Original Article

"You Are Not My Type": An Evaluation of Classification Methods for Automatic Phytolith Identification

José-Francisco Díez-Pastor[1], Pedro Latorre-Carmona[1], Álvar Arnaiz-González[1], Javier Ruiz-Pérez[2] and Débora Zurro[3]* (ID)

[1]Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad de Burgos, Burgos, Spain; [2]CaSEs – Culture and Socio-Ecological Systems Research Group, Department of Humanities, Pompeu Fabra University, Barcelona, Spain and [3]Institución Milá y Fontanals de Investigación en Humanidades – Consejo Superior de Investigaciones Científicas (IMF-CSIC), C. Egipciaques 15, 08001 Barcelona, Spain

Abstract

Phytoliths can be an important source of information related to environmental and climatic change, as well as to ancient plant use by humans, particularly within the disciplines of paleoecology and archaeology. Currently, phytolith identification and categorization is performed manually by researchers, a time-consuming task liable to misclassifications. The automated classification of phytoliths would allow the standardization of identification processes, avoiding possible biases related to the classification capability of researchers. This paper presents a comparative analysis of six classification methods, using digitized microscopic images to examine the efficacy of different quantitative approaches for characterizing phytoliths. A comprehensive experiment performed on images of 429 phytoliths demonstrated that the automatic phytolith classification is a promising area of research that will help researchers to invest time more efficiently and improve their recognition accuracy rate.

Key words: feature extraction, machine learning, microfossils, morphometry, proxy

(Received 18 March 2020; revised 4 August 2020; accepted 7 October 2020)

Introduction

Cell morphologies [with sizes mostly varying between 10 and 200 μm (Dunn, 1983)] vary according to the specific function of the tissue in which they develop. Phytoliths are particles of silica formed in cell walls, cell interiors, or intercellular spaces of living plants that can serve as archives of past vegetation in soils, as a record of human activities in archaeological contexts and materials, and in dental calculus and fecal materials as a result of food habits of past and present populations (Piperno, 2014; Shillito, 2018). Plant taxonomy is often related to specific cell morphologies which might produce diagnostic phytoliths. In many cases, phytolith morphology can provide information about the plant organ as well as the plant taxon in which it was formed.

Because phytoliths are composed largely of amorphous silica (SiO_2) and are more resistant to weathering than most other microfossils, phytolith analysis has been a very active and growing area of research during the last few decades (Hart, 2016; Zurro et al., 2016; Shillito, 2018). Phytoliths can provide hints and useful information about past vegetation and climate as well as past plant consumption, especially in circumstances where other sources of information are unclear or scarce (e.g., Lombardo et al., 2019).

Landscape anthropization and past and present modifications of the biosphere are currently hot research topics (Piperno et al., 2015), and methodologies combining different sources of information (multiproxy approaches) are becoming fundamental tools for disentangling the relationship between the social and the environmental systems (Mayle & Iriarte, 2014; Miehe et al., 2014). Because studies are commonly carried out either at regional or global scales, or with a long historical perspective, larger datasets are usually needed.

The development of an automatic classification framework for phytolith analysis would allow researchers to focus their investigation on other issues, such as improving the efficiency of recovery and calibration techniques, data integration, and interpretation of results. Automated classification would also foster standardized raw data production since data production depends, in several areas of knowledge, on the capability of the analyst to identify structures or patterns (Leighton et al., 2013). The identification of phytoliths and other microscopic and microfossil proxies (such as pollen or diatoms, for instance) is still carried out by researchers manually under the microscope, where observer bias and the relative experience of the analyst can lead to substantial identification errors and difficulty with replication of the results between labs (Peperzak, 2010; Mihlbachler et al., 2012).

An automated classification system would also help determine the minimum number of individuals required for significance for different research questions and in different environmental or archaeological contexts (e.g., pollen sum or phytolith sum), see a review in Strömberg (2009); Pearsall (2015); Zurro (2018).

Automatic classification has become a common tool in many scientific research areas, including remote sensing (Li et al., 2019), autonomous driving (Chen et al., 2018), and medicine (Selvikvåg-Lundervold & Lundervold, 2019). In addition, optical character recognition (OCR) is now a standard tool within the *humanistic* disciplines (Hockey, 1994; Crane et al., 2007; Traub et al., 2015). These methodologies are being increasingly adopted at the macro-scale level for archaeological research, including the analysis of landscapes using satellite images (Davis, 2019), the study of objects, such as ceramic typologies (Hein et al., 2018), or petroglyphs (Seidl et al., 2015). Despite the increase in its use in archaeology, automatic classification systems still remain an under-utilized tool when considering the potential of these methodologies within the discipline. In archaeobotanical studies, where the standard count number under the microscope has been fixed between 250 and 500 individuals per sample (depending on the technique, the research question, etc; Wright, 2010; Pearsall, 2015; Zurro, 2018), several attempts have been made to make this step of the research process much faster and more efficient, especially for microremains such as starches (Wilson et al., 2010; Arráiz et al., 2016) and pollen (France et al., 2000; Li et al., 2004; Treloar et al., 2004; Ticay-Rivas et al., 2011; Boser et al., 2020).

