Using machine learning on new feature sets extracted from 3D models of broken animal bones to classify fragments according to break agent *

Katrina Yezzi-Woodley† Alexander Terwilliger‡ Jiafeng Li ‡
Eric Chen § Martha Tappen¶ Jeff Calder ‡ Peter J. Olver‡

Abstract

Distinguishing agents of bone modification at paleoanthropological sites is at the root of much of the research directed at understanding early hominin exploitation of large animal resources and the effects those subsistence behaviors had on early hominin evolution. However, current methods, particularly in the area of fracture pattern analysis as a signal of marrow exploitation, have failed to overcome equifinality. Furthermore, researchers debate the replicability and validity of current and emerging methods for analyzing bone modifications. Here we present a new approach to fracture pattern analysis aimed at distinguishing bone fragments resulting from hominin bone breakage and those produced by carnivores. This new method uses 3D models of fragmentary bone to extract a much richer dataset that is more transparent and replicable than feature sets previously used in fracture pattern analysis. Supervised machine learning algorithms are properly used to classify bone fragments according to agent of breakage with average mean accuracy of 77% across tests.

Keywords— machine learning, taphonomy, zooarchaeology, bone fragment, marrow exploitation, fracture pattern analysis, hominin-carnivore interactions, subsistence, paleoanthropology, replicability, transparency

1 Introduction

Analyses of bone surface modifications and fracture patterns form the basis of substantial research focusing on early hominin subsistence patterns pertaining to the use of large animal food resources (i.e., meat

*We would like to thank the National Science Foundation NSF Grant DMS-1816917 and the University of Minnesota's Department of Anthropology for funding this research. JC was partially supported by an Alfred P. Sloan Research Fellowship and a McKnight Presidential Fellowship. Source code to reproduce all experimental results is available here: https://github.com/jwcalder/MachineLearningAMAAZE.

†Department of Anthropology, University of Minnesota, yezz0003@umn.edu (corresponding author)
‡School of Mathematics, University of Minnesota
§Wayzata High School, Plymouth, Minnesota
¶Department of Anthropology, University of Minnesota

and marrow). These methods are used to identify the agents of bone modification for the purpose of ascertaining the primary accumulating agent of zooarchaeological assemblages, exploring hominin-carnivore interactions, and determining hominin access order (i.e., hunting versus various forms of scavenging) (e.g. Bartholomew & Birdsell, 1953; Binford, 1981, 1984, 1985; Binford, Bunn, & Kroll, 1988; Binford et al., 1985, 1986; Blumenschine, 1986, 1988, 1989, 1995; Blumenschine et al., 1987; Blumenschine & Cavallo, 1992; Blumenschine, Cavallo, & Capaldo, 1994; Blumenschine & Selvaggio, 1991; Bunn & Ezzo, 1993; Bunn et al., 1986; Dominguez-Rodrigo, 1997; Domínguez-Rodrigo, 2002; Domínguez-Rodrigo & Barba, 2007; Domínguez-Rodrigo, Pickering, Semaw, & Rogers, 2005; Marean, Spencer, Blumenschine, & Capaldo, 1992; Pante, Blumenschine, Capaldo, & Scott, 2012; Pante et al., 2012; Plummer & Bishop, 2016; Pobiner, 2015; Potts, 1983; Selvaggio, 1994a, 1994b, 1998; Shipman, 1983, 1986). These long-standing debates are founded on the premise that the use of large animal food resources was a highly influential factor in our evolution. Conversely, Barr, Pobiner, Rowan, Du, and Faith (2022) downplayed the role of meat consumption altogether, stating that there is no evidence for increased carnivory after the appearance of *Homo erectus*. Clearly, debates continue on the extent to which large animal food resources informed our evolutionary past.

We have been unable to resolve these debates because bone surface modifications and fracture patterns are subject to equifinality which has not been overcome due to the limitations of current methods. This is only exacerbated by concerns over inter- and intra-analyst error and intense disagreement among research groups about the validity of currently used methods (e.g. Domínguez-Rodrigo, Saladié, et al., 2017; Domínguez-Rodrigo et al., 2019; Harris, Marean, Ogle, & Thompson, 2017; James & Thompson, 2015; Merritt, Pante, Keevil, Njau, & Blumenschine, 2019). Improving methods and ensuring that they are replicable could resolve long-standing debates over early hominin subsistence patterns at important paleoanthropological sites such as Dikika (Domínguez-Rodrigo, Pickering, & Bunn, 2010, 2011, 2012; McPherron et al., 2010; Thompson et al., 2015) and FLK Zinj (see Domínguez-Rodrigo, Bunn, & Yravedra, 2014; Pante et al., 2012; Pante, Scott, Blumenschine, & Capaldo, 2015; Parkinson, 2018, and citations contained therein).

Recently, Thompson, Carvalho, Marean, and Alemseged (2019) hypothesized that scavenging for in-bone nutrients may have led to the origin of the Human Predatory Pattern, whereby humans hunt animals larger than themselves. If this is the case, then marrow exploitation may factor into major changes that happened during the Late Pliocene ($3.6 - 2.6$ Ma) and the Early Pleistocene ($2.6 - 1.8$ Ma), such as the first appearance of our genus *Homo* (Bobe & Wood, 2021; Du, Rowan, Wang, Wood, & Alemseged, 2020), the first appearance of stone tools (3.3 Ma) and their subsequent technological advancements (Díez-Martín et al., 2015; Harmand et al., 2015), and geographic expansion (Prat, 2018; Zhu et al., 2018). Thompson et al. call for the development of new approaches and lines of analysis as a necessary step to successfully address these questions. Overcoming current methodological limitations has the potential to open avenues for advancing our understanding of the evolutionary implications of large animal food resource use on our genus, *Homo*.

In response to current methodological challenges, especially as they pertain to resolving equifinality and improving replication, researchers have developed and are continuing to develop various methods for analyzing bone surface modifications and fracture patterns through approaches such as geometric morphometrics (e.g. Arriaza et al., 2017; Courtenay, Huguet, Gonzalez-Aguilera, & Yravedra, 2019; Courtenay, Yravedra, Huguet, et al., 2019; Courtenay, Yravedra, Mate-González, Aramendi, & González-Aguilera, 2019; Maté-González, Courtenay, et al., 2019; Palomeque-González et al., 2017; Yravedra, Aramendi, Maté-González, Austin Courtenay, & González-Aguilera, 2018; Yravedra et al., 2017), confocal profilometry (e.g. Braun, Pante, & Archer, 2016; Gümrükçu & Pante, 2018; Pante et al., 2017; Schmidt, Moore, & Leifheit, 2012), and other digital data extraction methods (e.g. Bello, Verveniotou, Cornish, & Parfitt, 2011; O'Neill et al., 2020; Yezzi-Woodley et al., 2021). Many of these new methods rely on digital imaging, in particular 3D scanning, which has become another prominent avenue of research (e.g. Maté-González, González-Aguilera,

Linares-Matás, & Yravedra, 2019; Yezzi-Woodley, Calder, Olver, Sweno, & Siewert, n.d.) within the field.

The incorporation of digital imaging and data extraction opens opportunities for using powerful computational tools such as machine learning which has been employed in bone modification studies (Arriaza, Aramendi, Maté-González, Yravedra, & Stratford, 2021; Byeon et al., 2019; Cifuentes-Alcobendas & Domínguez-Rodrigo, 2019; Courtenay et al., 2020; Courtenay, Huguet, et al., 2019; Courtenay, Yravedra, Huguet, et al., 2019; Domínguez-Rodrigo, 2019; Domínguez-Rodrigo & Baquedano, 2018; Domínguez-Rodrigo, Fernández-Jaúregui, Cifuentes-Alcobendas, & Baquedano, 2021; Domínguez-Rodrigo, Wonmin, et al., 2017; Jiménez-García, Abellán, Baquedano, Cifuentes-Alcobendas, & Domínguez-Rodrigo, 2020; Jiménez-García, Aznarte, Abellán, Baquedano, & Domínguez-Rodrigo, 2020; Moclán, Domínguez-Rodrigo, & Yravedra, 2019; Moclán et al., 2020; Pizarro-Monzo & Domínguez-Rodrigo, 2020). However, most of these papers have focused on the use of machine learning for discriminating bone *surface* modifications.

The purpose of this paper is to investigate the application of machine learning methods as a means of classifying 3D meshes of bone fragments using new feature sets. Our collection of fragments consists of cervid appendicular long bones, that were broken either by hominins, using hammerstone and anvil, or by carnivores, specifically spotted hyenas. The 3D meshes were created using computed tomography (CT) and the Batch Artifact Scanning Protocol (Yezzi-Woodley et al., n.d.). We introduce new feature sets that provide more detailed information about each bone fragment and thus are more useful for distinguishing agents of bone breakage, and offer more precise data extracted in a highly replicable manner using the virtual goniometer (Yezzi-Woodley et al., 2021), Mesh Lab (Cignoni et al., 2008) and Python (Van Rossum, Guido and Drake, Fred L., 2009). A small set of qualitative features were also incorporated in the data. Together, these data were input into machine learning algorithms to classify each fragment based on the agent of breakage.

We trained machine learning classification models in two different ways. First, we trained the classifiers to classify individual breaks, which we call break-level classification, and the fragment prediction is given by majority voting of the breaks for that fragment. The models are evaluated by their accuracy at classifying fragments, not individual breaks. Second, we trained classifiers to classify the fragments directly, allowing the machine learning algorithm to determine the best way to combine information from the individual breaks. The classification accuracy for the break-level classifiers were only slightly higher than one could expect from random chance ($57.18\% - 66.16\%$). On the other hand, the classification rates of the fragment-level classifiers, which incorporate the entire ensemble of breaks associated with each fragment, were substantially improved ($72.82\% - 79.27\%$) into a statistically meaningful range.

Moclán et al. (2019) published the first paper to apply machine learning to distinguish agents of bone breakage using fracture pattern data (Moclán et al., 2019). They reported near perfect classification rates ($\geq 98^{\circ}$) for some of their testing models. However, we identified several aspects of concern regarding the data they used, the way in which they used it, and their failure to follow basic machine learning protocols. First, there were inconsistencies in the manner in which the data were recorded. Furthermore, it appears that they bootstrapped their sample prior to splitting it into training and test sets; moreover, they based their analysis on global, fragment-level variables, but split the sample at the break-level, which is not permitted in a proper application of machine learning analysis because there are several breaks per fragment. Handling the data in these ways has the effect of contaminating the training set with data from the test set which, in turn, can falsely inflate success rates in classification. As we discuss in detail in Section 4, these errors effectively invalidate their published results. We revisit the Moclán et al. (2019) analysis using both proper machine learning protocols, and, for illustrative and explanatory purposes, their inappropriate use of bootstrapping and global variables, thereby reproducing the latter misleading and overly optimistic results. We also provide the results of an experiment on randomized data showing that both bootstrapping and break-level train-test splits can arbitrarily inflate accuracy, in many cases up to 100%, even when no information is present in the dataset.

When used correctly, machine learning is a powerful tool that has the potential for advancing approaches for analyzing bone modifications and subsequently improving our understanding of early human evolution. Here we demonstrate that using richer data that capture more information about the features from individual breaks – more than has ever been captured before – offers better discriminatory power. These methods can be easily replicated by independent research teams. By comparing our application of machine learning to that of Moclán et al. (2019) we exemplify the ways in which machine learning can be used effectively. Finally, we argue that classifying fragmentary bone produces higher success rates when both global, fragment level and local, break level features are used. The ability to classify individual bone fragments holds promise for improving the resolution with which paleoanthropological sites can be interpreted and furnishes more useful information for interpreting our evolutionary past.

