selection of some of the best photographed marks from fossils found in the Ethiopian Pliocene and early Pleistocene areas from Sahle et al. (2017) and their classification by the Resnet50 and Jason DL models. Numbers show probability of classification per category (cut mark, crocodile tooth score).

**Table 6**
Accuracy rates of different ML algorithms applied to the complete dataset (i.e., intrinsic and extrinsic variables) in two studies before and after the use of bootstrapping. Underlined values show higher or equally high accuracy in algorithms that did not use bootstrapping. NN, neural networks; SVM, support vector machines; KNN, K-nearest neighbor; RF, random forests; MDA, mixture discriminant analysis; NB, naive Bayes; PLSDA, partial least square discriminant analysis.

| | Domínguez-Rodrigo & Baquedano (2018) | | Moclán et al. (2019) | |
|---|---|---|---|---|
| | Raw data accuracy | Bootstrapped data accuracy | Raw data accuracy | Bootstrapped data accuracy |
| NN | 0.99 | 1.00 | 0.89 | 0.93 |
| SVM | 0.98 | 0.99 | 0.88 | 0.92 |
| KNN | 0.95 | 0.99 | 0.82 | 0.87 |
| RF | 0.98 | 0.99 | 0.89 | 0.94 |
| MDA | 0.97 | 0.98 | 0.84 | 0.84 |
| NB | 0.97 | 0.96 | 0.82 | 0.78 |
| PLSDA | 0.96 | 0.96 | - | - |
| C5.0 | 0.98 | 0.99 | - | - |

The present study also shows that bootstrapping did not bias the accuracy of the ML classifiers in Domínguez-Rodrigo and Baquedano (2018) (Table 6). Recently, this has been suggested by McPherron et al.'s (2022) purported replication of the analysis using artificially-derived variables. McPherron et al. (2022) have suggested that our use of bootstrap overfitted the data, resulting in high accuracy estimates. They re-modelled the process by using artificial variables, which have no discriminatory power in their raw state, but which reach perfect classification after they are bootstrapped 10.000 times. We argue that their method did not reproduce our original analysis. These authors also argue that traditional statistical methods have similar accuracy as ML methods. All these claims are carefully scrutinized in Appendix A and proven inaccurate.

## 5. Conclusions

Multivariate analysis of structural microscopic features (i.e., intrinsic variables) of BSM can effectively be used to discriminate different types of marks and, more specifically, crocodile tooth marks from butchery cut marks. One does not need to use complex statistics and traditional multivariate discriminant methods could potentially be used (Domínguez-Rodrigo et al., 2009; Harris et al., 2017); however, these have shown to be less accurate than ML algorithms, especially when using variables that are predominantly categorical (Domínguez-Rodrigo, 2018). Machine learning may be seen by some as a black box set of methods; but the truth is quite different. If one knows the way each algorithm mathematically addresses problems, the computational procedure can be fully understood. There are now tools to make the mathematical process more understandable. White-box methods are being implemented even within the realm of DL (Landecker, 2000; Yang et al., 2019; Ayyar et al., 2021; Molnar, 2020).

Likewise, the intricate depth of DL methods can be fully understood by pulling out each layeŕs feature target (Brownlee, 2017). DCNN have revolutionized the way image-based science is done. Its preliminary application to taphonomic research is providing an objective and highly-confident platform from which assessment of BSM can be made with more reliance than in the past. The CV results shown in the present work regarding discrimination of crocodile tooth marks and butchery cut marks are enlightening and are a good example of its potential. They should encourage taphonomists to adopt them to address interpretation of BSM in the past.

## Declaration of Competing Interest

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

## Acknowledgements

We thank the Spanish Ministry of Science, Innovation and Universities for funding this research (PID2020-115452GB-C21). The code and the dataset used for the machine learning analysis can be found in Appendix A. We thank two anonymous reviewers for their in-depth comments on the original draft of this manuscript. We also thank Gabriel Cifuentes-Alcobendas for letting us use his cut mark visual library. The original dataset used for the DL analysis can be found at https://doi.org/10.7910/DVN/9NOD8W.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geobios.2022.07.001.

## References

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

Ayyar, M., Benois-Pineau, J., Zemmari, A., 2021. White Box Methods for Explanations of Convolutional Neural Networks in Image Classification Tasks. Journal of Electronic Imaging. arXiv:2104.02548.

Baquedano, E., Domínguez-Rodrigo, M., 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, 1728–1737.

Brownlee, J., 2017. Machine Learning mastery with Python. Machine Learning Mastery, Adelaide, Australia, p. 177.

Byeon, W., Domínguez-Rodrigo, M., Arampatzis, G., Baquedano, E., Yravedra, J., Maté-González, M.A., Koumoutsakos, P., 2019. Automated identification and deep classification of cut marks on bones and its paleoanthropological implications. Journal of Computer Science 32, 36–43.

Chernick, M.R., LaBudde, R.A., 2014. An Introduction to Bootstrap Methods with Applications to R. John Wiley & Sons, 225 p.

Chollet, F., 2017. Deep Learning with Python. Manning Publications Company, p. 597.

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.

