Machine Learning interspecific identification of mouse first lower molars (genus Mus Linnaeus, 1758) and application to fossil remains from the Estrecho Cave (Spain)

Abel Moclan, Ángel C. Domínguez-García, Emmanuelle Stoetzel, Thomas Cucchi, Paloma Sevilla, César Laplana

Quaternary Science Reviews, 299 : 107877.
https://www.sciencedirect.com/science/article/pii/S027737912200508X?dgcid=coauthor

Machine Learning interspecific identification of mouse first lower molars (genus Mus Linnaeus, 1758) and application to fossil remains from the Estrecho Cave (Spain)

Abel Moclan [a, b], Angel C. Domínguez-García [c], Emmanuelle Stoetzel [d], Thomas Cucchi [e], Paloma Sevilla [c], Cesar Laplana [f]

[a] Escuela de Posgrado, Universidad de Burgos, Don Juan de Austria 1, 09001, Burgos, Spain
[b] Institute of Evolution in Africa (IDEA), University of Alcala de Henares, Covarrubias 36, 28010, Madrid, Spain
[c] Departamento de Geodinamica, Estratigrafía y Paleontología, Facultad de Ciencias Geologicas, Universidad Complutense de Madrid, C/ Jose Antonio Novais, 12, 28040, Madrid, Spain
[d] Histoire Naturelle de L'Homme Prehistorique (HNHP), UMR 7194, Museum National D'Histoire Naturelle / CNRS / UPVD, Musee de L'Homme, Palais de Chaillot, 17 Place Du Trocadero, 75016, Paris, France
[e] Archeozoologie, Archeobotanique: Societes, Pratiques et Environnements (AASPE), UMR 7209, Museum National D'Histoire Naturelle / CNRS, CP56-43 Rue Buffon, 75005, Paris, France
[f] Museo Arqueologico y Paleontologico de La Comunidad de Madrid, Plaza de Las Bernardas S/n, 28801 Alcala de Henares, Madrid, Spain

Abstract

One of the first steps to address palaeontological studies is the taxonomic identification of fossils according to their morphology. Geometric Morphometric techniques together with multivariate statistical analysis are known to be precise tools to achieve this goal. More recent alternative techniques such as Machine Learning are still rarely used in Palaeontology, although it has been shown in various examples that they can offer powerful alternative statistical approaches to analyse quantitative morphometric data. Here we show how Machine Learning applied to two-dimensional geometric morphometric data from the outline shape of the lower first molars of Mus spp. has proven useful to overcome taxonomic problems. We collated a photographic database of 303 lower first molars from modern populations of Mus musculus domesticus and Mus spretus from southwestern Europe and North Africa to compare the performance between classic multivariate statistics and Machine Learning algorithms in identifying the two species from their dental morphology. We also include Late Holocene Mus specimens from the Estrecho Cave (east-central Spain) to predict their specific status. Our results suggest that Machine Learning is more efficient than classical statistical analyses in taxonomic identification of Mus molars, reaching 100% of correct classification. The application of such techniques to fossil material showed that ensemble/stacking algorithms provided robust identification of both M. m. domesticus and M. spretus in the Estrecho Cave assemblage and confirmed that both species colonised the Iberian Peninsula at a time prior to the formation of the site.

Keywords: Murinae, Rodentia, Western Mediterranean, Holocene, Geometric Morphometrics, Ensemble Learning.

1. Introduction

The study of Quaternary small mammals (Rodentia, Lagomorpha, Eulipotyphla, Chiroptera) has a special interest in vertebrate palaeontology. Among small mammals, rodents have a key role due to their great diversity, high speciation rate, and narrow ecological requirements of many of its species, making of them excellent biochronological and palaeoecological proxies (Andrews, 1995; Chaline et al., 1999; Avery, 2007; Cuenca-Besc_os et al., 2016). However, in more recent sites, particularly from the Middle Holocene onwards, the study of rodent and small mammal fossils in

archaeo/palaeontological sites decreases due to their comparative lower value as biochronological proxies, since alternative methods such as ceramics or numismatics can provide more precise relative ages (Domínguez García et al., 2019a, 2020; Laplana and Sevilla, 2013; L_opez García, 2011; Papayiannis, 2012). Therefore, it hinders the reconstruction of the recent dynamics of small mammal communities.

This general problem applies in the Holocene biogeographic history of rodents in the western Mediterranean region, where the scarcity of data has obscured the reconstruction of the dispersal process of species of the genus Mus Linnaeus, 1758. At present, two species occur sympatrically in most of the Iberian Peninsula, Balearic Islands, Mediterranean France and North Africa (Wilson et al., 2017): the western Mediterranean mouse (Mus spretus Lantz, 2013) and the western house mouse (Mus musculus domesticus Schwarz and Schwarz, 1943). According to recent studies (Domínguez García et al., 2019a; Lalis et al., 2019), M. spretus colonised Europe from the Maghreb during the Middle Holocene (Late Neolithic), whereas M. m. domesticus arrived later in the western Mediterranean region, both in Africa and in Europe, from the Levant during the Late Holocene (Iron Age) (Cucchi et al., 2005b; Bonhomme et al., 2011; Oueslati et al., 2020). These studies indicate that such dispersal processes may have been associated with accidental anthropogenic translocations through navigation routes in the Mediterranean Sea. However, many of the fossil occurrences of both species are taxonomically and/or chronologically imprecise, thus constraining the evidence for clarifying some details concerning the timing and mechanism involved in hese colonisations (Cucchi et al., 2005b; Domínguez García et al., 2019a; Lalis et al., 2019; Valenzuela-Lamas et al., 2011).

Mus spp. Show very similar skeletal and dental morphology and size, hindering often species identification. Based on sufficient material, a set of craniodental morphometric criteria can allow to discriminate between M. spretus and M. m. domesticus (Darviche et al., 2006; Darviche and Orsini, 1982; Gerasimov et al., 1990). However, given that many of these characters show a high intraspecific variability together with the fragmentary state in which the fossil material is usually preserved, taxonomic identification in palaeontological studies usually relies on the morphology of the first lower molar (m1) which is known to provide the most efficient criteria (Darviche et al., 2006; Darviche and Orsini, 1982). Thus, in M. spretus the anterior region of the m1 shows a tetralobate morphology whereas it is trilobate in the case of M. musculus. This difference has to do with the different development of the anterolabial tubercle (tE) in each species, which is well individualised in M. spretus and more reduced and not individualised in M. musculus. Additionally, an external cingular margin with a well-developed secondary cusp (c1) in the m1 is common in M. spretus, whereas it is infrequent or less developed in M. musculus (Darviche et al., 2006; Darviche and Orsini, 1982). However, these comparative morphological criteria are subject to a high intraspecific variability and are also affected by tooth wear (Darviche and Orsini, 1982; Cucchi et al., 2002), finally involving a high degree of subjectivity in the process of their observation.

Many efforts have been made to overcome these taxonomic difficulties by quantifying the shape differences in the m1 of Mus by means of Geometric Morphometrics (GMM). The approach in most of these studies involved the use of Elliptic Fourier Analysis (EFA) of the molar outline (Cucchi et al., 2002, 2005a; Michaux et al., 2007; Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011). More recently, landmark and semi-landmarks approaches to describe the molar outline have been employed (Cucchi et al., 2013, 2020; Weissbrod et al., 2017). All these studies have evidenced that GMM techniques,taking as reference the morphometric data obtained from modern populations, together with multivariate statistical analyses are powerful tools for intra- and interspecific discrimination of the genus Mus. Most of this research has applied the technique to material collected in sites from Southeastern Europe and Southwestern Asia, providing reliable taxonomic identifications and therefore a basis on which to build a clear picture of the biogeographic history of oriental taxa of the genus and to interpret how climatic and environmental variations or human activities have influenced in it (Cucchi et al., 2002, 2011, 2013, 2020; Cucchi, 2008; Cucchi and Vigne, 2006; Weissbrod et al., 2017). In contrast, there are fewer studies with a similar approach dealing with the western Mediterranean Mus records (Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011).

In recent times, the incorporation of Artificial Intelligence in different disciplines has improved significantly the results previously provided using classical Fisherian or Bayesian models. In the case of palaeontological research, these tools can be combined with the classical morphometric analyses in two ways: 1) using Convolutional Neural Networks (Krizhevsky et al., 2012; Kim and MacKinnon, 2018; Tan and Le, 2020) that can be used to classify taxa through the analysis of images containing relevant taxonomic information (Miele et al., 2020; Norouzzadeh et al., 2018; Romero et al., 2020; Sevillano et al., 2020; Villon et al., 2018); or 2) a combination of Artificial Intelligence methods and GMM using Machine Learning (ML) algorithms instead of the classical statistical approach with Linear Discriminant Analyses or Canonical Variance Analyses (Bellin et al., 2021; Courtenay et al., 2019; Cucchi et al., 2020; Herranz-Rodrigo et al., 2021; Monson et al., 2018; Quenu et al., 2020; Wills et al., 2021).

The first type of approximation (i.e. Convolutional Neural Networks) is probably the future for some different scientific disciplines, especially Palaeontology. However, the computational requirements of these methods and the high number of images needed to train the models lead us to consider that, at least nowadays, they are not useful tools to perform palaeontological studies. A good example of this type of approximation has been the differentiation between Mus musculus and Apodemus sylvaticus using Convolutional Neural Networks, which provided a perfect classification of these genera, otherwise easily differentiable by a traditional morphometric analysis (Miele et al., 2020).

Here we have tried an improved use of GMM for the taxonomic discrimination between recent M. m. domesticus and M. spretus by combining it with ML procedures, as an exploratory approach to test the effectiveness of this method when applied to isolated fossil teeth. Thus, we are pursuing to find the most accurate model to differentiate between both species using Machine Learning algorithms. For this purpose we employed some algorithm considered to be some of the most powerful methods in Machine Learning (Lantz, 2013), and subsequently applied Ensemble and Stacking methods in order to improve the classification rate of the base learners.

These methods were then used in order to classify the Mus materials from the Holocene small mammal assemblage of Estrecho Cave (Spain) where both species where previously proposed as probably identified (Domínguez García et al., 2019b, 2020).

2. Small mammal assemblage of the Estrecho Cave

The Estrecho Cave is located in the central-eastern area of the Iberian Peninsula (Villares del Saz, Cuenca) (Fig. 1) and is part of the karst system of Villares del Saz with other caves such as Cueva de las Monedas, Cueva de las Palomas or Cueva del Camino (Ortega Martínez and Martín Merino, 1992). The archaeological record documented has provided evidence of a funerary human activity inside the cave during the Bronze Age (Guisado Di Monti and Bern_andez G_omez, 2016).

The studies performed by Domínguez Gerasimov et al. (1990, 2019b)have revealed the palaeontological interest of this site where a rich small mammal assemblage of Late Holocene age was identified in the sedimentary deposits located at the eastern part of the current entrance to the cave. This assemblage is characterized by its high richness, both in the number of remains and in the number of taxa of rodents, lagomorphs, eulipotyphlans and chiropterans (Table 1). Its richness is probably a consequence of the fact that at least part of the micromammal assemblage is clearly related to the activity of a small to medium-sized mammal carnivore (Domínguez García et al., 2019b). Besides its richness, the assemblage is also relevant for having provided the earliest reliable record of Suncus etruscus in southwestern Europe (Domínguez García et al., 2020).

The 14C dating obtained from an Arvicola sapidus humerus provided two intervals of calibrated age: 2310e2290 (10.70%) and 2272e2149 (82.80%) years cal BP (Domínguez García et al., 2019b, 2020) for the small mammal assemblage. In this sense, it must be noted that the anthropic use of the cave was older than the accumulation of the small mammal remains, since the archaeological record documented in separate deposits inside the cave is clearly related to the Early Bronze Age (Guisado Di Monti and Bern_andez G_omez, 2016), whereas the radiocarbon age obtained place the small

mammal assemblage between the end of the Iron Age and the beginning of the Roman Period in Iberia.