There have been several studies that used quantitative phytolith morphometric size and shape parameters for the identification of morphometric characteristics (Ball et al., 2016; Out & Madella, 2016; Portillo et al., 2019). Recently, researchers have started to design computing methods for the automatic identification of phytoliths (Evett & Cuthrell, 2016; Cai & Ge, 2017; Gallaher et al., 2020).

Evett & Cuthrell (2016) established the conceptual basis for the application of semi-automated classification methods to morphometric phytolith analysis, describing detailed procedures and strategies to be tested while acknowledging current technical limitations. The authors examined functional aspects regarding image acquisition, morphometric parametrisation (e.g., geometric parameters and elliptic Fourier analysis), classification techniques (multivariate statistics versus supervised learning models), and the development of a semi-automated phytolith analysis system.

Cai & Ge (2017) extracted grass short cell phytoliths from the leaves of 23 taxa belonging to the subfamilies Ehrhartoideae, Bambusoideae, and Pooideae. They used morphometric data from scanning electron microscopy (SEM) images to train a classifier to successfully distinguish different genera within the Oryzeae even though they all produce the same phytolith morphotypes.

Gallaher et al. (2020) used three-dimensional (3D) geometric morphometrics and supervised classification algorithms to: (1) analyze the shape and size of modern grass phytoliths from 70 species of the subfamilies Anomochlooideae, Bambusoideae, Oryzoideae, Pharoideae, and Puelioideae; (2) build a classification model based on the short cells extracted from these modern samples; and (3) classify fossil grass phytoliths from Eocene sediments through the application of the previous resulting model. The results showed high classification scores among clades at different taxonomic levels even when different clades shared the same morphotypes.

This paper aims to analyze the applicability of machine learning algorithms for automatic phytolith classification, reducing time spent on phytolith identification under the microscope and human error. The scope of this paper can be summarized as follows:

- The classification of eight different phytolith morphotypes, corresponding to morphotypes commonly found in archaeological assemblages: spheroid, bilobate, cross, saddle, rondel-trapezoid, acute bulbosus, elongate, and bulliform flabellate (as defined in Neumann et al. (2019)). Although subtypes of some of these morphotypes have been widely recognized, only broad categories were taken into account. These morphotypes include (1) a wide variability in shapes and sizes, as well as (2) a degree of overlap (occurring in some cases), so that efficiency of the methods used can be tested accurately.
- Experimental comparison of six classification algorithms, including lazy learning techniques [k-nearest neighbors (k-NN)], to more advanced ones [support vector machines (SVM)].
- Two different feature extraction techniques were used, including geometric morphometric descriptors and elliptic Fourier descriptors (EFDs).

This paper is organized as follows: Section "Materials and Methods" presents the main steps of the computer-assisted morphometric-based phytolith analysis proposed, including how the images were collected and processed to obtain the features (Subsection "Generation of the Samples") and the different types and characteristics of classifiers used in the comparative analysis (Subsection "Computer-Assisted Morphometric-Based Phytolith Analysis"). Section "Results and Discussion" presents and discusses the classification results obtained using the different algorithms. The main conclusions are drawn in Section "Summary and Conclusions, and finally, future work is detailed in Section "Future Research Lines".

Materials and Methods

Phytolith classification was accomplished by applying the flowchart shown in Figure 1. A series of microscopic images was acquired, and the contour for each individual phytolith was drawn using an annotation tool. From the contours, a mask was created by the system in order to separate the phytolith from the background. By using these contours, a series of geometric and mathematical features (defining the shape of each phytolith) was computed and used to train a group of six classifiers, followed by a comparative performance analysis of these classifiers. Each methodological stage will be explained in deeper detail in the following subsections.

Generation of the Samples

Phytoliths were extracted from sediment samples collected in four different locations: archaeological site Isla Manechi, pre-Columbian raised fields and palaeosols from the Llanos de Moxos (Bolivia), and archaeological site Caldeirão from the state of Amazonas in Brazil.

We decided to work with material from soils as this material is part of our daily routine. Paleoecological and archaeological research, as well as studies related to agronomy or pedology (among others), constitute a fundamental part of current phytolith studies so that specialists working with phytoliths from soils constitute the majority of the research that is currently being carried out within this field of study.

Because substantial weathering of soil phytoliths can change their morphological characteristics, when choosing samples for this study, we discarded individuals showing any slight
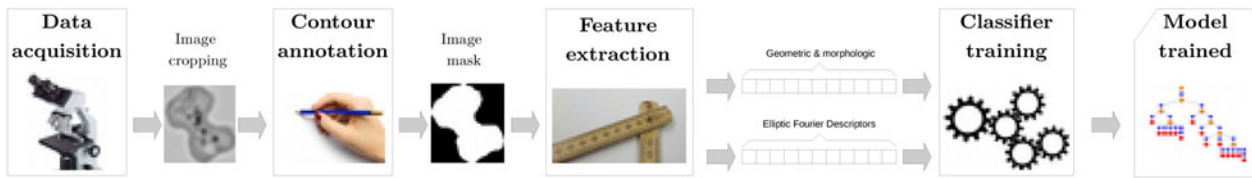
Fig. 1. Visual representation of the computer-assisted morphometric-based phytolith classification system.

mechanical breakage or a high degree of chemical dissolution (thus, altering their shape). Those phytoliths whose surface was affected by partial dissolution but had their shape unaltered were included in the study, without changing sample characteristics.