Domínguez-Rodrigo, M., 2018. Successful classification of experimental bone surface modifications (BSM) through machine learning algorithms: a solution to the controversial use of BSM in paleoanthropology? Archaeological and Anthropological Sciences 11, 2711–2725.

Domínguez-Rodrigo, M., Baquedano, E., 2018. Distinguishing butchery cut marks from crocodile bite marks through machine learning methods. Scientific Reports 8, 5786.

Domínguez-Rodrigo, M., de Juana, S., Galán, A.B., 2009a. A new protocol to differentiate trampling marks from butchery cut marks. Journal of Archaeological Science 36, 2643–2654.

Domínguez-Rodrigo, M., de Juana, S., Galán, A.B., Rodríguez, M., 2009b. A new protocol to differentiate trampling marks from butchery cut marks. Journal of Archaeological Science 36, 2643–2654.

Domínguez-Rodrigo, M., Pickering, T.R., Bunn, H.T., 2010. Configurational approach to identifying the earliest hominin butchers. Proceedings of the National Academy of Sciences 107, 20929–20934.

Domínguez-Rodrigo, M., Saladié, P., Cáceres, I., Huguet, R., Yravedra, J., Rodríguez-Hidalgo, A., Martín, P., Pineda, A., Marín, J., Gené, C., Aramendi, J., Cobo-Sánchez, L., 2017. Use and abuse of cut mark analyses: The Rorschach effect. Journal of Archaeological Science 86, 14–23.

Domínguez-Rodrigo, M., Cifuentes-Alcobendas, G., Jiménez-García, B., Abellán, N., Pizarro-Monzo, M., Organista, E., Baquedano, E., 2020. Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. Scientific Reports 10, 18862.

Domínguez-Rodrigo, M., Saladié, P., Cáceres, I., Huguet, R., Yravedra, J., Rodríguez-Hidalgo, A., Patricia, M., Antonio, P., Juan, M., Clara, G., Aramendi, J., Cobo-Sánchez, L., 2019. Spilled ink blots the mind: A reply to Merrit et al. (2018) on subjectivity and bone surface modifications. Journal of Archaeological Sciences 102, 80–86.

Gaudzinski-Windheuser, S., Kindler, L., Rabinovich, R., Goren-Inbar, N., 2010. Testing heterogeneity in faunal assemblages from archaeological sites. Tumbling and trampling experiments at the Early-Middle Pleistocene site of Gesher Benot Ya'aqov (Israel). Journal of Archaeological Sciences 37, 3170–3190.

Harris, J.A., Marean, C.W., Ogle, K., 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, 69–81.

Jiménez-García, B., Abellán, N., Baquedano, E., Cifuentes-Alcobendas, G., Domínguez-Rodrigo, M., 2020a. Corrigendum to "Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars". Journal of the Royal Society Interface 17, 20200782.

Jiménez-García, B., Aznarte, J., Abellán, N., Baquedano, E., Domínguez-Rodrigo, M., 2020b. Deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars. Journal of the Royal Society Interface 17, 20200446.

Landecker, W., 2000. Interpretable Machine Learning and Sparse Coding for Computer Vision. Dissertations and Theses. https://doi.org/10.15760/etd.1936.

McPherron, S.P., Alemseged, Z., Marean, C.W., Wynn, J.G., Reed, D., Geraads, D., Bobe, R., Béarat, H.A., 2010. Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. Nature 466, 857–860.

McPherron, S., Archer, W., Otarola-Castillo, E., Torquato, M., Keevil, T., 2021. Machine learning, bootstrapping, null models, and why we are still not 100% sure which bone surface modifications were made by crocodiles. Journal of Human Evolution 164, 103071.

Moclán, A., Domínguez-Rodrigo, M., 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 11, 4663–4680.

Molnar, C., 2020. Interpretable Machine Learning. Licensed under the Creative Commons Attribution-Non Commercial-Share Alike 4.0 International License, 320 p.

Pineda, A., Saladié, P., Vergès, J.M., Huguet, R., Cáceres, I., Vallverdú, J., 2014. Trampling versus cut marks on chemically altered surfaces: an experimental approach and archaeological application at the Barranc de la Boella site (la Canonja, Tarragona, Spain). Journal of Archaeological Sciences 50, 84–93.

Pineda, A., Cáceres, I., Saladié, P., Huguet, R., Morales, J.I., Rosas, A., Vallverdú, J., 2019. Tumbling effects on bone surface modifications (BSM): An experimental application on archaeological deposits from the Barranc de la Boella site (Tarragona, Spain). Journal of Archaeological Sciences 102, 35–47.

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.

Sahle, Y., El Zaatari, S., White, T.D., 2017. Hominid butchers and biting crocodiles in the African Plio-Pleistocene. Proceedings of the National Academy of Sciences 114, 13164–13169.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8, 25.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114.

Yang, J.H., Wright, S.N., Hamblin, M., McCloskey, D., Alcantar, M.A., Schrübbers, L., Lopatkin, A.J., Satish, S., Nili, A., Palsson, B.O., Walker, G.C., Collins, J.J., 2019. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell 177, 1649–1661.