This small mammal material is still under study, with several matters that remain to be cleared. For instance, solving some taxonomic issues such as reaching species level in the determination of several rodent pair of species (Microtus arvalis-agrestis, Microtus duodecimcostatus-lusitanicus, Apodemus sylvaticus-flavicollis, Mus spretus-musculus e Table 1), which would be of interest for environmental reconstructions or biogeographical purposes. However, the discrimination between these species that belong to the same genus is complicated due to their similar morphology and size overlap. Thus, we aim to search for methods to overcome this issue involves using more complex and modern techniques than those used by traditional taxonomy.

Within this context, we present here the results obtained after using a combination of ML and GMM methods with the aim of obtaining reliable taxonomic identifications that distinguish between Iberian specimens of the genus Mus (M. m. domesticus, M. spretus) and reply to the question of the potential co-occurrence of two Mus species in the site.

3. Material and methods

3.1. Fossil and modern samples

The fossil material of the genus Mus used for this study comes from the sampling carried out in the Estrecho Cave during 2016 at the uppermost level (CE-SE) of the sedimentary package located at the eastern side of the entrance to the cave. A water-screening system with two superimposed sieves of mesh sizes 2 and 0.5 mm was employed to retrieve all the microvertebrate remains. Among these, a total of 143 remains belonging to Mus were identified, of which 48 first lower molars (m1) were selected for this study. Fragmented and/or digested molars, or those that showed an advanced tooth wear stage were excluded.

The modern reference sample consisted of 303 m1 (belonging 157 to Mus musculus and 146 to Mus spretus) coming from different Spanish regions (Iberian Peninsula, Canary Islands and Mallorca), as well as some specimens from France, Algeria and Morocco (Table 2). This material belongs to the collections housed in the Estaci_on Biol_ogica de Do˜nana (EBD), the Institut des Sciences de l'Evolution de Montpellier (ISEM), Museo Nacional de Ciencias Naturales (MNCN) and specimens captured during field campaigns carried out by the team of the French ANR project MOdern Human installation in Morocco, Influence on the small terrestrial vertebrate biodiversity and Evolution (MOHMIE). The specimens from the ISEM and MOHMIE are genotyped, whereas the taxonomic assignations of the EBD and MNCN specimens were checked according to the set of cranial criteria currently used to distinguish between both species (Darviche et al., 2006; Darviche and Orsini, 1982; Gerasimov et al., 1990); together with the relative length of the tail, measured on their preserved skins, which is always shorter than the length of the head þ body in M. spretus and longer in M. m. domesticus (Britton et al., 1976). The detailed information of each specimen, including catalogue numbers and repositories, are provided in Supplementary File 1.

3.2. Geometric Morphometric Analysis (GMM)

The skeletal morphometric similarities between Mus sibling species (Darviche et al., 2006; Darviche and Orsini, 1982; Gerasimov et al., 1990) has triggered the application of GMM techniques in order to discriminate between both species, due to the differentiation of them using only morphological criteria is extremely difficult. Here, shape analysis of the first lower molar (m1) outline was carried out following Cucchi et al. (2020, 2013).

GMM data for the m1 outline analysis were obtained using 2D images of the occlusal view from photos taken of each m1 included in this study, keeping special care that all had the same orientation (occlusal surface horizontal), since changes in the tooth orientation may lead to alterations in outline morphology and therefore in landmark and semi-landmarks position (Fox et al., 2020). Teeth of

modern samples were grouped together independently of their age and sex since these factors have no significant influence on the molar outline in murids (Renaud, 2005; Valenzuela-Lamas et al., 2011). Right teeth of modern specimens were used whenever possible; however, all the fossil teeth were measured whether right or left. Following Renaud (1999) and Stoetzel et al. (2013), left specimens were mirrored.

We used the "landmark and sliding semi-landmarks" approach following Cucchi et al. (2013) to obtain the outline data. Accordingly, one landmark was positioned at the furthest point of the anterior lobe of the tooth, and 63 equally spaced semi-landmarks along the crown's external outline (Fig. 2) using tpsUtil v.1.76 (Rohlf, 2018) and tpsDig v.2.32 (Rohlf, 2010).

We have employed a Bending Energy Minimization (BEM) method for semi-landmark alignment applied on outlines defined by one landmark and 63 sliding semi-landmarks, since this has been demonstrated to be the most efficient GMM approach to capture the taxonomic signal for the genus Mus (Cucchi et al., 2020).

To standardise the position, orientation and scaling information of each specimen, a Generalized Procrustes Analysis (GPA) was conducted using tpsRelw v. 1.70 (Rohlf, 2019). By this procedure the BEM method was applied, where semi-landmarks are constrained to slide along an estimated tangent at each sliding point (Bookstein, 1997). With this methodology, we obtained the Procrustes coordinates which are the molar-shape variables set. With the Procrustescoordinates of each specimen, we performed two Principal Components Analysis (PCA) on the covariance matrix, one applied to the modern samples and the other to both the modern and fossil samples. The PC scores obtained were used as shape variables in the subsequent analysis.

Once the PCAs were calculated, a Linear Discriminant Analysis (LDA) was applied. This test has been performed to classify the specimens with the standard classic method in order to check if other Machine Learning models provide more accurate results than the classical methods. The LDA was also performed using a leaveone- out cross-validation method (LOOCV). LDA were calculated using the library 'MASS' (Ripley et al., 2020) of R (R Core Team, 2022). The performance of the LDA was developed using training and testing datasets in the sameway as for the other ML algorithms (see below).

In order to identify the best way to analyse the modern data we conducted the Linear Discriminant Analysis with four different approaches: 1) including 82 PC scores; 2) including 7 PC scores; 3) including 14 PC scores and 4) including 27 PC scores. This selection of PC scores is based on the percentage of the variance obtained for each sample. In the case of the 14 PC scores, they contain the 95% of the variance, while in the case of the 27 PC scores they contain the 99% of the variance. The selection of 82 PC scores is related to the fact that 82 is the maximum number of PC that can be used when a LDA is performed (if all PC scores are used, a statistical error appears because from PC 83, the variables appear to be constant [i.e., co-linear] within groups). The last case (7 PC scores) is related to the application of a broken stick test which indicated that the optimum number of PC scores is 7 (see Results section).

3.3. Machine Learning Analysis (ML)

ML algorithms are one of the most powerful statistical methods currently available (Lantz, 2013). This type of statistical approach allows the classification and prediction of labelled categories within analytical samples using a powerful system of data evaluation (Kuhn and Johnson, 2013).

A standard procedure in this type of predictive models is to improve the sample size using bootstrapping (Abell_an et al., 2022; Mocl_an et al., 2019, 2020) or even Generative Adversarial Networks (Courtenay and Gonz_alez-Aguilera, 2020). However, we have not applied this type of methods here since the sample size is large enough to carry out the study without the need of enlarging virtually the number of cases. Furthermore, this decision is supported on our previous experience dealing with GMM Mus species identifications, in which the accuracy rate was high although sample size improvement was not used (Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011).

In all performed ML analyses, we have evaluated the accuracy values as the estimator of the best models. The accuracy refers to the percentage of success in the classification of cases by the algorithm, varying between 0 and 1. Zero corresponds to a null classification and 1 being a perfect classification of the entire sample. However, Kappa statistic has been used in combination with accuracy in order to determine which model is the most accurate in any case. The Kappa statistic ranges from _1 to 1, with values of 0.80e1 providing “very good agreement” (Lantz, 2013).

We also calculated the sensitivity, the specificity and the balanced accuracy of all models in order to evaluate the performance of each algorithm. The sensitivity of a model measures the proportion of positive samples that were correctly classified, in contrast with specificity where the proportion of negative samples that were correctly classified is quantified. As Lantz (2013) points out these can be understood as “the true positive rate” and “the true negative rate”. Finally, balanced accuracy corrects this by averaging the results of the sensitivity and the specificity (Domínguez-Rodrigo, 2019).

In this paper, we have used a series of 11 algorithms that are considered a selection of the best available methods (Lantz, 2013). The selection is composed of different types of algorithms which evaluate the data in different ways. This aspect allows us the possibility of testing extremely different mathematically methods to analyse the same sample and discover the best way to obtain a correct taxonomic assignment of the fossil specimens. What's more, the application of Ensemble and Stacking methods needs the use of really different algorithms in order to improve the quality of the models (see below).

The specific algorithms we have used here are: Neural Networks (NNET), Linear Support Vector Machines (SVMl), Radial Support Vector Machines (SVMr), k-nearest neighbour (kNN), Logistic Regresion (LG), Decision Trees using the 5.0 algorithm (DTC5.0), Random Forest (RF), Gradient Boosting (GB), Naïve Bayes (NB), Linear Discriminant Analysis (LDA) and Partial Least Squares (PLS).

The different ML algorithms were trained in the same way. First, the sample (PC scores provided by the previous performing of PCA) was divided into two different parts: training (70% of the sample) and testing (30% of the sample). This methodology is used to check the reliability of a model, observing whether the tested model leads to the correct classification not only of the specimens in the studied sample, but also of additional unknown samples. This approach is similar to the approach performed with the ‘classical’ Linear Discriminant Analysis. In order to overcome the possible overfitting provided by the model, we added cross-validation methods (10-fold CV; repeats ¼ 10). In this way, we used the typical combination of training/cross-validation/testing in order to create accurate models.

All these methods have been developed using the ‘caret’ (Kuhn et al., 2020) and ‘caretEnsemble’ (Deane-Mayer, 2019; Deane-Mayer and Knowles, 2019) R libraries (R Core Team, 2022). These libraries allow us to perform the hyper-parameter configuration easily with the ‘tuneLength’ function used to generate 20 different models per algorithm. After ‘tuneLength’ application, accuracy and Kappa values were used to select the most accurate hyperparameter configuration. In addition to the previous procedures, we must point out that the performance of the models through the visualization of AUC-ROC curves developed using the ‘MLeval’ R library (John, 2020) were also evaluated.

‘caretEnsemble’ is a powerful library oriented to performing Ensemble Learning (Deane-Mayer, 2019; Deane-Mayer and Knowles, 2019). It contains methods that generate metaalgorithms that combine the classification of all the base learners previously trained. These techniques are used because they usually produce a better performance analysis because they give the opportunity to combine different algorithms that build classifications in clearly different ways (Dietterich, 2000; Opitz and Maclin, 1999; Rokach, 2010; Sagi and Rokach, 2018). To take advantage of these methods, it is important to know previously if there is no correlation between the classifications obtained using the different base learners. Here we have used the function ‘modelCor’ of the ‘caret’ library (Kuhn et al., 2020) to identify the existence of a possible correlation between the models.

For ensemble learning we have used the ‘caretEnsemble’ library (Deane-Mayer, 2019; Deane-Mayer and Knowles, 2019) since it provides an easier way to train the meta-learner from ‘caret’

models. Its only disadvantage is that it cannot provide classifications of three or more categories (Deane-Mayer and Knowles, 2019). For our aims here, however, this is not an issue because we are dealing only with two different species (i.e., Mus musculus domesticus and Mus spretus). Thus, we have used the 'caretEnsemble' function to create an ensemble algorithm that works using a Generalized Linear Model (GLM). Furthermore, we have used the function 'caretStack" to train three stacked models: a Neural Network, a Random Forest and a Gradient Boosting algorithm. The reason for this choice of algorithms to perform the stacked model was the high power of resolution that they have solving different types of ML problems and because they admit the highest number of hyperparameter combinations in 'caretEnsemble' library.