The extraction followed standard procedures (Madella et al., 1998; Lombardo et al., 2016). Phytoliths were mounted on microscope slides using Entellan® New (Merck), and images were obtained using an Olympus BX51 transmitted light microscope with an Olympus SC50 camera and the Olympus Stream Basic image processing software (version 1.9.4). Images were taken at ×500 on-scope magnification under automatic exposure and exported as *.jpeg* files with a resolution of 2,560 × 1,920 px.

Photomicrographs were taken according to the most recognizable view for each morphotype, that provided a clear morphological outline, discarding any surficial feature such as surcate or verrucate textures (see Madella et al., 2005; Neumann et al., 2019). Most phytoliths were photographed from apex/planar view (following a top-to-base perspective) except for rondel-trapezoid, acute bulbosus and a few bulliform flabellate that were captured from the side view.

The total number of photomicrographs obtained for each morphotype is shown in Table 1. Only nonarticulated (not attached to any other phytoliths) were considered and just one photomicrograph per phytolith was recorded. The total number of samples was 429. All the images are publicly available at https://reposi-tori.upf.edu/handle/10230/44939, all the morphotypes have at least 50 samples, and the dataset is fairly balanced (i.e., there is a similar number of samples per class).

The image of each phytolith was digitized, using an open-source web annotation tool called *VGG Image Annotator*.[1] This tool allows a researcher to create a *control-points* based contour of each object of interest in an image and then obtain the coordinates of each control point that defines the shape of the object. These coordinates were used to create a mask,[2] and these masks were then processed to obtain a series of geometrical features which define the feature vector for each sample object. The phytolith contour selection is a non trivial task, due to the 3D nature of phytoliths. For this reason, the contours are usually fuzzy and can be different depending on the person who is drawing their shape. With the aim of removing the bias associated with this issue, the digitization was made by only one researcher (for removing, at least, inter-person variation).

A mosaic composition of the images acquired of different phytoliths, and their corresponding masks obtained after processing

Table 1. Distribution of Phytolith Images Per Morphotype (Class).

Phytolith Class	Number of Images
Bilobate	55
Bulliform flabellate	60
Cross	63
Elongate	50
Spheroid	51
Rondel-trapezoid	50
Saddle	50
Acute bulbosus	50
Total number of images	429

the *.csv* files from the annotator, are shown in Figure 2. The geometric features used in the study are detailed in Table 2.

Computer-Assisted Morphometric-Based Phytolith Analysis

According to Evett & Cuthrell (2016), a computer-assisted phytolith automatic classification system should be formed by three blocks: data acquisition, classification, and database integration. In this paper, we focus on the classification stage.

The task of predicting the class of an unknown sample is called classification in the machine learning community. The classification task presented here aims to predict the class of a phytolith (i.e., its morphotype) from its photomicrograph, using algorithms or classifiers.

The classification scheme includes cropping of the phytolith image to isolate the phytolith; feature extraction—converting the cropped image into an array of features; and using the extracted features to train the classifier. Classifier training is the process by which the patterns of the different classes (morphotypes) are learnt by the algorithm. Once the classifier has been trained, the model "has learnt" how to distinguish between phytoliths of different classes, and this knowledge can be applied to classify unknown phytoliths.

Feature Extraction

Feature extraction involves transforming the picture into a set/array/vector of numeric features/attributes that summarizes the image characteristics.[3] The main problem with these techniques, is that they commonly need thousands or millions of examples to be trained.

Each phytolith image used for training or classification is converted into a *feature vector*, which is the input data to the

---

[1] VGG Image Annotator can be found at: http://www.robots.ox.ac.uk/~vgg/software/via/.

[2] In image processing, an object mask is an image with pixels of only two values (also called a *binary* image) where the pixels that belong to or define an object are assigned a value (e.g., zero) and the rest of the image pixels have a different value (e.g., one). This image representation is very useful in many image processing methodologies.