2 Materials and Methods

2.1 Our Experimental Sample

Our experimental sample (see Table 1) consisted of 463 bone fragments (3,218 breaks) from appendicular long bones (humeri, radius-ulnae, femora, tibiae, and metapodia) that were derived from *Cervus canadensis* (elk) ($n = 399$ fragments) and *Odocoileus virginianus* (white-tailed deer) ($n = 64$ fragments). Previous researchers have concluded that metapodia are not as useful for distinguishing agents of bone breakage (Capaldo & Blumenschine, 1994) but this is specific to fracture angles on notches as measured by a contact goniometer. Given that we used new feature sets and the more precise virtual goniometer (Yezzi-Woodley et al., 2021), we chose to include metapodia.

Table 1: Our Experimental Sample

	HSAnv Fragments (Breaks)	***Crocuta crocuta* Fragments (Breaks)**	**Total Fragments (Breaks)**
Cervus canadensis	**275 (1651)**	**124 (987)**	**399 (2638)**
FEM	63 (390)	71 (534)	134 (924)
HUM	28 (159)	27 (241)	55 (400)
MTPOD	71 (409)	0 (0)	71 (409)
UNIDENT LBSF	0 (0)	2 (15)	2 (15)
RAD-ULNA	38 (243)	13 (96)	51 (339)
TIB	75 (450)	11 (101)	86 (551)
Odocoileus virginianus	**0 (0)**	**64 (580)**	**64 (580)**
MTPOD	0 (0)	64 (580)	64 (580)
Total	**275 (1651)**	**188 (1567)**	**463 (3218)**

Abbreviations: femur (FEM), humerus (HUM), metapodial (MTPOD), unidentified long bone shaft fragment (Unindent LBSF), radius-ulna (RAD-ULN), tibia (TIB), hammerstone and anvil (HSAnv).

Of the 463 fragments, 275 (1,651 breaks) were produced by hominins and 188 (1,567 breaks) were produced by carnivores. Hammerstone and anvil were used to break the bones for the hominin sample. The carnivore sample was created by *Crocuta crocuta* (spotted hyenas) at the Milwaukee County Zoo (Wisconsin) and the Irvine Park Zoo in Chippewa Falls (Wisconsin). In some instances, articulated limbs were fed to the hyenas which resulted in two fragments that could not be identified to skeletal element. (See Coil, Tappen, & Yezzi-Woodley, 2017, for details regarding the experimental protocols.)

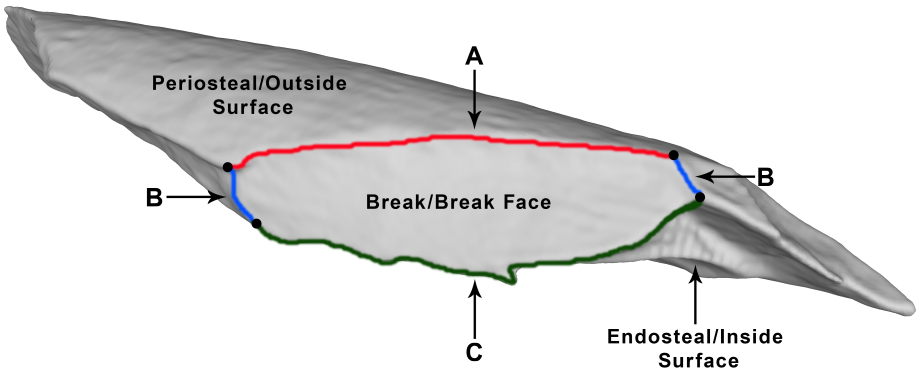


Figure 1: Fragment Features

Fracture (or break) ridges are used to delineate individual breaks. One fracture ridge separates the natural outside surface of the bone from the break surface (A) Two ridges on either side of the break serve as boundaries between adjacent breaks (B). The interior fracture ridge separates the break from the natural interior surface of the bone and, in some cases, other breaks (C).

Fragments were scanned via computed tomography (CT) using the streamlined Batch Artifact Scanning Protocol (Yezzi-Woodley et al., n.d.) that we developed to acquire 3D models of each of the fragments, which were stored as *.ply* files. Data were extracted from the 3D models of the bone fragments manually through the Graphical User Interface in Meshlab and automatically using Python scripts.

To know how each feature was extracted from the fragments, it is important to understand how we defined the features and in particular how we differentiated breaks, because, as researchers have previously acknowledged, identifying breaks is not always a straight forward task (Biddick & Tomenchuk, 1975; Bunn, 1982, 1989; Davis, 1985; Pickering, Domínguez-Rodrigo, Egeland, & Brain, 2005). As Bunn (1982, p.43) points out, the boundary between breaks is neither well-defined nor is it always obvious. Given a long bone, there is the natural outside (periosteal) surface of the bone and the natural inside (endosteal) surface of the bone. Fragments from broken long bones also have surfaces that expose what is referred to as breaks or break faces. Each break is surrounded by fracture (or break) ridges. One of those ridges constitutes the boundary between the periosteal surface and the break surface for that break (see Figure 1A). Two of the ridges connect adjacent break faces (see Figure 1B). The fourth ridge is the boundary between the break face and the endosteal surface of the bone fragment(see Figure 1C). In cases where the break face overlaps another break face(s) without extending through the entire thickness of the bone, the fourth ridge serves as a boundary between this break and the more internal break(s) (see Figure 2). Break curves can be extracted from those ridges (see Figure 3). In more complicated scenarios, the boundary of the break face may contain additional ridges of various types. It should be noted that ridges are not always easily detectable and one must decide at which scale to accept a ridge as a boundary and break surface as a separate face. Given the precision of the tools we used to extract data, we were able to accept small breaks as separate breaks. Though further discussion is necessary in the field to standardize the definition of an individual break, the extraction methods used here provide sufficient detail in the data to begin such research.

Angle measurements were taken along the exterior fracture ridge of each break on each fragment with the virtual goniometer (Yezzi-Woodley et al., 2021) using radii between $1-3$ using geodesic distance based on the size of the break (See Figure 4). When an angle measurement is taken, a colorized patch appears on the mesh. The center of subsequent angle measurement was placed on the fracture ridge at the border of the colorized patch indicated by the previous angle measurement. Therefore, measurements were $1-3$

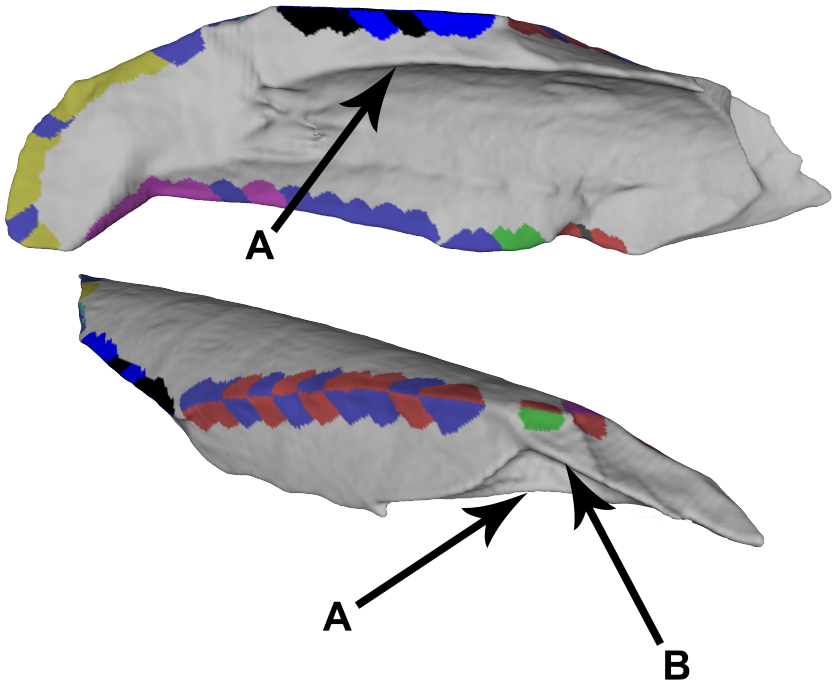


Figure 2: The Interior Ridge

The interior edge of some breaks border the endosteal surface (A) while others border another break (B).

geodesic units apart. When shifting to the next break, the patch colors change (see Figure 5). The endpoints of each break were chosen using the "Picked Points" tool in Meshlab. The angle and endpoint data were used to calculate the arc angle (see Figure 6) and break length variables (see Figure 7).

Additional variables were collected manually through observation of the models such as the presence or absence of notches (see Figure 8) or trabecula. We ran the models through a Python script to extract more global features such as the mesh volume and surface area. Data were recorded and calculated for each break, referred to here as break level data, and for the entire fragment, referred to here as fragment level data. A summary of these variables are as follows:

Break Level Variables:

1. **Number of Angles:** This refers to the number of fracture angle measurements per break. Fracture angle measurements were taken along each break curve. A minimum of one fracture angle measurement was recorded for each break. Type: natural number

2. **Angle data:** Because more than one angle measurement could be taken on each break curve, summary statistics were calculated for the angle measurements. This included the minimum, maximum, mean, median, standard deviation, and range. Type: continuous

3. **Interior Edge:** If the interior ridge of the break face transitioned to another break face it was categorized as "break". If it adjoined the endosteal surface, then it was categorized as "endosteal" (see

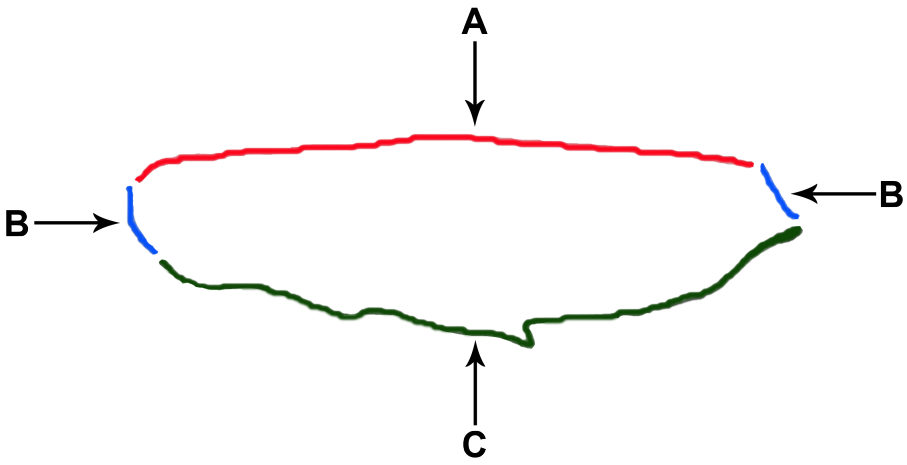


Figure 3: Break Curves

Break (or fracture) curves can be extracted from the fracture ridges surrounding a break face. These break curves correspond to the break ridges illustrated in Figure 1. The exterior curve separates the periosteal surface from the break surface (A) Two curves separate the break from adjacent breaks (B). One curve is the interior break curve (C).