It must be pointed out that in the case of the ensembled models we have used an 11-fold CV method. The reason is that if the same cross-validation value is used for the preparation of the basic algorithms and the meta-algorithms, an error can be generated by the library 'caretEnsemble' (Deane-Mayer, 2019).

All the above-mentioned methods were applied to the modern sample in order to find the most accurate model possible (see Results section).

As in the case of the LDA all analyses were performed four times: 1) including 82 PC Scores; 2) including 7 PC scores; 3) including 14 PC scores and 4) including 27 PC scores.

Additionally, once the modern samples had been tested, we repeated all the analyses adding the data provided by the Mus material of the Estrecho Cave. Subsequently, once all the possible issues had been clarified, the ML results of the most accurate models were used to identify the presence of Mus musculus domesticus and Mus spretus of the palaeontological sample. To evaluate the strength of the classification obtained, the posterior probability values (p) of belonging to one of the two species was calculated for each fossil m1. According to Cucchi et al. (2013), specimens classified with posterior probabilities above 0.9 are considered reliable taxonomic identifications. All four ensemble models were finally used together in the process of classifying the fossil specimens considering the posterior probabilities. A RStudio script is included in the supplementary material to show the different codes that were used to develop the ML analysis (Supplementary File 2).

4. Results

4.1. Analysis of the modern dataset (experimental approximation)

After calculating the GPA, a PCA was applied to the modern samples in order to create the new variables: PC scores. We performed a first approximation to the PCA by plotting the first two PC scores (Fig. 3). As can be seen in Fig. 3, both species are mainly separated by the x-axis, being the M. spretus sample mainly in the negative part of the axis whereas the M. m. domesticus are found mostly in the positive area. However, this plot shows a high degree of overlapping for both species and only explains 59.4% of the variance.

The application of the classical Linear Discriminant Analysis (i.e. non-ML analysis) to the PC scores of the modern dataset shows high accuracy values (Table 3), but it is clear that the use of control methods (i.e. LOOCV) and analysis of the testing dataset affects the results in a negative way (i.e. accuracy of the models decreases).

When 7 PC scores are used, the classification of the training and testing datasets are extremely similar, with a low reduction of 0.4% of the accuracy when the 'normal' LDA is applied and with an increase of 0.1% of the accuracy when LOOCV is applied. However, it must be noted that the accuracy values are lower than those provided by other number of PC scores analyses when LOOCV is not used and when the training dataset is classified with LOOCV.

If 14 PC scores are analysed, a better performance is generally shown by the LDA, except for the classification of the testing dataset with LOOCV, which provided an accuracy value of 0.933. This last result is especially interesting because it has been reduced by 3.4% from the accuracy of the training dataset with LOOCV.

The sample with 27 PC scores gave its best performance with a 'classic' LDA method. Although the classification of the testing sample has been reduced from the previous classification of the training sample, the classification of the testing with LOOCV gave an accuracy value of 0.955.

The case of 82 PC scores is especially interesting due to the different performance shown by the 'classical' LDA and the samples with LOOCV. When LOOCV is not used, the accuracy between the classification of the training and the testing datasets goes down 2.9%. However, if LOOCV is applied, the difference between both datasets is extremely pronounced (the classification of the testing sample having undergone a loss of accuracy of 19%).

The ML analysis on the other hand, provided highly accurate models with most of the algorithms used. First of all, when the AUC-ROC values were calculated (Table 4) all the algorithms provided values between 0.90 and 0.99 if 82 PC scores were used, between 0.95 and 0.99 with 7 PC scores, between 0.94 and 1 with 14 PC scores, and between 0.94 and 1 if 27 PC scores were used. As shown further below, other values also showed high confidence results (e.g. accuracy, kappa).

If the presence of non-correlation is used to test the performance of the analyses, the datasets with 82 and 27 PC scores must be considered the best options to create ensemble models.

When 82 PC scores were analysed, high algorithmic performance is obtained (Table 6). In the case of the ML algorithms the best performance is obtained by LDA (accuracy ¼ 0.989; Kappa ¼ 0.978) while the least are shown by NB (accuracy ¼ 0.822; Kappa ¼ 0.644). However, these results which are quite similar to those obtained previously with the Linear Discriminant Analysis, are highly improved using stacking methods with a Neural Network algorithm. This last approximation provides a perfect classification of the modern data (accuracy ¼ 1; Kappa ¼ 1).

When the analyses are done using 7 PC scores, the performance is similar to that provided by the analysis of 82 PC scores (Table 7). In this case, the lowest accuracy value provided by the algorithms is 0.944 (DTC5.0; kappa ¼ 0.888) while the highest accuracy value has been shown by SVMr and NB (accuracy ¼ 0.989; kappa ¼ 0.978). In this case, two ensemble models (NNET and GB) have provided a perfect classification (i.e., accuracy/kappa ¼ 1). When the analyses were calculated using 14 PC, the results obtained differed significantly from those from the previous mentioned ones (Table 8). In the first place, it is important remark that the best performance was obtained using the non-ensemble model. In this case, Ensemble Learning failed to get better results than those obtained after using the first series of algorithms. Now the most accurate models show an accuracy value of 0.978 and a kappa value of 0.955 (kNN).

Finally, when 27 PC scores were used to train the algorithms, the best performance was obtained by PLS and LG (accuracy ¼ 0.989; Kappa ¼ 0.978; Table 9). As in the former case, the ensemble models here failed in the attempt to improve the performance of the base learners (accuracy ¼ 0.978; kappa ¼ 0.955).

The hyperparameter configuration of the best performance models is shown in Table 10.

4.2. Analysis of the material from the Estrecho Cave

The performing of the Principal Component Analysis (Fig. 4) of all samples (both modern and Estrecho Cave Mus samples) has shown again an overall differential position of M. spretus and M. m. domesticus, which appear mostly standing apart on both sides of the x-axis, nevertheless showing a high degree of overlap.

From what is seen in the plot in which the sample from the Estrecho Cavewas included, most part of the fossil specimens seem to be closer to the M. spretus sample. Nevertheless, some specimens appear within the area occupied by the M. m. domesticus confidence interval, indicating that probably both species are represented in the sample.

As could be seen in the previous section, the best performance obtained classifying the modern sampleswas obtained with the use of 82 and 7 PC scores using ML algorithms. However, an important difference exists between these different samples.

Although in both cases, stacking techniques have provided accuracy values of 1, when 7 PC scores were used, there are strong correlations between base learners and this situation can generate

overfitting when the stacked model is performed. For this reason, here we decided to approach the classification of the fossil data only using 82 PC scores, excluding the classical LDA and analyses with the ML methods of the samples of 7, 14 and 27 PC scores.

It must be noted that the application of ensemble techniques is probably useful due to the general absence of pairwise correlation between the algorithms (Table 11). The higher correlation value was obtained was 0.69, found between LG and SVMl, a value that can be considered as low.

Similarly to what occurred with the classification of the modern sample, the best ML performance was given by the stacked model using a NNET, which provided a perfect classification rate of the modern sample (accuracy/kappa ¼ 1) (Table 12). The other ensemble models provided as well high accuracy values, between 0.978 (GLM and GB) and 0.989 (RF). The application of ensemble techniques was useful in the case of the stacked model using a NNET, as it improved the best performance shown by a base learner (accuracy ¼ 0.989).

However, although accuracy values among models are similar, it must be noted that the classification of the fossil specimens indicate clear differences of performance between these algorithms. For instance, when the base learners were used to identify the m1 sample of the Estrecho Cave, some of them assigned a high number of teeth to M. m. domesticus (DTC5.0, RF, GB and NB), while the most accurate model (NNET) only identified 4 teeth as belonging to M. m. domesticus. In the case of the Ensemble Learning algorithms, all classified as M. m. domesticus between 5 (GLM and NNET) and 3 (RF and GB) teeth. The hyperparameter configuration of these models is shown in Table 10.

As can be seen in Supplementary File 3, the classification results provided by the ensembled models are quite similar among them. However, they are showing certain discrepancies that must be considered.

Considering the posterior probability values (p) of assignation to one or another species obtained for each specimen using the ensemble models, between 43 and 48 out of the 48 (89.58%e100%) Mus m1 of the Estrecho Cave were classified with posterior probabilities above the 0.9 thresholds (Table 13) (Supplementary File 3). Thus, we think it best to use all the models together instead of only one if reliable results classifying fossil material is intended.

41 teeth have been indisputably identified as Mus spretus and only 2 teeth (specimens CE4 and CE17) as Mus musculus domesticus by all ensemble models. For these cases, all these teeth can be considered as successfully classified. However, the remaining specimens show some peculiarities that deserve being commented.

In the case of specimen CE1, GLM and NNET have classified the tooth ($p < 0.9$) as M. musculus domesticus while RF and GB have identified the specimen as Mus spretus (RF p ¼ <0.9; GB p ¼ > 0.9). Thus, the classification of this tooth shows too much problematic to classify it and it should remain for the time being as indeterminate.

A similar situation occurs with specimen CE14, which was classified as M. m. domesticus by GLM and NNET models (GLM p ¼ < 0.9; NNET p ¼ > 0.9) and as M. spretus by RF and GB (RF p ¼ < 0.9; GB p ¼ > 0.9). Thus, we also leave this tooth as indeterminate.

A different case is found in specimen CE18, which was clearly (p ¼ >0.9) classified as M. m. domesticus by NNET and GB, while although GLM and RF also classified the tooth as belonging to M. m. domesticus, these models gave a low p-value for this result. Therefore, we consider reasonable to classify this specimen as M. m. domesticus. A similar situation, but involving M. spretus is seen in specimens C128 and C135, which have been assigned to it (see Supplementary File 3).

Thus, after this assessment of the classification results, 43 specimens have been assigned to Mus spretus and only 3 to M. m. domesticus (Fig. 5), leaving only 2 specimens as indeterminate.

5. Discussion

The use of Machine Learning techniques has recently contributed to improve several approaches to palaeobiological issues such as the characterization of skeletal part profiles in archaeological assemblages (Arriaza and Domínguez-Rodrigo, 2016), the classification of different bone fracture

patterns (Mocl_an et al., 2019, 2020) or even the identification of differences among carnivore tooth marks (Courtenay et al., 2019).

The application of this type of techniques to the classification of fossils is very promising, since it provides objective mechanisms to proceed with the taxonomic assignments of material that, due to its fragmentary nature, usually provides less complete information compared to what is available in the classification of recent specimens, and reducing the subjective bias introduced by the taxonomist. Although this type of approach has been previously applied to different taxa with a high confidence performance (e.g. Miele et al., 2020; Monson et al., 2018; Villon et al., 2018; Wills et al., 2021), it is still rarely applied today in palaeontological studies.

In this paper we have shown how the use of stacking techniques with a Neural Network algorithm to distinguish between M. m. domesticus and M. spretus with modern samples has given a perfect accuracy performance. Here, the achievement attained after using Artificial Intelligence was an improved the accuracy rate from a 95.31% (Valenzuela-Lamas et al., 2011) to 100%.

This accomplishment is especially important in the fields of Palaeontology and Archaeology, since many relevant interpretations rely on the correct identification of the taxa present in the fossil assemblages, such as the reconstruction of past environmental conditions, their changes through time, or even in the case of more recent times, to clarify anthropogenic biological invasions of commensals such as the house mouse (Cucchi, 2008; Cucchi et al., 2005b, 2011, 2020; Cucchi and Vigne, 2006; Michaux et al., 2007; Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011).