[3] Other techniques, such as deep neural networks, have been recently proposed for avoiding feature extraction (Sun et al., 2013).
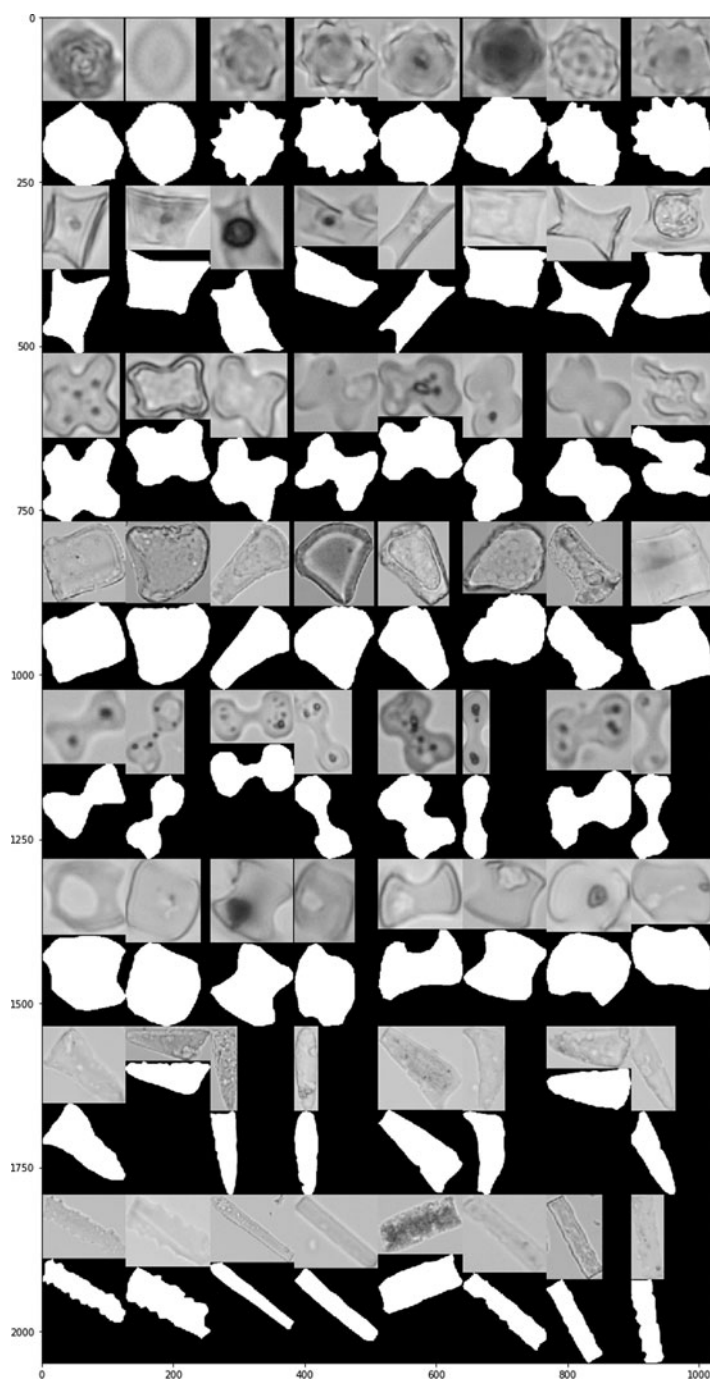
Fig. 2. Some examples of the images used in the study, and the corresponding created masks. Eight images of each of the eight phytolith classes, from top to down: spheroid, rondel-trapezoid, cross, bulliform flabellate, bilobate, saddle, acute bulbosus, and elongate. Each case shows the original image, as well as the binary mask used to obtain the set of features.

classifier. The following three sets of attributes were considered in this paper:

1. *Geometric and morphological attributes.* A great number of geometrical descriptors have been proposed in the literature to characterize the external shape of an object and are the basis of phytolith nomenclature (Madella et al., 2005; Ball et al., 2016; Neumann et al., 2019). These features are explained in Table 2 and were obtained from the created

image masks. Since 16 geometric and morphological descriptors were used, the length of the feature vector was also 16.

2. *Elliptic Fourier descriptors (EFDs)* These descriptors are very commonly used when representing the shape of a contour in a way that is invariant to rotation and size (Kuhl & Giardina, 1982). The PyEFD programming library was used to extract these descriptors.[4] The parameters associated with this extraction process were set as follows: the order of the

---

[4]https://pyefd.readthedocs.io/en/latest/.

Table 2. Geometrical and Morphological Descriptors to Characterize the External Shape of an Object (Ball et al., 2016).

Attribute	Description
Perimeter	Length of the boundary of an object
Convex perimeter	Perimeter of the convex hull that encloses the phytolith
Area	Simple area enclosed by the phytolith boundary
Convex area	Area of the convex hull that encloses the phytolith
Major axis length	Length of the major axis for the ellipse the phytolith is inscribed in
Minor axis length	Length of the minor axis for the ellipse the phytolith is inscribed in
Equivalent diameter	Diameter of a circle with the same area as the phytolith
Form factor	$= 4 \times \text{Area} \times \frac{\pi}{\text{Perimeter}^2}$. It is 1.0 for a perfect circle and diminishes for irregular shapes
Length	Longest cord within the phytolith
Width	The minor dimension of the phytolith. It can be obtained as the diameter of the smallest hole the object can pass through
Convexity	$= \frac{\text{convex} - \text{perimeter}}{\text{perimeter}}$; It is 1.0 for a perfectly convex shape, diminishes if there are irregularities in the boundary
Solidity	$= \frac{\text{Area}}{\text{convex} - \text{area}}$; It is 1.0 for a perfectly convex shape, diminishes if there are surface indentations
Aspect ratio	$= \frac{\text{Length}}{\text{Width}}$
Roundness	$= 4 \times \frac{\text{Area}}{\pi} \times \text{Length}^2$. It is 1.0 for perfect circle and diminishes with elongation of the phytolith
Compactness	$= \frac{\text{Equivalent} - \text{diameter}}{\text{length}}$