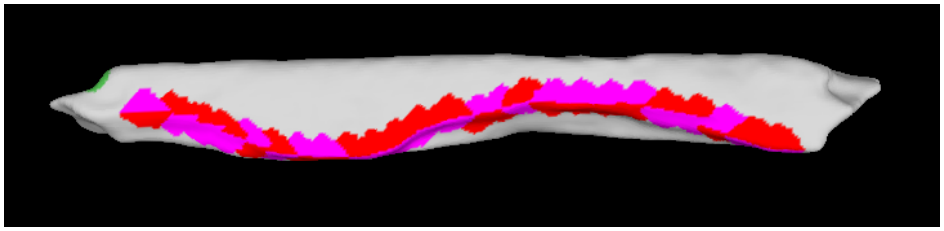


Figure 4: Angle Measurements Along Ridge

Angle measurements were taken along the entire fracture ridge between the natural outside surface of the bone and the break surface.

Figure 2). Type: Boolean

4. **Interrupted:** This is a TRUE/FALSE variable indicating whether or not the break curve was interrupted by another break. Type: Boolean

5. **Ridge Notch:** If the fracture ridge exhibited the arcuate indentation characteristic of a notch, then the break was classified as "TRUE" (see Figure 8A). Type: Boolean

6. **Interior Notch:** If the interior ridge of the break face exhibited the arcuate indentation(s) characteristic of a notch(es), then the break was classified as "TRUE" (see Figure 8B). Type: Boolean

7. **Break Lengths:** Two measures of break length were recorded. We calculated the Euclidean distance between the two endpoints of the break curve (see Figure 7C). When using the virtual goniometer, the location of each angle measurement is automatically recorded. Using those points in conjunction with the endpoints, we calculated the arc length of each break curve (see Figure 7D). Type: continuous

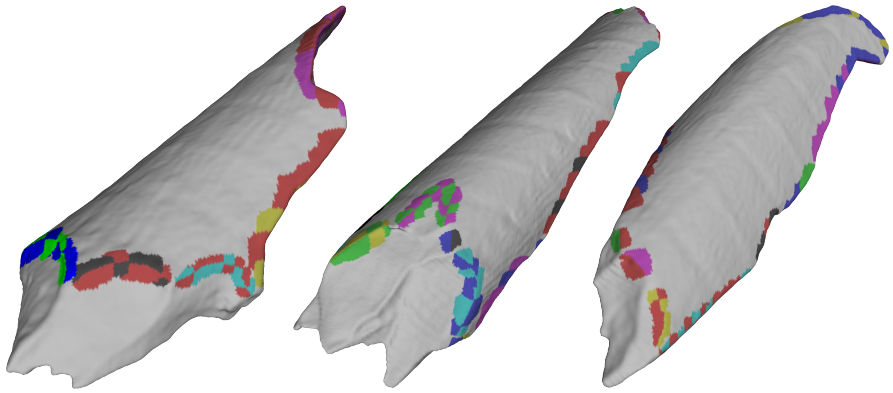


Figure 5: The Virtual Goniometer Captures More Detail

The virtual goniometer makes it possible to capture more information with a higher degree of detail. This includes the ability to measure small breaks. When transitioning from one break to the next, the colors of the patches where the angle measurements are taken change.

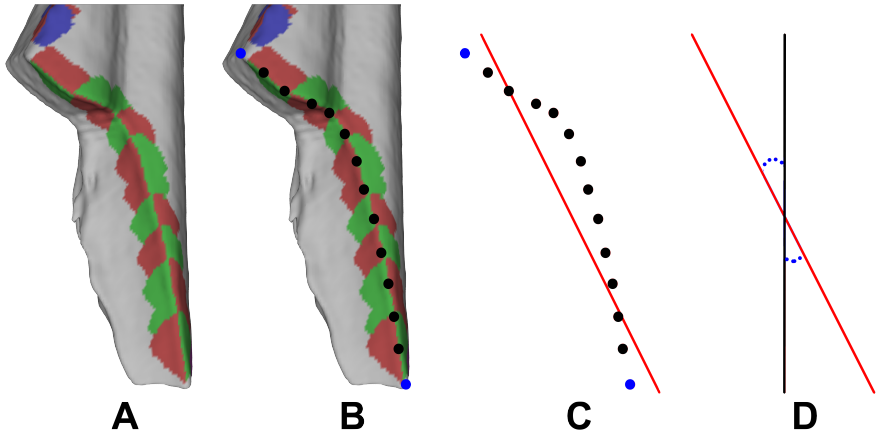


Figure 6: Arc Angle

For each break curve (A) the $x-$, $y-$, and $z-coordinates$ for both the endpoints (blue circles) and the locations of each angle measurement (black points) were recorded (B). The points were used to define a best fit line (C). The angle between the best fit line and the principal axis of the fragments was recorded (D).

8. **Arc Angle:** This is a calculation that we used in lieu of break plane, as defined by Gifford-Gonzalez (1989). We did this by calculating a best fit line to the ordered points along the break curve and then calculating the angle between the best fit line and the principal axis of the bone fragment. Again, the points were taken from the selected endpoints and the virtual goniometer data (see Figure 6). Type: continuous

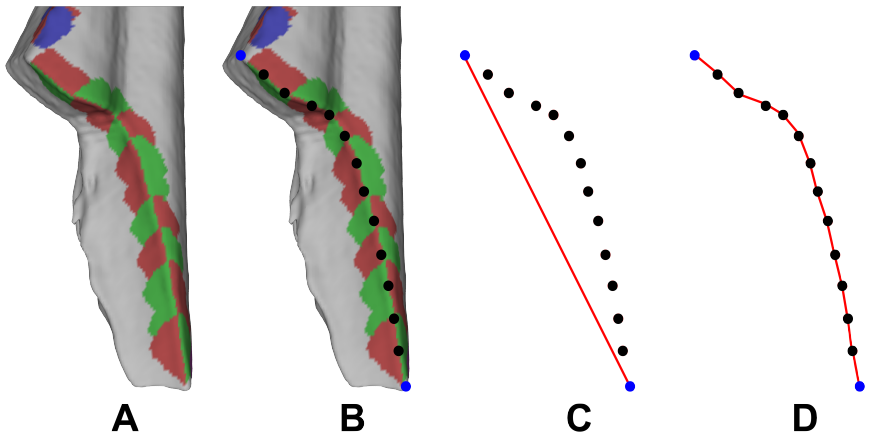


Figure 7: Break Lengths

For each break curve (A) the $x-$, $y-$, and $z-coordinates$ for both the endpoints (blue circles) and the locations of each angle measurement (black points) were recorded (B). The straight line (Euclidean) distance was measured between the endpoints of the break curve (C) and the arc length was measured using all the points (D).

Fragment Level Variables:

1. **Number of Breaks:** We recorded the number of break faces per fragment. Type: natural number
2. **Trabecula:** If there was trabecular bone on the fragment it was categorized as "TRUE". Type: Boolean
3. **Volume:** The volume of the domain bounded by the surface mesh was extracted in Python. Type: continuous
4. **Surface Area:** The surface area of the mesh was extracted in Python. Type: continuous
5. **Bounding Box Dimensions:** The bounding box dimensions were extracted using Python. This can be thought of as the fragment length, width, and depth. Type: continuous
6. **Angle Data:** The summary statistics were calculated from the summary statistics of the fracture angle data calculated at the break level. We chose to do this as opposed to summarizing the original angle data because we did not want each individual angle measurement to be weighted equally. We wanted the angle data to be weighted by how the angles were summarized for each break. Type: continuous
7. **Interior Edge:** The number of break faces with interior edges adjacent to another break were tallied per fragment as were those that were adjacent to the endosteal surface. Type: natural number
8. **Interrupted:** The number of break faces that were interrupted were tallied per fragment. Type: natural number
9. **Ridge Notch:** The number of break faces with fracture ridges classified as notches were tallied. Type: natural number
10. **Interior Notch:** The number of break faces that had notching on their interior ridge were tallied. Type: natural number

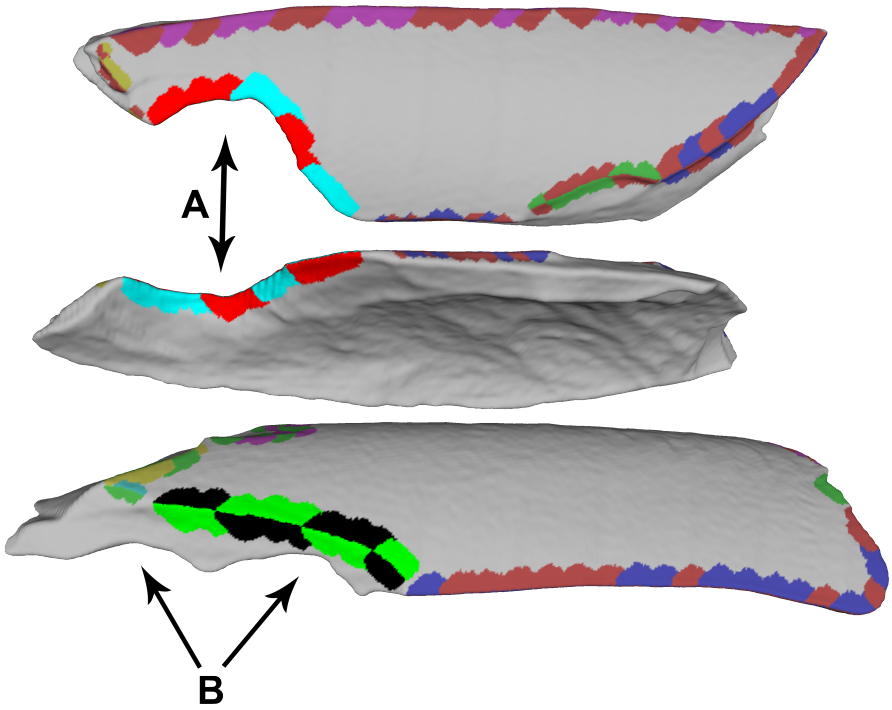


Figure 8: Notches

Notches are arcuate indentations on the bone created at the location of direct impact between the bone and the object used to break it. Some of the breaks had notches on the exterior fracture ridge (A). Some notches were found on the interior fracture ridge (B)

11. **Break Lengths:** Summary statistics were calculated for both measures of break curve length (Euclidean distance and arc length). Type: continuous

12. **Arc Angle:** Summary statistics were calculated for the arc angles of the break curves. Type: continuous.

2.2 Methods

Bone fragments were categorized using 7 different machine learning algorithms: random forest, linear support vector machine, support vector machine using the radial basis function, neural network, linear discriminant analysis, Gaussian naive Bayes, and k-nearest neighbor. High level descriptions of the machine learning methods we used here can be found in Yezzi-Woodley (2022, Chapter 4). For more detailed information about classical machine learning methods, we refer the reader to Bishop (2006), and for more information about deep learning and neural networks, we refer to LeCun, Bengio, and Hinton (2015). The code for all our experimental results can be found on GitHub.[1]

Data were split into training (75%) and testing (25%) sets. This was done at the fragment level for all tests so as to avoid contaminating the training set with data from the test set. This means that for the break level tests, 25% of the fragments were marked for the test set, and the test set was then populated by those

[1]Source Code: https://github.com/jwcalder/MachineLearningAMAAZE

fragment's breaks, ensuring that all breaks from a single fragment were either in the training set or the testing set. Because the train-test split was done at the fragment level, when classifying breaks, the breaks voted on which labels each fragment should receive based on the statistical mode of their predicted labels. Ties were broken at random. The accuracy reported for break-level tests is therefore the percentage of fragments that were assigned the correct label by their breaks for that algorithm in question. We emphasize that we *never* use information from the testing set when training any of the machine learning algorithms. As is standard in machine learning, the testing set must be kept independent of *all steps* in the model training procedure, so that the testing accuracy can give an unbiased evaluation of model performance on new data that was not seen during training.