However, several aspects must be considered in order to better understand the results obtained by the models used here. The sample size of the modern sample of this paper and its intraspecific variability seems to be more adequate than others (Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011) if we take into account that, with the “classical” LDA analysis, the accuracy value that was obtained when applied to classify the modern specimens reached better results (14 PC scores ¼ 0.955) than those previously obtained by other researchers (Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011). Nevertheless, the difference obtained with the latter is extremely weak (0.2%).

There is a clear improvement in the results obtained using ML algorithms compared to LDA. There are nine ML algorithms (with 82 PC scores) that provided a higher number of correct scores in the modern samples than those obtained through LDA with LOOCV, besides better classification results with all the ensembled/stacked models. This is in agreement with Dietterich (2000), who showed that ensemble methods can perform better classifications than any single classifier. A similar conclusion has been reached for the taxonomic identification of theropod teeth (Wills et al., 2021). Thus, from our results we consider that Ensemble Learning combined withGMMis the best method (excluding the DNA studies) to detect the presence of either M. m. domesticus or M. spretus, or both, in palaeontological and archaeological material.

However, it must be noted that some differences of algorithmic performance are detected when the ensembled/stacked models classified the palaeontological samples. Although all of them showed extremely high accuracy values (>0.97) classifying the modern sample, the number of specimens classified as M. m. domesticus/M. spretus in the Estrecho Cave varied depending on the model used.

In our opinion, the best performance of the stacked model is that which was trained using a Gradient Boosting algorithm. Although this model gives an accuracy value below 1, when it was used with the modern material, it classified correctly all the specimens giving a posterior probability above 0.9. Having this in mind, it is possible that the perfect classification provided by the stacked NNET model is due to overfitting, nevertheless to a low degree. In any case, as other studies have also pointed out (e.g. Domínguez-Rodrigo et al., 2020), if different classification rates provided by different algorithms are taken into account at the same time, the interpretation of the classification will be the more conservative solution and, this way, we will avoid making erroneous interpretations in most cases.

For the particular case of classifying isolated m1 of M. m. domesticus and M. spretus, our results show that the ML techniques applied to GMM data substantially outperform both LDA and traditional morphometric analyses in obtaining precise taxonomic assignations of fossil materials

specimens (Darviche et al., 2006; Darviche and Orsini, 1982; Gerasimov et al., 1990; Stoetzel et al., 2013; Valenzuela-Lamas et al., 2011).

Concerning the results achieved in the taxonomic identification of the Mus material from the Estrecho Cave, our results are relevant for providing new highly reliable information on the co-occurrence of M. m. domesticus and M. spretus at the site. Until now, M. m. domesticus in the Estrecho Cave had not been correctly confirmed (Domínguez García et al., 2019a, 2019b).

Through this methodology, to identify a few specimens of M. m. domesticus in a fossil assemblage dominated by M. spretus has been possible. Probably, the co-occurrence of two Mus taxa in other assemblages in which one of the two species is distinctly more abundant has been overlooked more than once due to misclassification (e.g. Cucchi et al., 2013). In this way, GMM and ML techniques are a promising new approach to obtain reliable taxonomic identifications in the numerous archaeological and palaeontological sites in which Mus is represented, both for new findings as for older material that remained undetermined due to imprecise taxonomic results with traditional methods/criteria (Domínguez García et al., 2019a).

Following previous works (Domínguez García et al., 2019b, 2020), the small mammal remains of the Estrecho Cave can be considered as a single and homogenous assemblage with the same origin and age according to stratigraphy, taphonomy and radiocarbon dating. Thus, the higher abundance of M. spretus in relation toM. m. domesticus could be explained by the ecology of each of the two species as well as the archaeological context of the site. The house mouse is a commensal species which hence is found associated to human settlements (Wilson et al., 2017), while the western Mediterranean mouse is not commensal with humans and is more abundant in natural environments. Moreover, the cave was just used by humans as a funerary place during the Bronze Age and not as habitat, a situation that it is not favourable to the presence of M. m. domesticus in the site. After to this use, mouse remains were accumulated from predation by a small to medium-sized mammal carnivore (Domínguez García et al., 2019b), which hunted the most abundant species in its low anthropised environment. This is in agreement with the few fossil occurrences of M. m. domesticus in the western Mediterranean region that can be considered reliable according to Domínguez García et al. (2019a). These records come only from human settlements of the Iron Age and early Roman Period (~2.8e2 kyr BP): at Alorda Park (Valenzuela-Lamas et al., 2011) and Estrets-Rac_o de Rata (Guillem Calatayud, 2011) in eastern Iberia, as well as at Rirha in Morocco (Oueslati et al., 2020), both species occurred with a majority of house mouse, while at La Mota or el Soto de la Medinilla (Morales Mu~niz et al.,1995) in Spain, and at Lattara in Mediterranean France (Poitevin and S_en_egas, 1999), only M. m. domesticus was identified.

The possibility to obtain a precise identification of both species in the small mammal assemblage of the Estrecho Cave using the methods explained in this paper, has thus added new valuable data concerning the history of the genus Mus in the western Mediterranean region. These reliable results demonstrate that both species were already present in central Iberia c. 2300-2150 years ago, supporting the available palaeontological and genetic data concerning the age of arrival and the way in which these two species colonised the Iberian Peninsula (Bonhomme et al., 2011; Boursot et al., 1985; Cucchi et al., 2005b; Domínguez García et al., 2019a; Lalis et al., 2019). The presence of M. m. domesticus points to a quick dispersal process that took place from coastal areas to inland Iberia, since it is supposed to have arrived after 3000 B P through navigation routes. The abundant presence of M. spretus in the Estrecho Cave provides one of the fewfossil records of this species in Europe, adding not only evidence of its colonisation and dispersal process, but also of the expansion in the Mediterranean climate region of the Iberian Peninsula since its arrival during the Late Neolithic.

Finally, all these results are most remarkable given the limited differentiation in the shape of m1s observed between these two closely related species, as shown by the high overlapping in the PCA (Figs. 3 and 4). This is most probably related to the relatively recent divergence between M. musculus and M. spretus estimated around 1e3 Ma (Chevret et al., 2005; She et al., 1990). In addition, genetic bidirectional introgressions have been demonstrated between natural populations of both species in Africa and Europe (Banker et al., 2022; Liu et al., 2015; Orth et al., 2002). Although this phenomenon is limited, these studies showed various ancient and recent interbreeding events. Thus, even though

genetic introgression does not necessarily affect phenotypes, it may generate significant morphological variations (Renaud et al., 2012), possibly affecting tooth morphology, an issue that is still to be addressed. In that case, our results could possibly be already influenced by unnoticed morphological variability consequences of hybridisation (e.g. could this be the reason for the disagreement obtained between algorithms and the indeterminate classification of some specimens?). Therefore, further research including genomic data in morphometric analysis using ML techniques should be addressed in order to test this hypothesis and to move forward improving systematic and taxonomic methodologies.

6. Conclusions

Machine Learning algorithms combined with the use of Geometric Morphometrics applied to the outline shape of the first lower molars of M. musculus domesticus and Mus spretus have provided the highest accuracy rate obtained so far to discriminate both species. Here we present a methodological approach to distinguish these two taxa using Artificial Intelligence techniques. Our results have provided a new procedure which gives a nearperfect classification rate based on the interspecific differences in molar shape from modern populations of the western Mediterranean.

Machine learning methods have provided high accuracy values when the maximum number of PC scores were employed, which implies that a higher percentage of the accumulated variance of the samplewas included in the analysis, instead of the reduced number of PC scores that usually have been employed in previous analyses.

The use of ensemble/stacking trained algorithms for the classification of Mus materials of the Estrecho Cave has provided robust identifications of both species in the assemblage. A combination of the results provided by the four ensembled methods used in this study has allowed a classification of 43 specimens as Mus spretus, 3 as M. m. domesticus and 2 as Mus sp.

Although the possible effects of hybridisation between these two closely related species on tooth morphology have not been tested yet, these could influence the application of this method for taxonomic purposes. Therefore, a new line of research combining genomic and morphometric data is open.

The good results obtained in this study leads us to encourage on the use of ensemble ML techniques as an alternative approach to be used single or combined with classical classification methods. The use of such methods on isolated Mus teeth demonstrates that high levels of predictive taxonomic accuracy are possible from GMM data. An example is given of how the application of these methods to fossil material of Mus spp. From a Spanish site provided sound results useful to clarify biogeographic issues concerning species dispersal in the past relevant to understand recent patterns of distribution of the species involved.

Credit author statement

Abel Mocl_an: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing-Original draft, Visualization. _Angel C. Domínguez-García: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing-Original draft, Visualization. Emmanuelle Stoetzel: Resources, Funding acquisition, Project administration, Writing – Review & Editing, Supervision. Thomas Cucchi: Resources, Writing - Review& Editing. Paloma Sevilla: Project administration,Writing - Review & Editing, Supervision. C_esar Laplana: Conceptualization, Project administration, Writing- Review & Editing, Supervision.

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.

Data availability

Data will be made available on request..

Acknowledgements
The authors are very grateful to Irene Prieto Saiz (University of Castilla la Mancha) for notifying the existence of the microvertebrate deposits at the Estrecho Cave and to “Lapis Specularis” team, especially to Juan Carlos Guisado di Monti and María Jos_e Bern_ardez G_omez for enabling to carry out the sampling of microvertebrate remains in the cave, among which were found the fossil Mus specimens used in this study.Wewould also like to thank C. Urdiales Alonso, curator of the Vertebrate Collections at the Estaci_on Biol_ogica de Do˜nana (EBD-CSIC, Sevilla) and _A.L. Garvía Rodríguez, curator of the Mammals Collections at the Museo Nacional de Ciencias Naturales (MNCN-CSIC, Madrid), who granted us access to modern mice collections from Spain. AMwas funded by a grant from the Junta de Castilla y Le_on financed in turn by the European Social Funds through the Consejería de Educaci_on (BDNS 376062). ACDG was funded by a Postdoctoral Grant (POP-UCMCT17/ 17-CT18/17) and by a Research Stay Grant (UCM 2020 e EB25/ 20), both financed by the Complutense University of Madrid and co-financed by Santander Bank. This work has also benefited from support from the SOuMed project “Approche pluridisciplinaire de la diffusion des souris commensales et sauvages dans l’Ouest de la M_editerran_ee” (E. Stoetzel dir.) from the D_epartement Homme & Environnement of the Mus_eum national d’Histoire naturelle of Paris. This work is a contribution of the Research Group UCM 970827 on Quaternary Ecosystems of the Complutense University of Madrid.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.quascirev.2022.107877.

References

Abellan, N., Baquedano, E., Domínguez-Rodrigo, M., 2022. High-accuracy in the classification of butchery cut marks and crocodile tooth marks using machine learning methods and computer vision algorithms. Geobios (Jodhpur).

Andrews, P., 1995. Mammals as palaeoecological indicators. Acta Zool. Cracov. (Engl. Transl.) 38, 59e72.

Arriaza, M.C., Domínguez-Rodrigo, M., 2016. When felids and hominins ruled at Olduvai Gorge: a machine learning analysis of the skeletal profiles of the nonanthropogenic Bed I sites. Quat. Sci. Rev. 139, 43e52.

Avery, D.M., 2007. Micromammals as palaeoenvironmental indicators of the southern African Quaternary. Trans. Roy. Soc. S. Afr. 62, 17e23.

Banker, S.E., Bonhomme, F., Nachman, M.W., 2022. Bidirectional introgression between Mus musculus domesticus and Mus spretus. Genome Biology and Evolution 14, evab288.

Bellin, N., Calzolari, M., Callegari, E., Bonilauri, P., Grisendi, A., Dottori, M., Rossi, V., 2021. Geometric morphometrics and machine learning as tools for the identification of sibling mosquito species of the Maculipennis complex (Anopheles). Infect. Genet. Evol. 95, 105034.