Fourier coefficients was fixed at 10, and normalization was set equal to True (recommended settings for shape classification). This method generates 4 coefficients for each order, which therefore forms a feature vector of length 40 ($= 4 \times 10$). However, when normalized, the first three coefficients are constant and can therefore be disregarded as valid descriptors. Consequently, the final length of the feature vector was 37.

3. *Both: the geometric and morphological features together with the EFDs*. The feature vector is created by joining (concatenating) the two previous feature vectors into a new one. Therefore, the length of the feature vector was 53 ($= 16 + 37$).

Principal component analysis (PCA) reduces the dimensionality of the datasets (i.e., the number of attributes/features) and has demonstrated its effectiveness in some studies (Cai & Ge, 2017). For this reason, we generated the PCA decomposition of the EFDs and formed the feature vector retaining 99% of the explained variance. Because the classification results obtained using the reduced feature vector were systematically worse, we did not report them in Section "Results and Discussion".

Classification Methods

The number of classification methods in machine learning is very large, mainly because there is no single classifier that outperforms all others for all problems. This is called the *no free lunch* theorem (Wolpert & Macready, 1997). For that reason, we tested several classification algorithms using the features defined above:

- **The k-nearest neighbors (*k*-NN)** algorithm (Fix & Hodges, 1951; Cover & Hart, 1967). This method assumes the intuitive reasoning that the nearest[5] points (representing the samples to be classified) in a dataset should be more similar than those that are further away, often assessed using a distance measure. In *k*-NN, an object is classified by assigning it to the class which is most common among its nearest neighbors (where *k* is the number of neighbors to account for, and is a positive integer, typically odd and relatively small). If $k = 1$, then the object is simply assigned to the class of its single nearest neighbor. A value of $k = 3$ was used in this study.

- **Support vector machines (SVM)** (Boser et al., 1992), with a Gaussian kernel (radial basis function—RBF). SVM is a classification method which has gained interest recently, due in part to its capability to deal with problems with a small number of samples in high-dimensional feature spaces. It was originally developed for linearly separable problems, aimed at obtaining the hyperplane whose distance to the two groups of data points (called *margin*), representing the two classes, was maximal. SVM was generalized later to deal with nonlinearly separable problems using the so-called (transformational) *Kernel trick* (Boser et al., 1992). A mathematical transformation function is applied to map the nonlinear separable dataset into a higher dimensional space where the samples can be linearly separated using an hyperplane. Under this mathematical framework, two parameters emerge (C, γ). Their value is usually obtained using a nested cross-validation with *k* folds. In our case, a commonly used value in the literature was selected, $k = 5$. The two-dimensional (2D) (C, γ) parameter space was explored using a grid search strategy, with C ranging from 1×10^{-2} to 1×10^{10} and γ ranging from 1×10^{-9} to 1×10^{3}, in both cases with 13 values equally spaced on a logarithmic scale.

- **Multilayer perceptron (MLP)** (Rosenblatt, 1958). A MLP is a type of *feedforward* artificial neural network (ANN), formed by at least three layers of nodes: (a) an input layer, (b) a hidden layer, and (c) an output layer. All nodes (except for the *input* ones) are neurons characterized by a nonlinear activation function. MLPs are trained using back-propagation of errors and are nonlinear versions of the *perceptron* classifier.[6] Several parameters must be tuned in ANNs. In our experiments, the number of iterations was fixed at 1,000. The regularization parameter, α (used to avoid the so-called overfitting problem) as well as the number of hidden neurons, were optimized using a nested five-fold cross-validation strategy (in a similar way as it was applied for SVM). The parameter space was explored using a grid search. Range in α was $[1 \times 10^{-5}, 1 \times 10^{-1}]$, and five values equally spaced on a logarithmic scale were considered. The number of hidden neurons were {50, 100, 200, 500, 1,000}.

- **Decision trees (Tree)** (Breiman, 2017). Decision trees are classification methods that use a tree-like model for making decisions. In the root of the tree, all examples are used to find which feature is the best to split the group of instances into two subsets, which are then assigned to two new nodes (that are called children nodes). This process is repeated in a recursive way until a stopping criterion is reached. The nodes that

---

[5]*k*-NN needs a distance function to perform classification, Euclidean distance (computed using the hyperspace defined by the feature vector of the examples/points) is commonly used because it is easy to understand. Nevertheless, any other distance function could be used.