Each test was repeated 300 times with a new train and test set computed from the original data-set. The mean accuracy across all repetitions was recorded as well as the standard deviation.

3 Our Results

The results of classifying bone fragments using break-level classifiers were only slightly above what can be expected from random chance (50%). The mean accuracy ranged from $57.18\% - 68.34\%$ with standard deviations ranging from $4.22\% - 5.06\%$ (see Table 2). In particular, the mean accuracies are all lower than the 69.8% accuracy we report from unsupervised learning in Section 3.1 below, indicating that there is very little information useful for classification in the break-level dataset.

Table 2: Machine Learning Results
(Break Level)

Algorithm	Mean Accuracy	Standard Deviation
Random forest (RF)	68.34%	4.24%
Support vector machine (SVM) – linear	62.60%	4.46%
Support vector machine - RBF	66.16%	4.22%
Neural network (NN)	65.30%	5.06%
Linear discriminant analysis (LDA)	64.47%	4.30%
Gaussian naive Bayes (GNB)	57.18%	4.76%
k-nearest neighbor (KNN)	65.19%	4.23%

300 repetitions

On the other hand, when we trained the machine learning classifiers at the fragment-level, giving the models access to summary statistics about each fragment's constituent breaks, the classification accuracy improved substantially. The mean accuracy across tests ranged from $72.82\% - 79.27\%$ with lower standard deviations ($3.42\% - 3.94\%$) (see Table 3). These results are substantially higher than the unsupervised results (69.8%) in Section 3.1, indicating that the machine learning methods are learning from the labeled information in a significant way.

Table 3: Machine Learning Results
(Fragment Level)

Algorithm	Mean Accuracy	Standard Deviation
Random forest (RF)	77.18%	3.51%
Support vector machine (SVM) – linear	77.24%	3.48%
Support vector machine - RBF	79.27%	3.42%
Neural network (NN)	77.95%	3.56%
Linear discriminant analysis (LDA)	76.19%	3.59%
Gaussian naive Bayes (GNB)	72.82%	3.94%
k-nearest neighbor (k-NN)	77.61%	3.51%

300 repetitions

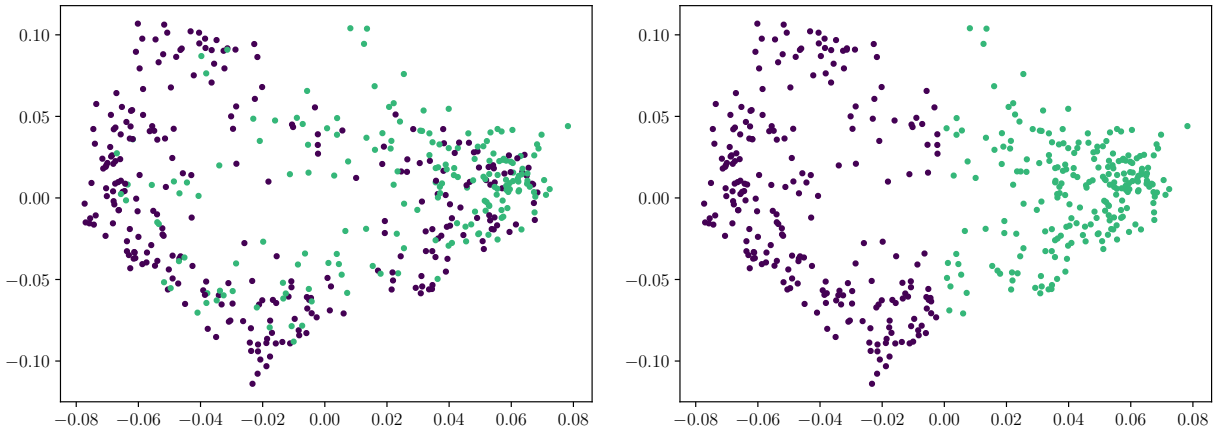
We did not tune hyperparameters for any of the methods. Hyperparameter optimization, with an appropriately chosen validation set, could have the potential to slightly improve results. For k-nearest neighbor, we fixed the value $k = 25$, which worked well for our replication of Moclán et al. 2019 discussed below. For the neural network, we used a fully connected neural network with three hidden layers with 100, 1000, and 5000 hidden nodes in each layer, respectively. We trained the network with the Adadelta (Zeiler, 2012) optimizer with batch size of 32, initial learning rate of 1 with scheduled decreases by 10% every epoch, and we trained the network over 100 epochs. To prevent overfitting, we used dropout layers (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) with dropout rate of 0.4 between the hidden layers of the neural network. We refer the reader to LeCun et al. (2015) for more details on deep learning.

3.1 Unsupervised Learning

In order to visualize our dataset and further explore its structure, we consider here the application of *unsupervised learning algorithms*. Unsupervised learning is a form of machine learning that does not utilize the labels of data points during its training process, and can include algorithms like clustering, dimensionality reduction, and ranking. Here, we used a spectral embedding for dimensionality reduction, and spectral clustering to detect clusters in the dataset. Spectral embeddings offer a way to embed a high dimensional dataset into a low dimensional space that is superior to linear techniques like principal component analysis (PCA). Spectral embeddings build a graph over the dataset based on similarities between datapoints, and the embedding into k dimensions involves computing the first k eigenvectors of the graph Laplacian. Spectral clustering clusters the data by running the k-means clustering algorithm on the embedded k-dimensional data. We refer to Von Luxburg (2007) for a tutorial on spectral clustering; we use the specific spectral clustering algorithm proposed in Ng, Jordan, and Weiss (2001).

In Figure 9a we show the spectral embedding of our fragment-level dataset into $k = 2$ dimensions, with the points colored by their true labels. We can see a small degree of separation between the classes, though there is significant overlap. In Figure 9b we show the labels predicted for each point by spectral clustering, which achieved 69.8% classification accuracy. In particular, the hominin broken fragments were classified at 67.3% accuracy, while the carnivore broken fragments were classified at 73.4% accuracy. These unsupervised accuracy values should be viewed as baseline accuracy scores that our fully supervised learning results can be compared to.

In Figure 10a we show the spectral embedding of our break-level dataset into $k = 2$ dimensions, with the points colored by their true labels. We see a clear cluster structure here with three well-separated clusters.



(a) Spectral Embedding

(b) Spectral Clustering

Figure 9: Spectral Clustering on Fragment-Level Data

In (a) we show the spectral embedding of our fragment-level dataset with the points colored based on their true labels of hominin or carnivore. In (b) we show the results of spectral clustering, which runs the k-means clustering algorithm on the spectral embedding. The accuracy of the spectral clustering in (b) is 69.8%.

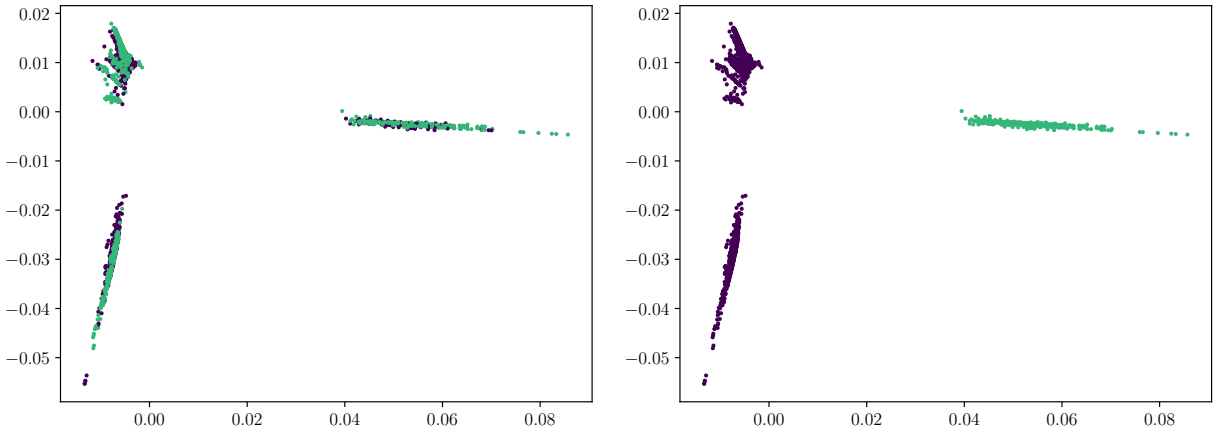
However, the clusters do not correspond with the classes hominin and carnivore. As a result of this, the spectral clustering on the break-level dataset achieved a total accuracy of 53.5%. Specifically, the hominin breaks were classified at 92.7% accuracy, while the carnivore breaks were classified at 12.2% accuracy. These results suggest that the break-level information is useful for classification only when it is compiled through summary statistics at the fragment level, and that considering information on a break-by-break basis yields less useful information for classification.

4 Comparing our Results to Moclán et al. 2019

In this section, we revisit the machine learning classification results in Moclán et al. (2019), and reanalyze their data using the preceding methods. We point out significant issues with their applications of machine learning and show that a correct application does not produce the seemingly impressive results they find. We further compare their data and analysis with ours, as discussed in the preceding section. Finally, we present the results of an experiment with randomized data showing that the issues we identified in Moclán et al. (2019) can arbitrarily inflate accuracy scores even when no patterns are present in the data.

4.1 The Moclán et al. 2019 Sample

In our analysis of the results in Moclán et al. (2019), we used their published dataset, which is provided as a *.csv* file in their supplemental information. According to the published *.csv* file, their sample consists of a total of 1,488 breaks comprised of 797 anthropogenic breaks, 177 breaks created by *Crocuta crocuta* and 514 breaks created by *Canis lupus*. According to the text in the main article the hyena sample consists of 66 bones and the wolf sample consists of 237 fragments. The anthropogenic sample was derived from 40 bones (10 humeri, 10 radii-ulnae, 10 femora, and 10 tibiae) that were broken, resulting in 1,497 fragments of which they selected 332. It should be noted that in the first paragraph of their Results Section, they report



(a) Spectral Embedding

(b) Spectral Clustering

Figure 10: Spectral Clustering on Break-Level Data

In (a) we show the spectral embedding of our break-level dataset with the points colored based on their true labels of hominin or carnivore. In (b) we show the results of spectral clustering, which achieved a total accuracy of 53.5%.

a total of 881 hominin produced breaks, 202 hyena produced breaks, and 610[2] breaks produced by wolves. It is not clear why there is a discrepancy between what is written in the main article and what is presented in the supplemental information. Dropping the transverse breaks from analysis does not account for this discrepancy (See Table 4).

Table 4: The Moclán et al. 2019 Sample

	Hominin	Hyena	Wolf	Total
Fragments reported in text	332	66	237	635
Breaks reported in text	881	202	610	1693
Breaks reported in SI	797	177	514	1488
Difference in reported breaks	84	25	96	205
longitudinal breaks reported in text	297	91	287	675
longitudinal breaks reported in SI	284	91	267	642
Difference in reported longitudinal breaks	13	0	20	33
oblique breaks reported in text	549	87	273	909
oblique breaks reported in SI	513	86	247	846
Difference in reported oblique breaks	36	1	26	63
transverse breaks reported in text	35	24	50	109
transverse breaks reported in SI	0	0	0	0
Difference in reported transverse breaks	35	24	50	109

[2]There is a typo. It is written as 61. However, once the breaks are summed for each fracture plane it is evident that this should be 610.