Bonhomme, F., Orth, A., Cucchi, T., Rajabi-Maham, H., Catalan, J., Boursot, P., Auffray, J.-C., Britton-Davidian, J., 2011. Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. Proc. Biol. Sci. 278, 1034e1043.

Bookstein, F.L., 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med. Image Anal. 1, 225e243.

Boursot, P., Jacquart, T., Bonhomme, F., Britton-Davidian, J., Thaler, L., 1985. Geographic differentiation of the mitochondrial genome in Mus spretus Lataste. C R acad sci III, serie III. Sciences de la vie 301, 161e166.

Britton, J., Pasteur, N., Thaler, L., 1976. Les souris du Midi de la France: caracterisation genetique de deux groupes de populations sympatriques. C R Acad Sci, Serie D 258, 515e518.

Chaline, J., Brunet-Lecomte, P., Montuire, S., Viriot, L., Courant, F., 1999. Anatomy of the arvicoline radiation (Rodentia): palaeogeographical, palaeoecological history and evolutionary data. Ann. Zool. Fenn. 36, 239e267.

Chevret, P., Veyrunes, F., Britton-Davidian, J., 2005. Molecular phylogeny of the genus Mus (Rodentia: murinae) based on mitochondrial and nuclear data. Biol. J. Linn. Soc. 84, 417e427.

Courtenay, L.A., Gonzalez-Aguilera, D., 2020. Geometric morphometric data augmentation using generative computational learning algorithms. Appl. Sci. 10, 9133.

Courtenay, L.A., Yravedra, J., Huguet, R., Aramendi, J., Mate-Gonzalez, M.A.,Gonzalez-Aguilera, D., Arriaza, M.C., 2019. Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks. Palaeogeogr. Palaeoclimatol. Palaeoecol. 522, 28e39.

Cucchi, T., 2008. Uluburun shipwreck stowaway house mouse: molar shape analysisand indirect clues about the vessel's last journey. J. Archaeol. Sci. 35, 2953e2959.

Cucchi, T., Balas¸escu, A., Bem, C., Radu, V., Vigne, J.-D., Tresset, A., 2011. New insights into the invasive process of the eastern house mouse (Mus musculus musculus): evidence from the burnt houses of Chalcolithic Romania. Holocene 21, 1195e1202.

Cucchi, T., Kovacs, Z.E., Berthon, R., Orth, A., Bonhomme, F., Evin, A., Siahsarvie, R., Darvish, J., Bakhshaliyev, V., Marro, C., 2013. On the trail of Neolithic mice and men towards Transcaucasia: zooarchaeological clues from Nakhchivan (Azerbaijan). Biol. J. Linn. Soc. 108, 917e928.

Cucchi, T., Orth, A., Auffray, J.-C., Renaud, S., Fabre, L., Catalan, J., Hadjisterkotis, E., Bonhomme, F., Vigne, J.-D., 2005a. A new endemic species of the subgenus Mus (Rodentia, Mammalia) on the Island of Cyprus. Zootaxa 1241, 1e36.

Cucchi, T., Papayianni, K., Cersoy, S., Aznar-Cormano, L., Zazzo, A., Debruyne, R., Berthon, R., Balaşescu, A., Simmons, A., Valla, F., Hamilakis, Y., Mavridis, F., Mashkour, M., Darvish, J., Siahsarvi, R., Biglari, F., Petrie, C.A., Weeks, L., Sardari, A., Maziar, S., Denys, C., Orton, D., Jenkins, E., Zeder, M., Searle, J.B., Larson, G., Bonhomme, F., Auffray, J.-C., Vigne, J.-D., 2020. Tracking the Near Eastern origins and European dispersal of the western house mouse. Sci Rep 10, 8276.

Cucchi, T., Vigne, J.-D., 2006. Origin and diffusion of the house mouse in the mediterranean. Human Evolution 21, 95.

Cucchi, T., Vigne, J.-D., Auffray, J.-C., 2005b. First occurrence of the house mouse (Mus musculus domesticus Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biological Journal of the Linnean Society 84, 429e445.

Cucchi, T., Vigne, J.-D., Auffray, J.-C., Croft, P., Peltenburg, E., 2002. Introduction involontaire de la souris domestique (Mus musculus domesticus) a Chypre des le Neolithique preceramique ancien (fin IXe et VIIIe millenaires av. J.-C.). Comptes Rendus Palevol 1, 235e241.

Cuenca-Bescos, G., Blain, H.-A., Rofes, J., Lopez-García, J.M., Lozano-Fernandez, I., Galan, J., Nú˜nez-Lahuerta, C., 2016. Updated Atapuerca biostratigraphy: smallmammal distribution and its implications for the biochronology of the Quaternary in Spain. Comptes Rendus Palevol, Biochronology, biostratigraphy, and paleoecology of the Quaternary Biochronologie, la biostratigraphie et la paleoecologie du Quaternaire 15, 621e634.

Darviche, D., Orsini, P., 1982. Criteres de differenciation morphologique et biom etrique de deux especes de souris sympatriques : Mus spretus et Mus musculus domesticus. Mammalia 46, 205e218.

Darviche, D., Orth, A., Michaux, J., 2006. Mus spretus et M. musculus (Rodentia, Mammalia) en zone mediterraneenne: differenciation biometrique et morphologique: application a des fossiles marocains pleistocenes/Mus spretus and M. musculus (Rodentia, Mammalia) in the Mediterranean zone: biometric and morphological differentiation: application to Pleistocene Moroccan fossils. Mammalia 70, 90e97.

Deane-Mayer, Z.A., 2019. A Brief Introduction to caretEnsemble [WWW Document]. CRAN. URL. https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html. accessed 11.12.21.

Deane-Mayer, Z.A., Knowles, J.E., 2019. Package 'caretEnsemble. '. Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 1e15.

Domínguez García, A.C., Laplana, C., Sevilla, P., 2020. Early reliable evidence of the Etruscan shrew (Suncus etruscus) in southwestern Europe during ancient times. Reconstructing its dispersal process along the Mediterranean Basin. Quaternary Science Reviews 250, 106690.

Domínguez García, A.C., Laplana, C., Sevilla, P., Blain, H.-A., Zumajo, N.P., Benítez de Lugo Enrich, L., 2019a. New data on the introduction and dispersal process of small mammals in southwestern Europe during the Holocene: castillejo del Bonete site (southeastern Spain). Quaternary Science Reviews 225, 106008.

Domínguez García, A.C., Laplana, C., Sevilla, P., Guisado Di Monti, J.C., Bernandez Gomez, M.J., 2019b. Tafonomía y cronología de la asociacion de micromamíferos de la Cueva del Estrecho (Villares del Saz, Cuenca, Espa~na). Spanish Journal of Palaeontology 34, 241e256.

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? Archaeol Anthropol Sci 11, 2711e2725.

Domínguez-Rodrigo, M., Cifuentes-Alcobendas, G., Jimenez-García, B., Abellan, 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.

Fox, N.S., Veneracion, J.J., Blois, J.L., 2020. Are geometric morphometric analyses replicable? Evaluating landmark measurement error and its impact on extant and fossil Microtus classification. Ecology and Evolution 10, 3260e3275.

Gerasimov, S., Nikolov, H., Mihailova, V., Auffray, J.-C., Bonhomme, F., 1990. Morphometric stepwise discriminant analysis of the five genetically determined European taxa of the genus Mus. Biological Journal of the Linnean Society 41, 47e64.

Guillem Calatayud, P.M., 2011. Els paisatges ramaders en epoca iberica. Una reconstruccio a partir dels micromamífers. Arqueo Mediterrania 12, 117e121.

Guisado Di Monti, J.C., Bernandez Gomez, M.J., 2016. Cueva del Estrecho en Villares del Saz. Adaptacion de la Cueva a Uso Turístico. In: Ruiz-Checa, J.M., Cristini, V. (Eds.), Actuaciones Sobre El Patrimonio Historico y Medioambiental. Plan de Mejoras Turísticas, Provincia de Cuenca (Plamit 2011-2015). Diputacion Provincial de Cuenca, Cuenca, pp. 59e61.

Herranz-Rodrigo, D., Tardaguila-Giacomozzi, S.J., Courtenay, L.A., Rodríguez-Alba, J.-J., Garrucho, A., Recuero, J., Yravedra, J., 2021. New geometric morphometric insights in digital taphonomy: analyses into the sexual dimorphism of felids through their tooth pits. Applied Sciences 11, 7848.

John, C.R., 2020. MLeval: Machine Learning Model Evaluation. Kim, D.H., MacKinnon, T., 2018. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 73, 439e445.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Presented at the 26th Annual Conference on Neural Information Processing Systems. Currant Associates Inc., U.S.A., pp. 1097e1105 Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer-Verlag, New York.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., , R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2020. Caret: Classification and Regression Training.

Lalis, A., Mona, S., Stoetzel, E., Bonhomme, F., Souttou, K., Ouarour, A., Aulagnier, S., Denys, C., Nicolas, V., 2019. Out of Africa: demographic and colonization history of the Algerian mouse (Mus spretus Lataste). Heredity 122, 150e171.

Lantz, B., 2013. Machine Learning with R. Packt Publishing, Birmingham.

Laplana, C., Sevilla, P., 2013. Documenting the biogeographic history of Microtus cabrerae through its fossil record. Mammal Review 43, 309e322.

Liu, K.J., Steinberg, E., Yozzo, A., Song, Y., Kohn, M.H., Nakhleh, L., 2015. Interspecific introgressive origin of genomic diversity in the house mouse. Proceedings of the National Academy of Sciences 112, 196e201.

Lopez García, J.M., 2011. Los micromamíferos del Pleistoceno superior de la Península Iberica: Evolucion de la diversidad taxonomica y cambios paleoambientales y paleoclimaticos. Editorial Academica Española, Madrid.

Michaux, J., Cucchi, T., Renaud, S., Garcia-Talavera, F., Hutterer, R., 2007. Evolution of an invasive rodent on an archipelago as revealed by molar shape analysis: the house mouse in the Canary Islands. Journal of Biogeography 34, 1412e1425.

Miele, V., Dussert, G., Cucchi, T., Renaud, S., 2020. Deep Learning for Species Identification of Modern and Fossil Rodent Molars.

Moclan, A., Domínguez-Rodrigo, M., Yravedra, J., 2019. Classifying agency in bone breakage: an experimental analysis of fracture planes to differentiate betweenhominin and carnivore dynamic and static loading using machine learning (ML) algorithms. Archaeol Anthropol Sci 11, 4663e4680.

Moclan, A., Huguet, R., Marquez, B., Laplana, C., Arsuaga, J.L., Perez-Gonzalez, A., Baquedano, E., 2020. Identifying the bone-breaker at the Navalmaíllo Rock Shelter (Pinilla del Valle, Madrid) using machine learning algorithms. Archaeol Anthropol Sci 12, 46.

Monson, T.A., Armitage, D.W., Hlusko, L.J., 2018. Using machine learning to classify extant apes and interpret the dental morphology of the chimpanzee-human last common ancestor. PaleoBios 35, 1e20.

Morales Mu˜niz, A., Cereijo Pecharroman, M.A., Hernandez Carrasquilla, F., Liesau von Lettow-Vorbeck, C., 1995. Of mice and sparrows: commensal faunas from the Iberian iron age in the duero valley (central Spain). International Journal of Osteoarchaeology 5, 127e138.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. PNAS 115, E5716eE5725.

Opitz, D., Maclin, R., 1999. Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research 11, 169e198.