[6]The first perceptron classifier initially proposed by Rosenblatt (1958) now includes several improvements related to ANNs.

have not got any children are called leafs, and they make the decision of the class that will be assigned to an example. The decision tree can be seen as a sequence of if/else sentences that determines the decision process of the classifier.

- **Random forest (RF)**. Ensemble learning methods do not train one single model (e.g., one tree), but several of them. The idea behind ensemble learning is the way an expert committee works in real-life, i.e., it is usually easier to properly predict something when the prediction is made by more than one expert, and a consensus is obtained from them. RF generates a group of decision trees (called base classifiers in ensemble learning) during the training stage. In the prediction stage, each base classifier predicts a class, and the class selected the most (the mode) is the final prediction of the ensemble. RFs are used to correct the tendency of the decision trees to *overfit*.[7] The main parameter of a RF is its size (i.e., the number of trees that are generated into the ensemble); in our study, 100 decision trees were used, a common value for this ensemble method.

- **Gradient boosting trees (GBT)** (Friedman, 2001). GBT, as well as RFs, is an ensemble technique. Nevertheless, GBT trains the base classifiers in a different way than RF: whereas RF trains each base classifier independently of each other, GBT trains the base classifiers in a sequential order, one by one. Specifically, GBT generates base classifiers iteratively: the first classifier predicts the original class labels of the samples, the second classifier predicts the error made by the first classifier, the third classifier predicts the error made by the classifier formed by the first two base classifiers, and so on. The idea of GBT is to focus the learning process on those examples that are more difficult to predict. Like RF, GBT has an essential parameter: the number of base classifiers, which was set to 100.

Some of the main advantages and disadvantages of the classifiers explained above are summarized in Table 3. The characteristics that all the classifiers had in common were not included in the table for readability purposes.

All the experiments were performed using the Scikit-learn Python library (version 0.23.1) (Pedregosa et al., 2011); the source code can be publicly accessed on Github.[8] Unless stated otherwise, the program's default classifier parameters were used. For each feature, data values were transformed and standardized to mean 0 and standard deviation 1. The classification accuracy was assessed by applying a 10-fold cross-validation strategy, using the three types of feature vectors mentioned above.

Results and Discussion

The accuracy for the six classification algorithms trained with the three sets of features is shown in Table 4. The best result for each column (i.e., for each set of features) is highlighted in bold. The best result overall the table is highlighted in italics and bold. The 95% confidence intervals were obtained using the 10-fold cross-validation strategy.

Classifiers trained with EFDs alone performed worse than those trained with geometric and morphological descriptors. The combined EFDs together with geometric and morphological descriptors outperformed the solo descriptors for all the classifiers but the SVM classifier. The best result was achieved by SVM

(exclusively trained with the geometric and morphological attributes) and by RF (trained with both: EFDs and geometric and morphological). Although this may seem counterintuitive, basic geometric attributes are in fact the basis for phytolith identification, classification, and nomenclature (Madella et al., 2005; Neumann et al., 2019).

The Kruskal–Wallis H test (Kruskal & Wallis, 1952) and the Wilcoxon signed-rank test (Wilcoxon, 1945) were used in order to assess whether the differences between the classifiers were significant or not. This was done in two different ways: one versus one (using Wilcoxon signed-rank test) and all versus all (using Kruskal–Wallis H test).

First of all, Kruskal–Wallis was used to determine whether there was significant differences across all methods overall (multiple comparison), showing that the differences between the methods were significant at a 95% confidence level. Then, Kruskal–Wallis was performed by columns (for each one of the feature sets) giving the same conclusion, i.e., the differences between the accuracies on a column were significantly different at 95% of confidence.

In the same way, the Wilcoxon signed-rank test was first used to compare the best result overall against all the other classifiers. Finally, the Wilcoxon signed-rank test was applied to compare the best classifier of each column in Table 4 (i.e., each set of features), against all the other classifiers in the same column.

For the first column (morphological and geometric features), the best classifier was SVM and it was significantly better than any other classifier trained with this set of features. For the second column (EFDs), RF was the best classifier and it was statistically indistinguishable at 95% of confidence from GBT (this was highlighted with a △ symbol close to these results). Finally, in the last column (all features), RF was the best classifier and it was statistically indistinguishable from SVM (this was highlighted with a • symbol close to these results). Moreover, the best overall result (SVM trained with morphological and geometric features) was compared against all the other results using Wilcoxon, showing that it was statistically indistinguishable from SVM, GBT, and RF trained with all the features (this was highlighted by enclosing the results into a box).

The accuracy per morphotype/class for the six classifiers trained with the geometric and morphological descriptors are gathered in Table 5. Note that the *Macro avg.* value does not match the accuracy of the classifier of Table 4 because the former is the mean of accuracies per class, not the global accuracy. Table 5 shows that certain phytolith morphotypes, including bulliform flabellate, elongate, and spheroid are more easily identified than others for most of the classifiers, probably due to their distinctive shape. On the other hand, some morphotypes, such as saddle, are more problematic to identify for all classifiers.