Their sample consisted of fragments from animals that weighed $80-200$kg. The fragments from the anthropogenic sample all came from *Cervus elaphus* (red deer). The carnivore samples, which were gathered from a hyena den in Tanzania and a natural park in Spain, included unidentifiable fragments that were not metapodial fragments. They chose not to include metapodia, stating that they were not diagnostic due to the thick cortical bone, citing Capaldo and Blumenschine (1994). The wolf-created sample included fragments from *Cervus elaphus* and *Sus scrofa*. Fragments they chose were ≥ 4 cm in maximum dimension and bore, at minimum, one measurable break.

Moclán et al. (2019)'s dataset contains 12 variables: Epiphysis, Length, Interval, Number of Planes, Fracture Plane, Plane (Fracture Angle), Type of Angle, >4 cm, Notch, Notch A, Notch B, and Notch D (See Table 7). Epiphysis refers to the presence/absence of some or all of the epiphyseal surface on the fragment. Length refers to the measured length (mm) of the fragment. The Length category is derived by parsing the fragments into bins based on their measured lengths. Number of Planes refers to the number of measurable breaks on the fragment, including transverse breaks. Fracture Plane measures the angle that the fracture plane makes with the longitudinal axis of the bone, as defined by Gifford-Gonzalez (1989). Though transverse breaks were included in the Number of Planes, only breaks that were longitudinal and oblique were input into the machine learning algorithms. In this count, we interpret "Plane" to mean the fracture angle, i.e., the measured angle of transition between the periosteal surface and the break surface along each break curve. In previous studies, the angle was taken at the center along the edge of the break (Alcántara-García et al., 2006; Coil et al., 2017; Pickering et al., 2005). Here the authors state that it is "measured at the point of maximum angle. In cases where both acute and obtuse angles are present, the latter was used" (Moclán et al., 2019, p.3). "Maximum angle" suggests that the largest angle value was used. However, given the caveat that obtuse angles were used in instances where both acute and obtuse angles were present on a break, the meaning of maximum in this case may refer to the distance from $90°$. We assume the measurement was taken with a handheld, contact goniometer; therefore assessing where to take the measurement was likely done, in large part, by eye. (In contrast, our use of the virtual goniometer makes our angle measurements both more accurate and completely reproducible.) Type of angle is derived from parsing the fragments into bins based on their measured angles, where angles $<85°$ are categorized as "acute", angles between $85°-95°$ are categorized as "right", and angles $>95°$ are categorized as "obtuse". We interpret ">4 cm" as meaning the presence or absence of breaks on the fragment that are greater than 4cm in length. The last four variables identify the presence or absence of notches (in general) on the fragment and the notch types: A, C, and D. (See Capaldo & Blumenschine, 1994, for details on notch types).

Table 5: Comparison of Sample Sizes

	PERCUSSION	*Crocuta*	*Canis*	TOTAL
MOCLAN	332 (797)	66 (177)	237 (514)	637 (1,488)
OUR EXPERIMENTAL SAMPLE	275 (1,651)	188 (1,567)	0(0)	463 (3,218)

THE FIRST VALUE IS THE NUMBER OF FRAGMENTS. THE VALUE IN PARENTHESES IS THE NUMBER OF BREAKS.

4.2 Replicating Moclán et al. 2019's Machine Learning Analysis

Moclán et al. (2019) applied six different algorithms (neural networks, support vector machines, k-nearest neighbor, random forests, mixture discriminant analysis, and naive Bayes) to their dataset. They ran these tests with and without bootstrapping (1,000 times) the raw data (see also Moclán et al., 2020, p. 7). They separated both the original dataset and the bootstrapped dataset into a $70/30$ training/testing split. It should be emphasized that bootstrapping prior to splitting the sample into training and test sets is not allowed

in machine learning applications because it contaminates the training set with test data; see Section 5 for further details. The classification success rate for the original sample ranged between $82\% - 89\%$. The classification rates for the bootstrapped sample ranged from $78\% - 94\%$. They separated out the breaks according to fracture planes and whether or not the breaks were greater or less than 90°. When applied to the longitudinal fractures with fracture angles $< 90^\circ$, the classification rates on the original sample were between $75\% - 83\%$ and the classification rates for the bootstrapped samples were between $73\% - 99\%$. The classification rates for the longitudinal fractures with fracture angles $> 90^\circ$ showed a success rate of $72\% - 82\%$ for the original sample and $81\% - 98\%$ for the bootstrapped sample. For the oblique fracture with fracture angles $< 90^\circ$ classification rates ranged from $68\% - 86\%$ for the original sample and $69\% - 98\%$ for the bootstrapped sample. Finally, for oblique fractures with fracture angles $> 90^\circ$, classification rates ranged between $86\% - 90\%$ and $89\% - 96\%$ (see Table 6).

Table 6: Summary of Moclán et al. 2019's ML Results

	ORIGINAL	BOOTSTRAPPED
ALL	$82\% - 89\%$	$78\% - 94\%$
LONGITUDINAL $< 90^\circ$	$75\% - 83\%$	$73\% - 99\%$
LONGITUDINAL $> 90^\circ$	$72\% - 82\%$	$81\% - 98\%$
OBLIQUE $< 90^\circ$	$68\% - 86\%$	$69\% - 98\%$
OBLIQUE $> 90^\circ$	$86\% - 90\%$	$89\% - 96\%$

We replicated their machine learning approach on the entire dataset provided in their supplemental information. Because they did not include specimen information in their dataset, we were unable to replicate our method of splitting by fragment to ensure breaks from the same fragment were not contaminating the testing set. For their dataset, we used repeated k-fold cross validation to ensure each data point was included in the test set at least once for each replication.

We applied random forest, linear support vector machine, neural network, linear discriminant analysis, Gaussian naive Bayes, and k-nearest neighbor machine learning algorithms to their dataset. Our use of linear discriminant analysis was a substitution of their mixture discriminant analysis, and we do not expect major differences in algorithm performance.

As in Moclán et al. (2019), we ran the test with and without their inappropriate bootstrapping protocol. However, in the discussion section, we will elaborate on why it is not appropriate to use bootstrapping in this manner when applying machine learning methods, and we chose to run the bootstrapped version here purely for comparison with Moclán et al. 2019's work and as a tool for discussion. Additionally, it is important to bear in mind the discrepancies in reported sample sizes mentioned previously in so much that we are making the assumption that they ran the machine learning algorithms on the samples as provided in the supplemental *.csv* file, which could explain any discrepancies with the results we report here.

Unlike Moclán et al. (2019), we did not run tests using subsets of the data based on break plane and fracture angle. This is unnecessary when using machine learning which can parse out which features and relationships among features are useful for classification. Subsetting the data in this way reduces the sample size which exacerbates the issues stemming from the mixing of test data into the training data and the data recording errors.

As noted above, we have some concerns about Moclán et al.'s data. Some of the variables were corrupted due to what appears to be recording errors. For instance, Epiphysis is a Boolean (present/absent) variable, but 264 observations were categorized with a 2, while 78 observations were categorized as a 3, plus 37 observations were categorized as a 4, while 233 observations were categorized as "present", and 876 observations were categorized as "absent". Likewise, Notch A and Notch C are Boolean variables and in

addition to "present"/"absent", contained a third value "2". Interval length and the type of angle are redundant variables. Indeed, the information contained in these variables is provided by the measured lengths and the measured angles and are therefore unnecessary (see Table 7).

Table 7: Summary of Moclán et al. 2019's Variables

Variable	Level	Type	Entered Values	Notes
Epiphysis	Frag	Boolean	2, 3, 4, Absent, Present	Corrupted
Length (mm)	Frag	Numerical	Whole numbers ranging from 40-267	–
Interval (length)	Frag	Categorical	Bins: 40-49...90-99, 100-149, 150-199, >199	Redundant
Number of planes	Frag	Count	Whole numbers ranging from 1-6	Transverse breaks included in counts
Fracture plane	Break	Boolean	Longitudinal, Oblique	–
Plane/Fracture angle	Break	Numerical	Whole numbers ranging from $20^\circ - 161^\circ$	–
Type of angle	Break	Categorical	Acute ($< 85^\circ$), Right ($85 - 95^\circ$), Obtuse ($> 95^\circ$)	Redundant
$>$ 4cm	Frag	Boolean	Absent, Present	–
Notch	Frag	Boolean	Absent, Present	–
Notch a	Frag	Boolean	2, Absent, Present	Corrupted
Notch c	Frag	Boolean	2, Absent, Present	Corrupted
Notch d	Frag	Boolean	Absent, Present	–

Of additional concern are the levels at which the data were collected. Some of the variables were collected at the fragment level. However, the training/testing splits were made at the break level which, as noted above, has the potential for contaminating the training set with data from the test set. It is possible to use data with variables at different levels. However, the data must be split into training/testing splits at the highest level, in this case the fragment level. We will elaborate on this problem in the discussion section.

Given these data challenges, we ran additional tests on their data, but, first we dropped all corrupted, redundant, and fragment-level variables. We were unable to use the fragment level variables here because the *.csv* file did not identify from which fragment each break was derived so we could only split the sample at the break level. Cleaning the data reduced the variables to break plane and fracture angle. In the first iteration, we maintained the three groups in order to compare the results of the properly run machine learning test against the results of the previous tests. We then pooled the carnivore samples in order to compare the results with our experimental sample.

Despite the inherent issues with the dataset, we were able to run the machine learning algorithms on the entire dataset with and without bootstrapping to offer a point of comparison between the results that Moclán et al. (2019) achieved and the results one can expect to achieve when the data are cleaned and the machine learning algorithms are properly applied. When bootstrapping was used we achieved successful classifications rates ranging from $86.93\% - 95.52\%$ with standard deviations ranging from $1.33\% - 2.35\%$ (see Table 8). These results are similar to that achieved by Moclán et al. (2019).

It is expected that the Random Forest algorithm will perform well on bootstrapped data, as it is able to leverage the duplication contamination of the testing dataset due to the nature of decision tree classifiers. The k-nearest neighbor algorithm also performs well when we set $k = 1$, which outperforms all other k for this level of bootstrapping. This is precisely because the bootstrapping replicates datapoints and ensures that every datapoint appears multiple times in both the training and testing set, so the nearest neighbor is always

the duplicated point with the correct label. Without bootstrapping, classification rates dropped, ranging from 80.74% – 87.83% with standard deviations ranging from 2.60% – 3.15%; see Table 9. Again, these results are similar to those achieved by Moclán et al. (2019).

Table 8: ML Results Using Moclan's Data
(With Bootstrapping)

Algorithm	Mean Accuracy	Standard Deviation
Random forest (RF)	95.52%	1.33%
Support vector machine (SVM) – linear	87.80%	2.12%
Neural network (NN)	87.56%	2.08%
Linear discriminant analysis (LDA)	86.93%	2.10%
Gaussian naive Bayes (GNB)	81.24%	2.35%
k-nearest neighbor (k-NN)	94.25%	1.47%

300 repetitions, 10 folds, 1,000 boostraps

Table 9: ML Results Using Moclan's Data
(Without Bootstrapping)

Algorithm	Mean Accuracy	Standard Deviation
Random forest (RF)	87.83%	2.60%
Support vector machine (SVM) – linear	86.60%	2.60%
Neural network (NN)	82.57%	3.03%
Linear discriminant analysis (LDA)	86.16%	2.71%
Gaussian naive Bayes (GNB)	80.74%	3.15%
k-nearest neighbor (k-NN)	82.32%	3.04%

300 repetitions, 10 folds

Due to the fragment level contamination of the data, Random Forest is still expected to do well. We used $k = 25$ for the k-nearest neighbor algorithm, which did decently, but could potentially be optimized by modifying the value of k. We used a neural network with three hidden layers of size 100, 200, and 400, respectively, and trained it in a similar way as we described earlier in the paper.