Ortega Martínez, A.I., Martín Merino, M.A., 1992. Informe sobre el descubrimiento de tres cuevas arqueologicas en el termino municipal de Villares del Saz (Cuenca). Servicio de Investigaciones Espeleologicas. Diputacion provincial de Burgos, Spain.

Orth, A., Belkhir, K., Britton-Davidian, J., Boursot, P., Benazzou, T., Bonhomme, F., 2002. Hybridation naturelle entre deux especes sympatriques de souris Mus musculus domesticus L. et Mus spretus Lataste. Comptes Rendus Biologies 325, 89e97.

Oueslati, T., Kbiri Alaoui, M., Ichkhakh, A., Callegarin, L., de Chazelle, C.-A., Rocca, E., Carrato, et al., 2020. 1st century BCE occurrence of chicken, house mouse and black rat in Morocco: socio-economic changes around the reign of Juba II on the site of Rirha. Journal of Archaeological Science: Reports 29, 102162.

Papayiannis, K., 2012. The micromammals of Minoan Crete: human intervention in the ecosystem of the island. Palaeobio Palaeoenv 92, 239e248.

Poitevin, F., Senegas, F., 1999. Les micromammiferes du site de Lattara. In: Py, M. (Ed.), Recherches Sur La Quatrieme Siecle Avant Notre Ere a Lattes, Lattara. CNRS Editions, Paris, pp. 609e635.

Quenu, M., Trewick, S.A., Brescia, F., Morgan-Richards, M., 2020. Geometric morphometrics and machine learning challenge currently accepted species limits of the land snail Placostylus (Pulmonata: bothriembryontidae) on the Isle of Pines, New Caledonia. Journal of Molluscan Studies 86, 35e41.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. Renaud, S., 2005. First upper molar and mandible shape of wood mice (Apodemus sylvaticus) from northern Germany: ageing, habitat and insularity. Mammalian Biology 70, 157e170.

Renaud, S., 1999. Size and shape variability in relation to species differences and climatic gradients in the African rodent Oenomys. Journal of Biogeography 26, 857e865.

Renaud, S., Alibert, P., Auffray, J.-C., 2012. Modularity as a source of new morphological variation in the mandible of hybrid mice. BMC Evolutionary Biology 12, 141.
Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., 2020. MASS: Support Functions and Datasets for Venables and Ripley's MASS.
Rohlf, F.J., 2019. Tps Relative Warps.
Rohlf, F.J., 2018. tpsUtil.
Rohlf, F.J., 2010. tpsDig.
Rokach, L., 2010. Ensemble-based classifiers. Artif Intell Rev 33, 1e39.
Romero, I.C., Kong, S., Fowlkes, C.C., Jaramillo, C., Urban, M.A., Oboh-Ikuenobe, F., D'Apolito, C., Punyasena, S.W., 2020. Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. PNAS 117, 28496e28505.
Sagi, O., Rokach, L., 2018. Ensemble learning: a survey. WIREs Data Mining and Knowledge Discovery 8, e1249.
Sevillano, V., Holt, K., Aznarte, J.L., 2020. Precise automatic classification of 46 different pollen types with convolutional neural networks. PLOS ONE 15, e0229751.
She, J.X., Bonhomme, F., Boursot, P., Thaler, L., Catzeflis, F., 1990. Molecular phylogenies in the genus Mus: comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. Biological Journal of the Linnean Society 41, 83e103.
Stoetzel, E., Denys, C., Michaux, J., Renaud, S., 2013. Mus in Morocco: a Quaternary sequence of intraspecific evolution. Biological Journal of the Linnean Society 109, 599e621.
Tan, M., Le, Q.V., 2020. EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning.
Valenzuela-Lamas, S., Baylac, M., Cucchi, T., Vigne, J.-D., 2011. House mouse dispersal in Iron Age Spain: a geometric morphometrics appraisal. Biological Journal of the Linnean Society 102, 483e497.
Villon, S., Mouillot, D., Chaumont, M., Darling, E.S., Subsol, G., Claverie, T., Villeger, S., 2018. A Deep learning method for accurate and fast identification of coral reef fishes in underwater images. Ecological Informatics 48, 238e244.
Weissbrod, L., Marshall, F.B., Valla, F.R., Khalaily, H., Bar-Oz, G., Auffray, J.-C., Vigne, J.-D., Cucchi, T., 2017. Origins of house mice in ecological niches created by settled hunter-gatherers in the Levant 15,000 y ago. PNAS 114, 4099e4104.
Wills, S., Underwood, C.J., Barrett, P.M., 2021. Learning to see the wood for the trees: machine learning, decision trees, and the classification of isolated theropod teeth. Palaeontology 64, 75e99.
Wilson, D.E., Lacher Jr., T.E., Mittermeier, R.A. (Eds.), 2017. Handbook of the Mammals of the World. Rodents II, vol. 7. Lynx Edicions, Barcelona, Spain.

FIGURES

Fig. 1. Geographic location of the Estrecho Cave in the Iberian Peninsula (created with Natural Earth, www.naturalearthdata.com).

Fig. 2. Example of the orientation and location of the landmark and semi-landmarks on a first lower molar of Mus spp.

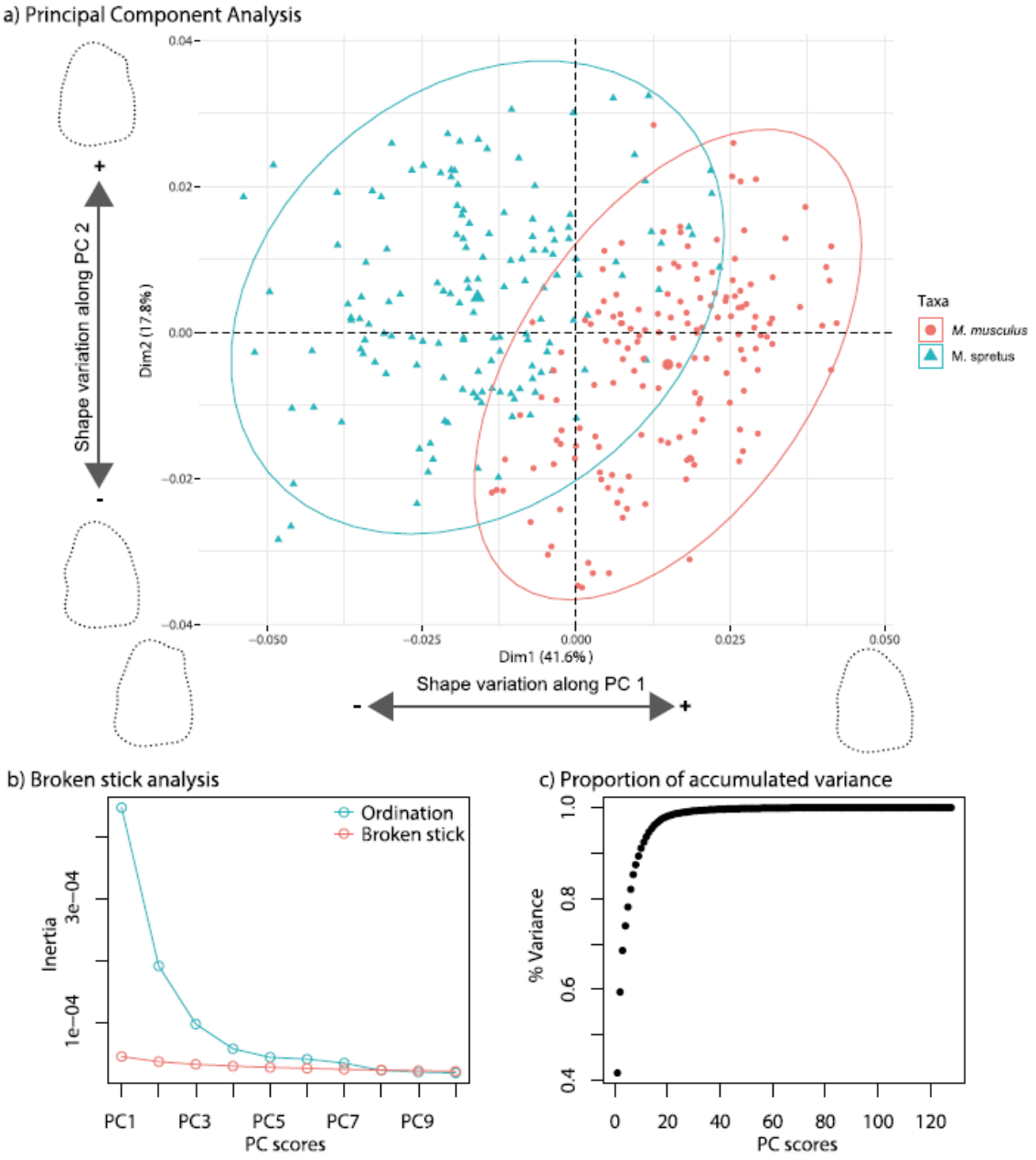

Fig. 3. a) Biplot of the two first principal component scores of the modern sample; b) results provided by the application of the Broken Stick test to the PCA results; c) percentage of the accumulated variance of the PC scores.

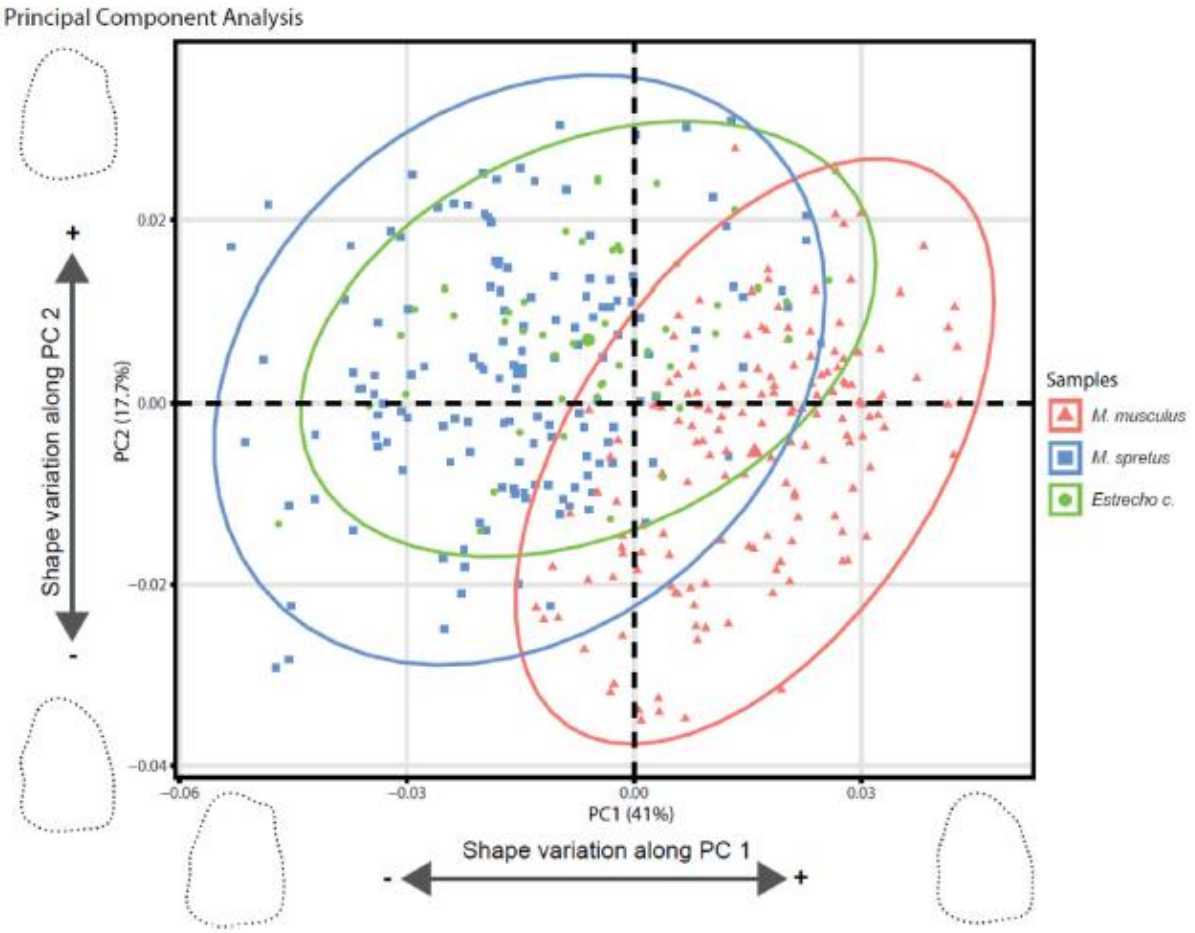

Fig. 4. Biplot of the two first principal components of the modern sample and the Estrecho Cave specimens.