The confusion matrix, which presents the hits (images properly identified) and misses (images wrongly predicted), obtained for SVM trained with the geometrical and morphological features, is shown in Figure 3a. The classification rate is high for most of the morphotypes, being the cross morphotype the most problematic with 49 hits and 14 misses. Moreover, the same information as that shown in the confusion matrix, is presented in percentage terms in Figure 3b.

Images of four examples of incorrect classification results are illustrated in Figure 4.

Cai & Ge (2017), instead of working with general types, focused their research specifically on the classification of short cell phytoliths, aiming at taxonomical identification and obtaining

---

[7]The overfit is the ability/pathology of some models to adapt too close to the original data losing their ability to generalize or to predict future observations reliably.
[8]Repository at: https://github.com/alvarag/AutomaticPhytolithClassification.

Table 3. Main Properties of the Different Classifiers Used in the Study.

Characteristic	*k*-NN	SVM	MLP	Tree	RF	GBT
Training time	Fast	Slow	Slow	Fast	Medium	Medium
Testing time	Slow	Fast	Fast	Fast	Fast	Fast
Tendency to overfit	Medium	High	High	High	Medium	Medium
Number of parameters	Low	High	High	Low	Low	Low
Ease of interpretation	Yes	No	No	Yes	No	No
Deal with many features	Bad	Good	Good	Good	Good	Good

Table 4. Accuracy Scores and 95% Confidence Intervals for the Six Classifiers Trained with the Three Different Sets of Features: Morphological and Geometric, EFDs, and Both (All). The best classifier for each column is highlighted in boldface, and the best classifier overall is highlighted in italics. All the results that are statistically indistinguishable from the best overall are shown in a box. By columns, the symbols (Δ and •) represent those classifiers statistically indistinguishable from the best classifier of its column: using EFDs (Δ) and using all features (•).

Classifier	Features Used for Training		
	Morph. and Geom.	EFDs	All
SVM	*0.8741*±0.032	0.7087±0.041	0.8460± 0.030 •
RF	0.7880±0.029	**0.7760**±0.050Δ	**0.8741**±0.034 •
GBT	0.7716±0.043	0.7621±0.055Δ	0.8344±0.046
k-NN	0.7064±0.029	0.6037±0.053	0.7460±0.040
Tree	0.6993±0.046	0.6060±0.018	0.7366±0.031
MLP	0.6994±0.049	0.6317±0.033	0.7552±0.056

results comparable to ours, with SVM performing the best of the models they tested.

In a similar way, Gallaher et al. (2020) carried out an automatic classification process of grass short cells. However, they considered a 3D morphometric approach, instead of a 2D strategy. They showed that linear discriminant analysis (LDA) gave the best classification results, while SVM and MLP algorithms exhibited the worst.

Within scientific praxis, classification constitutes a problem limitation *per se* and categories are established according to research questions, meaning they can focus on specific aspects of research material, thus obviating many others. Morphotypes are somehow an idealized and simplified result (an abstract version) of what constitutes reality on the basis of chosen criteria that allow distinguishing a type from the rest of types under consideration (Rovner & Russ, 1992). In addition, existing phytoliths classifications have been developed in different areas of the globe, so that classification systems are neither standardized, nor necessarily compatible or covering completely the phytolith variability.

We have used general categories that all specialists recognize, but that can be further subdivided. The case of grass short cells is paradigmatic, and several researches carry out subclassification processes when those morphotypes can provide further taxonomic information (Barboni & Bremond, 2009; Gallaher et al., 2020). The existing internal variability within our samples is probably producing a loss of accuracy on the results. Incorrect

classifications may have occurred because of the similarity in some of the morphologies. For example, some acute bulbosus have a quadrangular form similar to elongates (see Fig. 2). Another important reason for error might be that the classifiers only consider phytoliths as 2D bodies and the identification of phytoliths might depend on their spatial orientation when mounted on the slides since even slight differences in the position of a morphotype could affect how the descriptors describe its shape up to a point that it might lead to a misclassification result.

Approaches to archaeobotanical remains, such as the one presented here, are still scarce. Regarding Evett & Cuthrell (2016), Cai & Ge (2017), or Gallaher et al. (2020) although they all develop similar methodologies to identify phytoliths, the objectives, as well as the selected criteria, are not comparable to the present study.

Summary and Conclusions

Phytolith identification and classification is a time-consuming task subject to human errors. Automatic classification techniques have been recently proposed to help solve these problems (Evett & Cuthrell, 2016; Cai & Ge, 2017).

Our study presents a computer-assisted morphometric-based model for automatic recognition of phytoliths using photomicrographs of phytoliths preserved in archaeological samples. Only non-weathered phytoliths were selected to carry out this pilot study.

Morphological features were extracted from a dataset of 429 phytolith images composed of eight phytolith morphotypes. Three feature extraction methods were applied, and six classification techniques were tested.