These results are extremely appealing, however, given the errors in data collection and the application of bootstrapping, the results are unreliable. Therefore, we cleaned the data by dropping all the variables that had recording errors and dropping all the fragment level variables. We were not able to incorporate fragment level variables because the published dataset does not contain fragment labels, and thus we were unable to divide the set into training and test sets at the fragment level and we were unable to classify at the fragment level.

We ran the machine learning algorithms initially with three labels: hominin, hyena, and wolf. Then we repeated the process after pooling the hyena and wolf into a carnivore class so that we could make direct comparisons with the results we achieved using our experimental data.

Using three labels, one can expect a classification accuracy of approximately 33% by random guessing. We achieved classification rates between 53.74% – 57.59% with standard deviations ranging from 3.73% – 4.29% (see Table 10). Though slightly better than what can be expected with random choice, the proper application of machine learning results, not surprisingly, in a dramatic decline in the overall accuracy.

Table 10: ML Results Using Moclan's Data
(Break Level Only - 3 Actors)

Algorithm	Mean Accuracy	Standard Deviation
Random forest (RF)	55.00%	3.83%
Support vector machine (SVM) – linear	53.74%	4.29%
Neural network (NN)	58.33%	3.92%
Linear discriminant analysis (LDA)	57.59%	3.77%
Gaussian naive Bayes (GNB)	54.92%	4.05%
k-nearest neighbor (k-NN)	56.08%	3.73%

300 repetitions, 10 folds

Without the fragment level contamination to leverage, the Random Forest algorithm no longer leads in mean accuracy, and other algorithms that aren't based on Decision Tree methods overtake it.

When carnivores were pooled the classification mean accuracy ranged from 59.25% – 64.21% with standard deviations ranging from 3.70% – 4.41% (see Table 11). Our mean accuracy, using only break-level data, was, on average, a bit higher (64% as opposed to 61%), had a slightly wider range (57.18 – 68.64%) and slightly higher standard deviations (4.22% – 5.06%). Using both break level and fragment level variables to classify fragments, our classification rates improved with mean accuracy ranging from 72.82% – 79.27% and lower standard deviations 3.42% – 3.94%

Table 11: ML Results Using Moclan's Data
(Break Level Only - 2 Actors)

Algorithm	Mean Accuracy	Standard Deviation
Random forest (RF)	61.19%	3.70%
Support vector machine (SVM) – linear	61.20%	4.41%
Neural network (NN)	64.21%	3.76%
Linear discriminant analysis (LDA)	62.80%	3.78%
Gaussian naive Bayes (GNB)	59.25%	3.73%
k-nearest neighbor (k-NN)	61.63%	3.73%

300 repetitions, 10 folds

4.3 An experiment with randomized data

In order to further illuminate the issues we have identified from Moclán et al. (2019) with bootstrapping and break-level train-test splits, we applied machine learning algorithms to a random dataset that we constructed of a similar size to the dataset used in Moclán et al. (2019). Our random synthetic dataset has 200 fragments, each with 7 breaks, yielding 1400 breaks, which is comparable to the 1488 used in Moclán et al. (2019). Each fragment is assigned 34 random numerical features, and each break is assigned 6 random numerical features. The number of features is similar to the number of break and fragment-level features used in Moclán et al. (2019), after the categorical features are converted to numerical features through one-hot encodings. This yields a dataset of 1400 breaks, each of which has 40 numerical features (34 fragment-level and 6 break-level). Each fragment is then assigned a label of 0 or 1 uniformly at random, and that label is

transferred to the break. We emphasize that this dataset is constructed completely at random, so there is no information in the dataset from which a machine learning algorithm can learn. Any proper application of machine learning should achieve on average 50% classification accuracy.

Table 12: Results of the randomized machine learning experiment.

Algorithm	Break-Level Split	Frag-Level Split with Bootstrapping	Frag-Level Split
LDA	65.1 (4.1)	96.7 (4.1)	50.1 (6.9)
Random Forest	100.0 (0.0)	100.0 (0.0)	50.7 (7.3)
Linear SVM	66.5 (3.7)	100.0 (0.0)	50.2 (7.4)
RBF SVM	99.6 (0.7)	100.0 (0.0)	49.6 (7.0)
Nearest neighbor	100.0 (0.1)	100.0 (0.0)	50.0 (6.4)
Neural Network	63.3 (3.8)	100.0 (0.0)	50.7 (5.8)

100 repetitions, standard deviation in parentheses

We applied six machine learning algorithms to this dataset: linear discriminant analysis (LDA), random forest, linear SVM, SVM with radial basis function (RBF) kernel, a nearest neighbor classifier, and a neural network. We considered three separate experiments. First, we performed machine learning with a train-test split at the break level, as was done in Moclán et al. (2019). Second, we bootstrapped the fragment-level data 100 times before doing a fragment-level train-test split. Third, we applied machine learning correctly by simply performing the train-test split at the fragment level. We show the results of these three experiments in Table 12. All experiments were averaged over 100 trials of randomized train-test splits, and we report the mean accuracy with the standard deviation in parentheses.

We can see that both the break-level train-test split, and bootstrapping 100 times, which is less than the 1000 times used in Moclán et al. (2019), both yield accuracy scores near 100% for many algorithms. These accuracy scores are artificially inflated, since the dataset is random. Any properly applied machine learning algorithm can achieve no better than 50% accuracy on average. We note that some algorithms, like LDA, linear SVM, and neural networks, do not achieve very high accuracies with the break-level split. The duplication of datapoints in the training and testing set can only be exploited by machine learning models that can easily overfit. Linear SVM and LDA are very low complexity models that cannot easily overfit, without a larger amount of duplication, like in the second experiment which was bootstrapped 100 times. The break-level split can be thought of as bootstrapping 7 times, since there are 7 breaks per fragment. Neural networks have the capacity to overfit, due to the number of parameters in the model, but the specific techniques used in training, such as stochastic gradient descent, offer some protection against overfitting, though the acutal mechanisms by which this occurs are currently an open problem in deep learning (Zhang, Bengio, Hardt, Recht, & Vinyals, 2021).

5 Discussion

When applied properly, machine learning can be an effective tool that can be used to advance our understanding of human evolution through the analysis of fracture patterns. However, the data need to be clean (i.e. properly and consistently recorded), there needs to be a clear understanding of the ways in which the data can be properly used, as well as a sense for the quality of the data. The quality of the data can be assessed from the perspective of how much information can be conveyed as well as replicability and transparency. Independent replication is key for testing the ability of the data to answer the question of interest.

The quality of the features presented here surpasses what has been achieved in prior work in terms of the amount of information that is extracted, transparency, and replicability.

Two limitations to this study are the sample size and the sample distribution in respect to species and skeletal element which are not uniformly balanced. These are fundamental limitations that we must continually address within the discipline and is continually improved through the expansion of the samples as research continues. More importantly, we have set a reliable groundwork from which to build future research by properly applying machine learning methods on more robust feature sets.

One of the primary challenges is how to subdivide the broken surface of a fragment into separate breaks. This issue has been acknowledged in the published research since the 1970s (Biddick & Tomenchuk, 1975; Bunn, 1982, 1989; Davis, 1985; Pickering et al., 2005). The selection of break endpoints and the measurement of angles using the virtual goniometer results in visualizations that clearly communicate how we chose to subdivide the fragments in our sample, which can then be the basis for conversations that can move beyond the recognition of the challenge toward the development of a quantifiable, replicable convention. These visualizations also offer a level of transparency by which other researchers can independently evaluate the extent to which an established protocol was followed.

Additionally, the points along the edge were used to create two measures of distance for each break and, given that the x, y, and z coordinates for each point are recorded, the measurements can be replicated with high precision. This is a substantial improvement over using calipers which are subject to variations resulting from the physical interaction of the measuring implement and the fragment which leads to higher error ranges. And, the arc length cannot be measured with the calipers alone. Likewise, using the bounding box dimensions extracted from 3D models using a Python script for measures of fragment length, width, and depth are more precise than those obtained using calipers, while additional features such as volume and surface area can be extracted as well. This also holds true for the angle measurements; the precision, accuracy, and reproducibility of the virtual goniometer far surpasses that of the handheld contact goniometer (Yezzi-Woodley et al., 2021). And, multiple measurements were taken along the entire break curve which provides substantially more information about the fracture angles.

Another commonly used variable is the break plane which has been defined as the angle of the break relative to the longitudinal axis of the bone. This is a qualitative, categorical variable with three possible labels: longitudinal, transverse, and oblique. Here we have replaced that with a continuous variable that can be calculated with precision using the points along the break curve and the principal axis of the bone fragment.

Though we have not completely abandoned the use of qualitative variables, we have used a substantial number of variables that offer precise quantitative values and can be easily replicated. Transparency is promoted through the recording of metadata and visualizations of the data on the 3D models. Because most of the data are automatically extracted using programming scripts or as a built-in feature to plug-ins using a graphical user interface, the potential for errors during data recording are minimized. As a consequence, we are satisfied that our dataset is clean, with no recording errors, is transparent and replicable. Moreover, we have improved the quality of the features relative to previous work by extracting features that offer more information about the bone fragment than ever before.

The next consideration then, is how to properly use the data when employing machine learning. Firstly, bootstrapping is used in machine learning (e.g. random forest). However, when bootstrapping is used, it is a hyperparameter built into the algorithm and not something that should be done separately. It is an egregious error to bootstrap a sample prior to subdividing it into training and test sets. This is because sampling with replacement followed by subdivision of the bootstrapped sample will lead to duplicates shared between the training and testing sets, which, of course, leads to exact matches and can then falsely inflate the accuracy of the model. In a proper application of machine learning, the testing set is completely independent of the training set, and is not used in any way during training. This allows the test accuracy to serve as an unbiased

measure of how well the model generalizes to new data.

Another important consideration is the scale at which data are classified relative to the scale of the features used in classification and how the data are split. In this case, there are two scales: the entire fragment and the individual breaks on each fragment. Once again, this is an issue of contaminating the training set with duplicates from the test set. To offer a simplified hypothetical example, consider a dataset in which there is only one fragment that is exactly 10 cm long that has 5 break faces, and the goal is to classify all the breaks in the dataset. If 3 of those breaks are added to the training set and the other 2 land in the test set, then the model will identify 10 cm as belonging to a specific label. The model will be built from this and when the test samples are used to validate the model, they will most likely classify on the basis of this feature alone. The key here is to ensure that the data are split at the appropriate level (i.e. fragment level) so as not to contaminate the training set with information from the testing set. Algorithms with many degrees of freedom (like Decision Trees or Random Forest) can easily discern a correspondence between a variable like dimension and actor, provided each fragment has a different value for dimension. The algorithm memorizes that relationship, which is not useful when new fragments are introduced to the model. Moclán et al. (2019) should have classified at the fragment level, used only variables that were specific to the breaks, or split the data at the fragment level. We chose to use two levels of data in our classification. However, we also chose to split the data at the fragment level and classify entire fragments as opposed to individual breaks. Though our results were not as compelling as those produced by Moclán et al. (2019), they are based on proper applications of machine learning and therefore are valid and reliable results. Our results are promising in that they are able to discriminate between the agents of breakage.