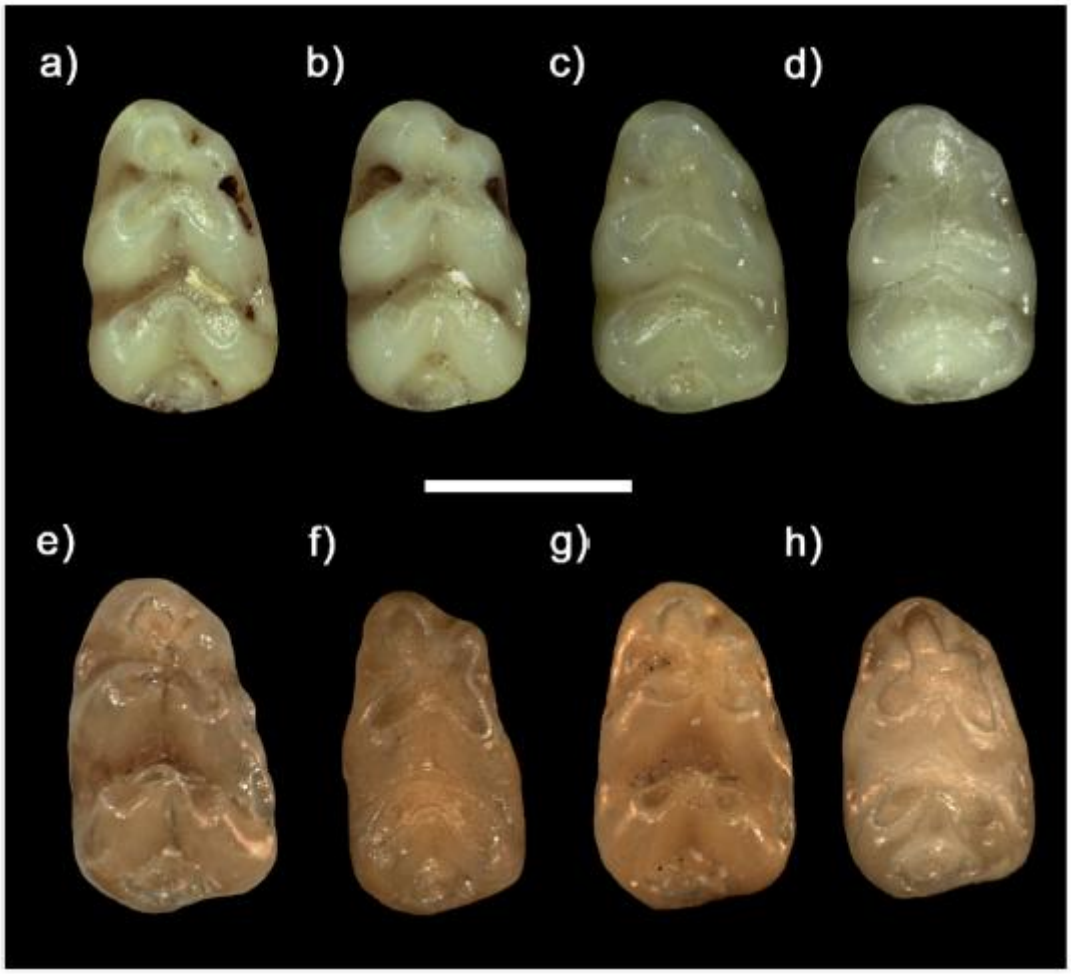

Fig. 5. Selected Mus spp. M1 in occlusal view of the modern and the Estrecho Cave samples. Modern Mus spretus (a: MNCN_9982eb: MNCN_9979); modern Mus musculus domesticus (c: MNCN_3490-d: MNCN_18,920); Mus spretus from the Estrecho Cave (e: CE3-f: CE19); Mus musculus domesticus from the Estrecho Cave (g: CE4-h: CE17). Scale = 1 mm.

TABLEAUX

Order	Species
Rodentia	
	Eliomys quercinus
	Arvicola sapidus
	Microtus cabrerae
	Microtus arvalis-agrestis
	Microtus duodecimcostatus-lusitanicus
	Apodemus sylvaticus-flavicollis
	Mus spretus-musculus
Lagomorpha	
	Oryctolagus cuniculus
Eulipotyphla	
	Erinaceus europaeus
	Crocidura russula
	Suncus etruscus
Chiroptera	
	Rhinolophus ferrumequinum
	Rhinolophus euryale
	Myotis escalerai
	Myotis myotis-blythii

Table 1. Small mammals identified in the assemblage of the Estrecho Cave, taken from Domínguez Gerasimov et al. (1990, 2020)and unpublished data.

Population	Collection	*n*
Modern *Mus musculus domesticus*		
Gran Canaria	EBD, MNCN	14
Lanzarote	EBD, MNCN	11
Iberian Peninsula	EBD, MNCN	61
France	ISEM	18
Algeria	ISEM	17
Morocco	ISEM	36
TOTAL		157
Modern *Mus spretus*		
Mallorca	MNCN	10
Iberian Peninsula	EBD, MNCN	57
France	ISEM	14
Morocco	ISEM, MOHMIE	65
TOTAL		146
Fossils		
Estrecho Cave		48

Table 2. Modern and fossil samples used in this study, according to species, populations, collections and number of specimens studied (n).

	LDA		LDA (LOOCV)	
	Training	**Testing**	**Training**	**Testing**
7 PC scores	0.948	0.944	0.943	0.944
14 PC scores	0.981	0.966	0.967	0.933
27 PC scores	0.99	0.97	0.986	0.955
82 PC scores	0.995	0.966	0.934	0.744

Table 3. Results provided by the Linear Discriminant Analysis when 7,14, 27 and 82 PC scores are used. Note that an important reduction of the accuracy is produced when the testing dataset is classified, and the leave-one-out cross-validation method is applied (especially when 82 PC scores are used). The best results are obtained when the training dataset is analysed (except for the sample with 7 PC scores with LOOCV).

	NNET	SVMl	SVMr	kNN	LG	DTC5.0	RF	GB	NB	LDA	PLS
7 PC scores	0.99	0.98	0.99	0.99	0.98	0.95	0.98	0.97	0.98	0.99	0.99
14 PC scores	1	1	0.99	0.98	1	0.94	0.98	0.99	0.98	1	1
27 PC scores	0.99	0.98	1	0.98	0.98	0.94	0.98	0.98	0.97	0.99	1
82 PC scores	0.99	0.97	0.98	0.98	0.96	0.93	0.96	0.98	0.9	0.98	0.99

Table 4. Results obtained after calculating the AUC-ROC values. Before presenting the results obtained after using the different Machine and Ensemble algorithms, it must be noted that in most of the pairwise comparison, non-correlation between algorithms was detected, and when it was detected, the correlation was in most cases low (Table 5). Only when 7 PC scores were used, correlation appeared to be present in most cases, sometimes giving results indicating strong correlation values, even higher than 80%.

82 PC scores										
	NNET	SVMl	SVMr	kNN	LG	DTC5.0	RF	GB	NB	LDA
SVMl	0.10	–	–	–	–	–	–	–	–	–
SVMr	−0.02	0.44	–	–	–	–	–	–	–	–
kNN	0.41	0.19	0.11	–	–	–	–	–	–	–
LG	0.14	**0.71**	0.39	0.06	–	–	–	–	–	–
DTC5.0	0.17	0.21	0.09	0.41	0.15	–	–	–	–	–
RF	0.24	0.11	0.05	0.50	0.05	**0.53**	–	–	–	–
GB	0.28	0.23	0.14	**0.57**	0.19	0.44	**0.75**	–	–	–
NB	0.15	0.17	0.17	0.29	0.22	0.26	0.27	0.10	–	–
LDA	0.25	0.43	0.42	0.28	0.47	0.24	0.23	0.32	0.24	–
PLS	**0.57**	0.03	0.05	0.45	0.03	0.18	0.17	0.20	0.12	0.28
7 PC scores (Broken stick)										
	NNET	**SVMl**	**SVMr**	**kNN**	**LG**	**DTC5.0**	**RF**	**GB**	**NB**	**LDA**
SVMl	**0.82**	–	–	–	–	–	–	–	–	–
SVMr	**0.62**	**0.63**	–	–	–	–	–	–	–	–
kNN	**0.70**	**0.56**	0.53	–	–	–	–	–	–	–
LG	**0.87**	**0.79**	**0.68**	**0.66**	–	–	–	–	–	–
DTC5.0	0.46	0.47	0.46	0.49	0.48	–	–	–	–	–
RF	**0.64**	**0.59**	**0.61**	**0.72**	**0.72**	**0.58**	–	–	–	–
GB	**0.65**	**0.59**	**0.59**	**0.68**	**0.64**	**0.60**	**0.77**	–	–	–
NB	**0.53**	0.50	**0.70**	**0.66**	**0.61**	0.44	**0.74**	**0.69**	–	–
LDA	**0.87**	**0.82**	**0.66**	**0.65**	**0.80**	0.42	**0.63**	**0.60**	**0.57**	–
PLS	**0.87**	**0.82**	**0.66**	**0.65**	**0.80**	0.42	**0.63**	**0.60**	**0.57**	**1**
14 PC scores (95% Variance)										

82 PC scores										
	NNET	SVMl	SVMr	kNN	LG	DTC5.0	RF	GB	NB	LDA
	NNET	**SVMl**	**SVMr**	**kNN**	**LG**	**DTC5.0**	**RF**	**GB**	**NB**	**LDA**
SVMl	0.36	–	–	–	–	–	–	–	–	–
SVMr	0.44	0.43	–	–	–	–	–	–	–	–
kNN	0.41	0.16	0.47	–	–	–	–	–	–	–
LG	**0.67**	**0.55**	**0.51**	0.30	–	–	–	–	–	–
DTC5.0	0.29	0.20	0.27	0.38	0.25	–	–	–	–	–
RF	0.41	0.35	0.43	**0.51**	0.42	0.39	–	–	–	–
GB	**0.61**	0.35	0.50	**0.55**	0.47	0.47	**0.69**	–	–	–
NB	0.45	0.37	**0.55**	**0.60**	**0.52**	0.30	**0.65**	**0.55**	–	–
LDA	0.39	**0.61**	**0.52**	0.25	**0.51**	0.10	0.27	0.29	0.40	–
PLS	0.38	**0.65**	**0.51**	0.24	**0.51**	0.09	0.30	0.28	0.36	**0.95**
27 PC scores (99% Variance)										
	NNET	**SVMl**	**SVMr**	**kNN**	**LG**	**DTC5.0**	**RF**	**GB**	**NB**	**LDA**
SVMl	0.50	–	–	–	–	–	–	–	–	–
SVMr	0.38	0.40	–	–	–	–	–	–	–	–
kNN	0.40	0.27	0.33	–	–	–	–	–	–	–
LG	0.39	**0.63**	0.15	0.08	–	–	–	–	–	–
DTC5.0	0.32	0.16	0.11	0.39	0.01	–	–	–	–	–
RF	0.36	0.27	0.41	0.46	0.11	0.47	–	–	–	–
GB	**0.56**	0.45	0.47	0.49	0.22	0.42	**0.55**	–	–	–
NB	0.32	0.36	0.37	0.33	0.12	0.34	0.40	**0.59**	–	–
LDA	0.45	**0.53**	0.34	0.33	**0.52**	0.08	0.13	0.39	0.24	–
PLS	**0.51**	0.28	0.41	0.16	0.14	0.17	0.37	0.37	0.17	0.32

Table 5. Results obtained after calculating the pairwise correlation of the different algorithms applied to the modern sample. Values in bold show the presence of correlation between algorithms.

	Accuracy	Kappa	Acc. Lower	Acc. Upper	Sensitivity	Specificity	Bal. Acc.
NNET	0.978	0.955	0.922	0.997	1	0.954	0.977
SVM linear	0.956	0.911	0.89	0.988	0.957	0.954	0.956
SVM radial	0.978	0.956	0.922	0.997	0.979	0.977	0.978
kNN	0.967	0.933	0.906	0.993	1	0.93	0.965
LG	0.956	0.911	0.89	0.988	0.957	0.954	0.956
DTC5.0	0.922	0.844	0.846	0.968	0.957	0.884	0.921
RF	0.978	0.956	0.922	0.997	0.957	1	0.979
GB	0.967	0.933	0.906	0.993	0.957	0.977	0.967
NB	0.822	0.644	0.727	0.895	0.83	0.814	0.822
LDA	0.989	0.978	0.94	0.999	1	0.977	0.988
PLS	0.967	0.933	0.906	0.993	1	0.93	0.965
Ensemble: GLM	0.989	0.978	0.94	0.999	1	0.977	0.988
Stacking: NNET	**1**	**1**	**0.960**	**1**	**1**	**1**	**1**
Stacking: RF	0.978	0.955	0.922	0.997	1	0.954	0.977
Stacking: GB	0.978	0.955	0.922	0.997	1	0.954	0.977

Table 6. Results provided by ML learning and Ensemble Learning algorithms with 82 PC scores of the modern sample. In bold is shown the best algorithmic performance.

	Accuracy	Kappa	Acc. Lower	Acc. Upper	Sensitivity	Specificity	Bal. Acc.
NNET	0.967	0.933	0.906	0.993	1	0.93	0.965
SVM linear	0.967	0.933	0.906	0.993	1	0.93	0.965
SVM radial	0.989	0.978	0.94	0.999	1	0.977	0.988
kNN	0.978	0.955	0.922	0.997	1	0.954	0.977
LG	0.967	0.933	0.906	0.993	1	0.93	0.965
DTC5.0	0.944	0.888	0.875	0.982	0.979	0.907	0.943
RF	0.967	0.933	0.906	0.993	0.957	0.977	0.967
GB	0.967	0.933	0.906	0.993	0.957	0.977	0.967
NB	0.989	0.978	0.94	0.999	1	0.977	0.988
LDA	0.967	0.933	0.906	0.993	1	0.93	0.965
PLS	0.967	0.933	0.906	0.993	1	0.93	0.965
Ensemble: GLM	0.989	0.978	0.94	0.999	1	0.977	0.988
Stacking: NNET	**1**	**1**	**0.96**	**1**	**1**	**1**	**1**
Stacking: RF	0.978	0.955	0.922	0.997	1	0.954	0.977
Stacking: GB	**1**	**1**	**0.96**	**1**	**1**	**1**	**1**

Table 7. Results provided by ML learning and Ensemble Learning algorithms with 7 PC scores of the modern sample. In bold is shown the best algorithmic performance.

	Accuracy	Kappa	Acc. Lower	Acc. Upper	Sensitivity	Specificity	Bal. Acc.
NNET	0.967	0.933	0.906	0.993	1	0.93	0.965
SVM linear	0.967	0.933	0.906	0.993	1	0.93	0.965
SVM radial	0.922	0.844	0.846	0.968	0.957	0.884	0.921
kNN	**0.978**	**0.955**	**0.922**	**0.997**	**1**	**0.954**	**0.977**
LG	0.956	0.911	0.89	0.988	1	0.907	0.954
DTC5.0	0.933	0.866	0.861	0.975	0.979	0.884	0.931
RF	0.944	0.889	0.875	0.982	0.936	0.954	0.945
GB	0.956	0.911	0.89	0.988	0.979	0.93	0.955
NB	0.944	0.889	0.875	0.982	0.957	0.93	0.944
LDA	0.967	0.933	0.906	0.993	1	0.93	0.965
PLS	0.967	0.933	0.906	0.993	1	0.93	0.965
Ensemble: GLM	0.967	0.933	0.906	0.993	1	0.93	0.965
Stacking: NNET	0.956	0.911	0.89	0.988	0.979	0.93	0.955
Stacking: RF	0.967	0.933	0.906	0.993	1	0.93	0.965
Stacking: GB	0.956	0.911	0.89	0.988	1	0.907	0.954

Table 8. Results provided by ML learning and Ensemble Learning algorithms with 14 PC scores of the modern sample. In bold is shown the best algorithmic performance.

	Accuracy	Kappa	Acc. Lower	Acc. Upper	Sensitivity	Specificity	Bal. Acc.
NNET	0.978	0.955	0.922	0.997	1	0.954	0.977
SVM linear	0.967	0.933	0.906	0.993	1	0.930	0.965
SVM radial	0.967	0.933	0.906	0.993	0.979	0.954	0.966
kNN	0.967	0.933	0.906	0.993	1	0.930	0.965
LG	**0.989**	**0.978**	**0.940**	**0.999**	**1**	**0.977**	**0.988**
DTC5.0	0.922	0.844	0.846	0.968	0.957	0.884	0.921
RF	0.967	0.933	0.906	0.993	0.936	1	0.968
GB	0.967	0.933	0.906	0.993	0.957	0.977	0.967
NB	0.911	0.822	0.832	0.961	0.936	0.884	0.910
LDA	0.967	0.933	0.906	0.993	1	0.930	0.965
PLS	**0.989**	**0.978**	**0.940**	**0.999**	**1**	**0.977**	**0.988**
Ensemble: GLM	0.978	0.955	0.922	0.997	1	0.954	0.977
Stacking: NNET	0.978	0.955	0.922	0.997	1	0.954	0.977
Stacking: RF	0.978	0.955	0.922	0.997	1	0.954	0.977
Stacking: GB	0.978	0.955	0.922	0.997	1	0.954	0.977

Table 9. Results provided by ML learning and Ensemble Learning algorithms with 27 PC scores of the modern sample. In bold is shown the best algorithmic performance.

		82 PC scores	7 PC scores	14 PC scores	27 PC scores	Estrecho Cave
NNET	**Size**	3	5	3	33	5
	Decay	0.0002154435	0.006812921	0.0006812921	0.0004641589	0.001
SVMl	**C**	1	1	1	1	1
SVMr	**Sigma**	0.006572848	0.09959687	0.04646318	0.02026129	0.006722123
	C	4	0.5	256	64	128
kNN	***k***	7	21	7	7	7
RF	**mtry**	23	2	2	4	6
GB	**n.trees**	300	150	900	900	900
	interaction.depth	3	11	2	1	2
	shrinkage	0.1	0.1	0.1	0.1	0.1
	n.minobsinnode	10	10	10	10	10
NB	**fl**	0	0	0	0	0
	usekernel	False	False	False	False	False
	adjust	1	1	1	1	1
PLS	**ncomp**	8	6	7	5	8
Stacking: NNET	**Size**	33	29	23	23	27
	Decay	0.002154435	0.003162278	0.006812921	1×10^{-4}	0.006812921
Stacking: RF	**mtry**	2	10	4	3	4
Stacking: GB	**n.trees**	900	1000	700	700	950
	interaction.depth	7	8	10	5	6
	shrinkage	0.1	0.1	0.1	0.1	0.1
	n.minobsinnode	10	10	10	10	10

Table 10. Hyperparameter configuration of the best performance model provided by the Machine Learning algorithms. Note that the configuration used to classify the sample of the Estrecho Cave has also been included.

	NNET	SVMl	SVMr	kNN	LG	DTC5.0	RF	GB	NB	LDA
SVMl	0.06	–	–	–	–	–	–	–	–	–
SVMr	0.08	0.43	–	–	–	–	–	–	–	–
kNN	0.40	0.13	0.27	–	–	–	–	–	–	–
LG	0.12	**0.69**	0.31	−0.03	–	–	–	–	–	–
DTC5.0	0.22	0.18	0.20	0.29	0.09	–	–	–	–	–
RF	0.13	0.14	0.24	0.37	0.01	0.49	–	–	–	–
GB	0.23	0.16	0.21	**0.53**	0.12	0.47	**0.52**	–	–	–
NB	0.06	0.09	0.29	0.26	−0.01	0.14	0.35	0.13	–	–
LDA	0.20	0.45	0.45	0.27	0.45	0.25	0.16	0.32	0.04	–
PLS	**0.62**	0.09	0.14	0.50	0.01	0.20	0.18	0.21	0.10	0.23

Table 11. Correlation values obtained by calculating the pairwise correlation of the different algorithms applied to the modern sample and the Estrecho Cave. Values in bold show the presence of correlation between algorithms.

	Accuracy	Kappa	Acc. Lower	Acc. Upper	Sensitivity	Specificity	Bal. Acc.	*M. musculus*	*M. spretus*
NNET	0.989	0.978	0.940	1.000	1	0.977	0.988	4	44
SVM linear	0.956	0.911	0.890	0.988	0.957	0.954	0.956	4	44
SVM radial	0.967	0.933	0.906	0.993	0.957	0.977	0.967	3	45
kNN	0.978	0.955	0.922	0.997	1	0.954	0.977	7	41
LG	0.956	0.911	0.890	0.988	0.957	0.954	0.956	3	45
DTC5.0	0.944	0.889	0.875	0.982	0.915	0.977	0.946	11	37
RF	0.944	0.889	0.875	0.982	0.915	0.977	0.946	13	35
GB	0.967	0.933	0.906	0.993	1	0.930	0.965	10	38
NB	0.844	0.689	0.753	0.912	0.830	0.861	0.845	26	22
LDA	0.978	0.955	0.922	0.997	1	0.954	0.977	4	44
PLS	0.967	0.933	0.906	0.993	1	0.930	0.965	3	45
Ensemble: GLM	0.978	0.955	0.922	0.997	1	0.954	0.977	5	43
Stacking: NNET	**1**	**1**	**0.960**	**1**	**1**	**1**	**1**	**5**	**43**
Stacking: RF	0.989	0.978	0.940	1.000	1	0.977	0.988	3	45
Stacking: GB	0.978	0.955	0.922	0.997	1	0.954	0.977	3	45

Table 12. Results provided by ML learning and Ensemble Learning algorithms with all PC scores of the modern sample and the Estrecho Cave. First, the results provided by the training process of the algorithms are shown while in the last two columns the number of identified teeth by species is shown.

	Ensemble: GLM	Stacking: NNET	Stacking: RF	Stacking: GB
M. spretus	41	42	43	45
M. m. domesticus	2	4	2	3
Indeterminate ($p < 0.9$)	5	2	3	0

Table 13. Taxonomic classification of 48 m1 of the Estrecho Cave according to Ensemble Learning algorithms. p: the posterior probability of classification.