6 Conclusion and Future Research

The driving motivation for this line of research is to better understand early hominin evolution and how marrow exploitation factored into subsistence patterns that influenced that evolution. The results achieved here through the proper application of machine learning show promise for identifying agents of bone breakage in the archaeological record, which in turn will be useful for addressing higher level questions focused on subsistence. Future research directions that can further develop and strengthen this approach include, first and foremost, increasing the comparative samples and testing these methods on models that emulate the effects of taphonomic processes, e.g., erosion. The hominin-carnivore debate has, as the name suggests, been treated as a binary problem when, in fact, other agents can break bones, e.g., geological processes, so including samples derived from other methods of bone breakage will be an important next step.

It is also important to understand how machine learning algorithms make their predictions, to shed some light on which features are important for classification, and which are redundant or can be viewed as noise. While this interpretability problem in machine learning is well-documented and is often cited as an issue in the field, key factors in the decision making can be determined by probing the machine learning models in appropriate ways, and doing so in the anthropological context may yield interesting and important information. It would also be interesting to understand better the clustering structure of our dataset, and what characterizes the clusters in, for example, Figure 10a.

Another interesting avenue of research would be to explore fragments that consistently classify poorly to examine in more detail the sources of equifinality. Success in this direction would aid researchers in either finding features sets that can address equifinality or identify specimens that cannot be used for classification. This holds promise for much higher resolution analyses of paleoanthropological sites whereby zooarchaeological collections are not restricted to analysis at the assemblage level. The classification of individual fragments opens possibilities for spatial analysis and reconstructing more complex details of the dynamic systems and interactions that formed important paleoanthropological sites.

The present work represents a strong first step in the direction of leveraging richer datasets and the

application of advanced computational tools that hold promise for the future of taphonomic research in paleoanthropology.

References

Alcántara-García, V., Egido, R. B., del Pino, J. M. B., Ruiz, A. B. C., Vidal, A. I. E., Aparicio, Á. F., ... Domínguez-Rodrigo, M. (2006). Determinación de procesos de fractura sobre huesos frescos: Un sistema de análisis de los ángulos de los planos de fracturación como discriminador de agentes bióticos. *Trabajos de Prehistoria*, *63*(1), 37–45.

Arriaza, M. C., Aramendi, J., Maté-González, M. Á., Yravedra, J., & Stratford, D. (2021). The hunted or the scavenged? Australopith accumulation by brown hyenas at Sterkfontein (South Africa). *Quaternary Science Reviews*, *273*, 107252.

Arriaza, M. C., Yravedra, J., Domínguez-Rodrigo, M., Mate-González, M. Á., Vargas, E. G., Palomeque-González, J. F., ... Baquedano, E. (2017). On applications of micro-photogrammetry and geometric morphometrics to studies of tooth mark morphology: The modern Olduvai Carnivore Site (Tanzania). *Palaeogeography, Palaeoclimatology, Palaeoecology*, *488*, 103–112.

Barr, W. A., Pobiner, B., Rowan, J., Du, A., & Faith, J. T. (2022). No sustained increase in zooarchaeological evidence for carnivory after the appearance of *Homo erectus*. *Proceedings of the National Academy of Sciences*, *119*(5).

Bartholomew, G. A., & Birdsell, J. B. (1953). Ecology and the protohominids. *American Anthropologist*, *55*(4), 481–498.

Bello, S. M., Verveniotou, E., Cornish, L., & Parfitt, S. A. (2011). 3-dimensional microscope analysis of bone and tooth surface modifications: Comparisons of fossil specimens and replicas. *Scanning*, *33*(5), 316–324.

Biddick, K. A., & Tomenchuk, J. (1975). Quantifying continuous lesions and fractures on long bones. *Journal of Field Archaeology*, *2*(3), 239–249.

Binford, L. R. (1981). *Bones: Ancient men and modern myths*. Academic Press, New York.

Binford, L. R. (1984). *Faunal remains from Klasies River Mouth*. Academic Press, New York.

Binford, L. R. (1985). Human ancestors: Changing views of their behavior. *Journal of Anthropological Archaeology*, *4*(4), 292–327.

Binford, L. R., Bunn, H. T., & Kroll, E. M. (1988). Fact and fiction about the *Zinjanthropus* floor: Data, arguments, and interpretations. *Current Anthropology*, *29*(1), 123–149.

Binford, L. R., Ho, C. K., Aigner, J. S., Alimen, M.-H., Borrero, L. A., Te-K'un, C., ... Yi, S. (1985). Taphonomy at a distance: Zhoukoudian, "the cave home of Beijing Man"? [and comments and reply]. *Current Anthropology*, *26*(4), 413–442.

Binford, L. R., Stone, N. M., Aigner, J. S., Behrensmeyer, A. K., Haynes, G., Olsen, J. W., ... Yu-Zhu, Y. (1986). Zhoukoudian: A closer look. *Current Anthropology*, *27*(5), 453–475.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blumenschine, R. J. (1986). Carcass consumption sequences and the archaeological distinction of scavenging and hunting. *Journal of Human Evolution*, *15*(8), 639–659.

Blumenschine, R. J. (1988). An experimental model of the timing of hominid and carnivore influence on archaeological bone assemblages. *Journal of Archaeological Science*, *15*(5), 483–502.

Blumenschine, R. J. (1989). A landscape taphonomic model of the scale of prehistoric scavenging opportunities. *Journal of Human Evolution*, *18*(4), 345–371.

Blumenschine, R. J. (1995). Percussion marks, tooth marks, and experimental determinations of the timing of hominid and carnivore access to long bones at FLK *Zinjanthropus*, Olduvai Gorge, Tanzania. *Journal of Human Evolution*, *29*(1), 21–51.

Blumenschine, R. J., Bunn, H. T., Geist, V., Ikawa-Smith, F., Marean, C. W., Payne, A. G., ... van der Merwe, N. J. (1987). Characteristics of an early hominid scavenging niche. *Current anthropology*, *28*(4), 383–407.

Blumenschine, R. J., & Cavallo, J. A. (1992). Scavenging and human evolution. *Scientific American*, *267*(4), 90–97.

Blumenschine, R. J., Cavallo, J. A., & Capaldo, S. D. (1994). Competition for carcasses and early hominid behavioral ecology: A case study and conceptual framework. *Journal of Human Evolution*, *27*(1-3), 197–213.

Blumenschine, R. J., & Selvaggio, M. M. (1991). On the marks of marrow bone processing by hammerstones and hyenas: Their anatomical patterning and archaeological implications. In *Cultural beginnings: Approaches to understanding early hominid life-ways in the African savanna* (Vol. 19, pp. 17–32). Dr Rudolf Habelt GMBH Bonn.

Bobe, R., & Wood, B. (2021). Estimating origination times from the early hominin fossil record. *Evolutionary Anthropology: Issues, News, and Reviews*.

Braun, D. R., Pante, M., & Archer, W. (2016). Cut marks on bone surfaces: Influences on variation in the form of traces of ancient behaviour. *Interface Focus*, *6*(3), 20160006.

Bunn, H. T. (1982). Meat-eating and human evolution: Studies on the diet and subsistence patterns of Plio-Pleistocene hominids in East Africa [PhD dissertation]. *Berkeley: University of California, Berkeley*.

Bunn, H. T. (1989). Diagnosing Plio-Pleistocene hominid activity with bone fracture evidence. In *Bone modification* (pp. 299–315). Center for the Study of First Americans, University of Maine, Orono, Maine.

Bunn, H. T., & Ezzo, J. A. (1993). Hunting and scavenging by Plio-Pleistocene hominids: Nutritional constraints, archaeological patterns, and behavioural implications. *Journal of Archaeological Science*, *20*(4), 365–398.

Bunn, H. T., Kroll, E. M., Ambrose, S. H., Behrensmeyer, A. K., Binford, L. R., Blumenschine, R. J., ... Wymer, J. (1986). Systematic butchery by Plio/Pleistocene hominids at Olduvai Gorge, Tanzania. *Current Anthropology*, *27*(5), 431–452.

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 Computational Science*, *32*, 36–43.

Capaldo, S. D., & Blumenschine, R. J. (1994). A quantitative diagnosis of notches made by hammerstone percussion and carnivore gnawing on bovid long bones. *American Antiquity*, *59*(4), 724–748.

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*(1), 1–12.

Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., & Ranzuglia, G. (2008). Meshlab: An open-source mesh processing tool. In *Eurographics Italian Chapter Confer-*

ence (Vol. 2008, pp. 129–136).

Coil, R., Tappen, M., & Yezzi-Woodley, K. (2017). New analytical methods for comparing bone fracture angles: A controlled study of hammerstone and hyena (*Crocuta crocuta*) long bone breakage. *Archaeometry*, *59*(5), 900–917.

Courtenay, L. A., Herranz-Rodrigo, D., Huguet, R., Maté-González, M. Á., González-Aguilera, D., & Yravedra, J. (2020). Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy. *PLoS One*, *15*(10), e0240328.

Courtenay, L. A., Huguet, R., Gonzalez-Aguilera, D., & Yravedra, J. (2019). A hybrid geometric morphometric deep learning approach for cut and trampling mark classification. *Applied Sciences*, *10*(1), 150.

Courtenay, L. A., Yravedra, J., Huguet, R., Aramendi, J., Maté-González, M. Á., González-Aguilera, D., & Arriaza, M. C. (2019). Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *522*, 28–39.

Courtenay, L. A., Yravedra, J., Mate-González, M. Á., Aramendi, J., & González-Aguilera, D. (2019). 3D analysis of cut marks using a new geometric morphometric methodological approach. *Archaeological and Anthropological Sciences*, *11*(2), 651–665.

Davis, K. L. (1985). *A taphonomic approach to experimental bone fracturing and applications to several South African Pleistocene sites* (Unpublished doctoral dissertation). State University of New York at Binghamton.

Díez-Martín, F., Yustos, P. S., Uribelarrea, D., Baquedano, E., Mark, D. F., Mabulla, A., … Domínguez-Rodrigo, M. (2015). The origin of the Acheulean: The 1.7 million-year-old site of FLK West, Olduvai Gorge (Tanzania). *Scientific Reports*, *5*(1), 1–9.

Dominguez-Rodrigo, M. (1997). Meat-eating by early hominids at the FLK 22 *Zinjanthropus* site, Olduvai Gorge (Tanzania): An experimental approach using cut-mark data. *Journal of Human Evolution*, *33*(6), 669–690.

Domínguez-Rodrigo, M. (2002). Hunting and scavenging by early humans: The state of the debate. *Journal of World Prehistory*, *16*(1), 1–54.

Domínguez-Rodrigo, M. (2019). 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*(6), 2711–2725.

Domínguez-Rodrigo, M., & Baquedano, E. (2018). Distinguishing butchery cut marks from crocodile bite marks through machine learning methods. *Scientific Reports*, *8*(1), 1–8.

Domínguez-Rodrigo, M., & Barba, R. (2007). Five more arguments to invalidate the passive scavenging version of the carnivore-hominid-carnivore model: A reply to Blumenschine et al.(2007a). *Journal of Human Evolution*, *53*(4), 427–433.

Domínguez-Rodrigo, M., Bunn, H. T., & Yravedra, J. (2014). A critical re-evaluation of bone surface modification models for inferring fossil hominin and carnivore interactions through a multivariate approach: Application to the FLK Zinj archaeofaunal assemblage (Olduvai Gorge, Tanzania). *Quaternary International*, *322*, 32–43.

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. *Applied Sciences*, *11*(11), 5237.

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*(49), 20929–20934.

Domínguez-Rodrigo, M., Pickering, T. R., & Bunn, H. T. (2011). Reply to McPherron et al.: Doubting Dikika is about data, not paradigms. *Proceedings of the National Academy of Sciences*, *108*(21), E117–E117.

Domínguez-Rodrigo, M., Pickering, T. R., & Bunn, H. T. (2012). Experimental study of cut marks made with rocks unmodified by human flaking and its bearing on claims of 3.4-million-year-old butchery evidence from Dikika, Ethiopia. *Journal of Archaeological Science*, *39*(2), 205–214.

Domínguez-Rodrigo, M., Pickering, T. R., Semaw, S., & Rogers, M. J. (2005). Cutmarked bones from Pliocene archaeological sites at Gona, Afar, Ethiopia: Implications for the function of the world's oldest stone tools. *Journal of Human Evolution*, *48*(2), 109–121.

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

Domínguez-Rodrigo, M., Saladié, P., Cáceres, I., Huguet, R., Yravedra, J., Rodríguez-Hidalgo, A., ... 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 Science*, *102*, 80–86.

Domínguez-Rodrigo, M., Wonmin, B., Arampatzis, G., Varela, L., Tambusso, S., Fariña, R. A., ... Koumoutsakos, P. (2017). Revising the timing of arrival of humans in America through deep classification of cut marks on bones. *Nature Communications*.

Du, A., Rowan, J., Wang, S. C., Wood, B. A., & Alemseged, Z. (2020). Statistical estimates of hominin origination and extinction dates: A case study examining the *Australopithecus anamensis–afarensis* lineage. *Journal of Human Evolution*, *138*, 102688.

Gifford-Gonzalez, D. (1989). Ethnographic analogues for interpreting modified bones: Some cases from East Africa. In R. Bonnichsen & M. Sorg (Eds.), *Bone modification* (pp. 179–246). Center for the Study of the First Americans Orono.

Gümrükçu, M., & Pante, M. C. (2018). Assessing the effects of fluvial abrasion on bone surface modifications using high-resolution 3-D scanning. *Journal of Archaeological Science: Reports*, *21*, 208–221.

Harmand, S., Lewis, J. E., Feibel, C. S., Lepre, C. J., Prat, S., Lenoble, A., ... Roche, H. (2015). 3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya. *Nature*, *521*(7552), 310–315.

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.

James, E. C., & Thompson, J. C. (2015). On bad terms: Problems and solutions within zooarchaeological bone surface modification studies. *Environmental Archaeology*, *20*(1), 89–103.

Jiménez-García, B., Abellán, N., Baquedano, E., Cifuentes-Alcobendas, G., & Domínguez-Rodrigo, M. (2020). 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*(171), 20200782.

Jiménez-García, B., Aznarte, J., Abellán, N., Baquedano, E., & Domínguez-Rodrigo, M. (2020).

Deep learning improves taphonomic resolution: High accuracy in differentiating tooth marks made by lions and jaguars. *Journal of the Royal Society Interface*, *17*(168), 20200446.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Marean, C. W., Spencer, L. M., Blumenschine, R. J., & Capaldo, S. D. (1992). Captive hyaena bone choice and destruction, the schlepp effect and Olduvai archaeofaunas. *Journal of Archaeological Science*, *19*(1), 101–121.

Maté-González, M. Á., Courtenay, L. A., Aramendi, J., Yravedra, J., Mora, R., González-Aguilera, D., & Domínguez-Rodrigo, M. (2019). Application of geometric morphometrics to the analysis of cut mark morphology on different bones of differently sized animals. does size really matter? *Quaternary International*, *517*, 33–44.

Maté-González, M. Á., González-Aguilera, D., Linares-Matás, G., & Yravedra, J. (2019). New technologies applied to modelling taphonomic alterations. *Quaternary International*, *517*, 4–15.

McPherron, S. P., Alemseged, Z., Marean, C. W., Wynn, J. G., Reed, D., Geraads, D., ... Béarat, H. A. (2010). Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. *Nature*, *466*(7308), 857–860.

Merritt, S. R., Pante, M. C., Keevil, T. L., Njau, J. K., & Blumenschine, R. J. (2019). Don't cry over spilled ink: Missing context prevents replication and creates the Rorschach effect in bone surface modification studies. *Journal of Archaeological Science*, *102*, 71–79.

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*(9), 4663–4680.

Moclán, A., Huguet, R., Márquez, B., Laplana, C., Arsuaga, J. L., Pérez-González, A., & Baquedano, E. (2020). Identifying the bone-breaker at the Navalmaíllo Rock Shelter (Pinilla del Valle, Madrid) using machine learning algorithms. *Archaeological and Anthropological Sciences*, *12*(2), 46.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, *14*.

O'Neill, R. C., Angulo-Umana, P., Calder, J., Hessburg, B., Olver, P. J., Shakiban, C., & Yezzi-Woodley, K. (2020). Computation of circular area and spherical volume invariants via boundary integrals. *SIAM Journal on Imaging Sciences*, *13*(1), 53–77.

Palomeque-González, J. F., Maté-González, M. Á., Yravedra, J., San Juan-Blazquez, M., Vargas, E. G., Martín-Perea, D. M., ... Domínguez-Rodrigo, M. (2017). Pandora: A new morphometric and statistical software for analysing and distinguishing cut marks on bones. *Journal of Archaeological Science: Reports*, *13*, 60–66.

Pante, M. C., Blumenschine, R. J., Capaldo, S. D., & Scott, R. S. (2012). Validation of bone surface modification models for inferring fossil hominin and carnivore feeding interactions, with reapplication to FLK 22, Olduvai Gorge, Tanzania. *Journal of Human Evolution*, *63*(2), 395–407.

Pante, M. C., Muttart, M. V., Keevil, T. L., Blumenschine, R. J., Njau, J. K., & Merritt, S. R. (2017). A new high-resolution 3-D quantitative method for identifying bone surface modifications with implications for the Early Stone Age archaeological record. *Journal of Human Evolution*, *102*, 1–11.

Pante, M. C., Scott, R. S., Blumenschine, R. J., & Capaldo, S. D. (2015). Revalidation of bone

surface modification models for inferring fossil hominin and carnivore feeding interactions. *Quaternary International*, *355*, 164–168.

Parkinson, J. A. (2018). Revisiting the hunting-versus-scavenging debate at FLK Zinj: A GIS spatial analysis of bone surface modifications produced by hominins and carnivores in the FLK 22 assemblage, Olduvai Gorge, Tanzania. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *511*, 29–51.

Pickering, T. R., Domínguez-Rodrigo, M., Egeland, C. P., & Brain, C. (2005). The contribution of limb bone fracture patterns to reconstructing early hominid behaviour at Swartkrans Cave (South Africa): Archaeological application of a new analytical method. *International Journal of Osteoarchaeology*, *15*(4), 247–260.

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*(1), 1–13.

Plummer, T. W., & Bishop, L. C. (2016). Oldowan hominin behavior and ecology at Kanjera South, Kenya. *Journal of Anthropological Sciences*, *94*.

Pobiner, B. L. (2015). New actualistic data on the ecology and energetics of hominin scavenging opportunities. *Journal of Human Evolution*, *80*, 1–16.

Potts, R. (1983). Foraging for faunal resources by early hominids at Olduvai Gorge, Tanzania. *Animal and Archaeology. 1. Hunters and Their Prey; BAR International Series*, *163*, 51–62.

Prat, S. (2018). First hominin settlements out of Africa. tempo and dispersal mode: Review and perspectives. *Comptes Rendus Palevol*, *17*(1-2), 6–16.

Schmidt, C. W., Moore, C. R., & Leifheit, R. (2012). A preliminary assessment of using a white light confocal imaging profiler for cut mark analysis. In *Forensic Microscopy for Skeletal Tissues* (pp. 235–248). Springer.

Selvaggio, M. M. (1994a). Carnivore tooth marks and stone tool butchery marks on scavenged bones: Archaeological implications. *Journal of Human Evolution*, *27*(1-3), 215–228.

Selvaggio, M. M. (1994b). *Evidence from carnivore tooth marks and stone-tool-butchery marks for scavenging by hominids at FLK Zinjanthropus, Olduvai Gorge, Tanzania* (Unpublished doctoral dissertation). Rutgers University.

Selvaggio, M. M. (1998). Evidence for a three-stage sequence of hominid and carnivore involvement with long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania. *Journal of Archaeological Science*, *25*(3), 191–202.

Shipman, P. (1983). Early hominid lifestyle: Hunting and gathering or foraging and scavenging. *Animals and Archaeology*, *1*, 31–49.

Shipman, P. (1986). Scavenging or hunting in early hominids: Theoretical framework and tests. *American Anthropologist*, *88*(1), 27–43.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Thompson, J. C., Carvalho, S., Marean, C., & Alemseged, Z. (2019). Origins of the human predatory pattern: The transition to large-animal exploitation by early hominins. *Current Anthropology*, *60*(1), 1–23.

Thompson, J. C., McPherron, S. P., Bobe, R., Reed, D., Barr, W. A., Wynn, J. G., ... Alemseged, Z. (2015). Taphonomy of fossils from the hominin-bearing deposits at Dikika, Ethiopia. *Journal of Human Evolution*, *86*, 112–135.

Van Rossum, Guido and Drake, Fred L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395–416.

Yezzi-Woodley, K. (2022). *Applying a new 3D data extraction method and machine learning to identify hominin bone marrow exploitation* (Unpublished doctoral dissertation). University of Minnesota.

Yezzi-Woodley, K., Calder, J., Olver, P., Sweno, M., & Siewert, C. (n.d.). *The batch artifact scanning protocol: A new method using computed tomography (CT) to rapidly create three-dimensional models of objects from large collections.* (In preparation)

Yezzi-Woodley, K., Calder, J., Olver, P. J., Cody, P., Huffstutler, T., Terwilliger, A., ... Tostevin, G. (2021). The virtual goniometer: Demonstrating a new method for measuring angles on archaeological materials using fragmentary bone. *Archaeological and Anthropological Sciences*, *13*(7), 1–16.

Yravedra, J., Aramendi, J., Maté-González, M. Á., Austin Courtenay, L., & González-Aguilera, D. (2018). Differentiating percussion pits and carnivore tooth pits using 3D reconstructions and geometric morphometrics. *PLoS One*, *13*(3), e0194324.

Yravedra, J., Garcia-Vargas, E., Maté-González, M. Á., Aramendi, J., Palomeque-González, J. F., Valles-Iriso, J., ... Dominguez-Rodrigo, M. (2017). The use of micro-photogrammetry and geometric morphometrics for identifying carnivore agency in bone assemblages. *Journal of Archaeological Science: Reports*, *14*, 106–115.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

Zhu, Z., Dennell, R., Huang, W., Wu, Y., Qiu, S., Yang, S., ... Ouyang, T. (2018). Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago. *Nature*, *559*(7715), 608–612.