The most accurate results were obtained with SVM and RFs. This result was not unexpected because both SVM and RF have shown high classification accuracy results in other applications. Interestingly, SVM behaved quite differently compared with most of the other classifiers. While SVM performed better when trained with only morphological and geometric features, the other classifiers performed better when using both characteristics (EFDs combined with morphological and geometric features), probably because irrelevant attributes can seriously affect SVM performance (Weston et al., 2001). Some elliptical Fourier descriptors can be substantially discriminatory, while others might not, which could improve the performance of classifiers such as RFs or GBTs (classifiers that deal better with irrelevant attributes), while damaging the SVM classifier.

We found that the accuracy results obtained by SVM, RF, and GBT, trained with all features (EFDs and morphological and geometric features) were statistically indistinguishable at 95% of

Table 5. Accuracy Per Morphotype for Each One of the Classifiers Used, Trained with the Geometric and Morphological Features.

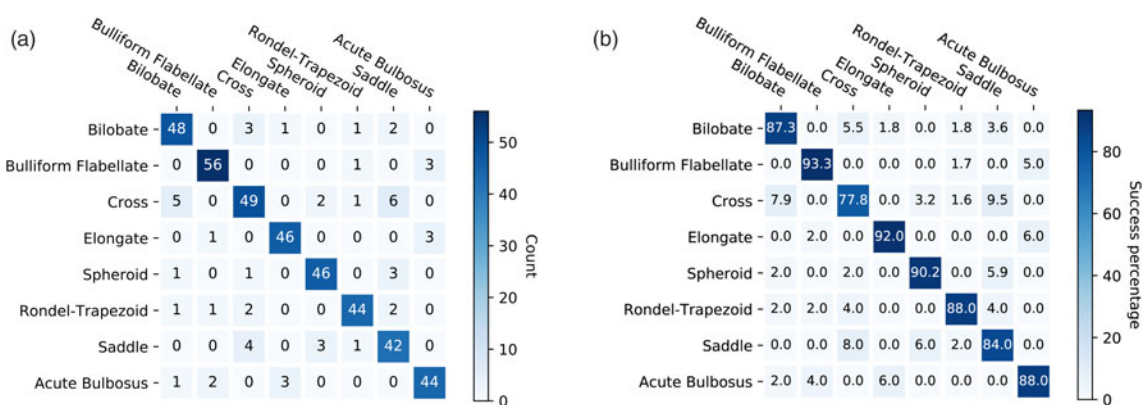| Phytolith Class	*k*NN	SVM	Tree	RF	GBT	MLP
Bilobate	0.5091	**0.8727**	0.5818	0.7455	0.7636	0.5273
Bulliform flabellate	0.9333	0.9333	0.9333	0.9167	0.9000	**0.9500**
Cross	0.5873	**0.7778**	0.5714	0.6984	0.6984	0.7460
Elongate	0.7800	**0.9200**	0.6800	0.8400	0.8000	0.7800
Spheroid	0.7451	**0.9020**	0.7255	**0.9020**	0.8431	0.7255
Rondel-trapezoid	0.7200	**0.8800**	0.7200	0.7800	0.7400	0.6400
Saddle	0.6600	**0.8400**	0.6200	0.6000	0.6400	0.5400
Acute bulbosus	0.7200	**0.8800**	0.7600	0.8200	0.7800	0.6400
Macro avg.	0.7069	0.8757	0.6990	0.7878	0.7706	0.6936
Std. dev.	0.1276	0.0491	0.1173	0.1059	0.0812	0.1383

Fig. 3. (**a**) Confusion matrix of the SVM classifier trained with morphological and geometric features. (**b**) Matrix obtained by normalizing the results in the confusion matrix, to present them on a success percentage basis. Darker colors represent higher values. The darker the diagonal, the better the classification result.
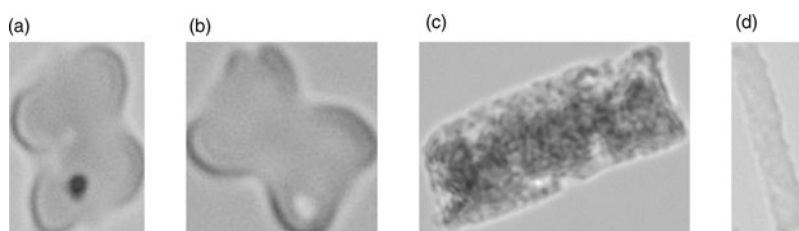

Fig. 4. Four examples of incorrect classifications results: (**a,b**) cross phytoliths, classified as saddle, and (**c,d**) elongate phytoliths, classified as acute bulbosus.

confidence (using Wilcoxon) from the SVM trained only with morphological and geometric features.

While deep learning needs thousands of images per class, machine learning methodologies basically require a balanced dataset. Despite the fact that the sample size was relatively small in this study, this was not an unachievable hindrance, in the sense that machine learning methodologies are much more adaptable tools than nonspecialists could expect.

Studies such the one presented here are crucial to advance in research using proxies like phytoliths or other botanical microfossils. At the moment, we are experiencing a trial and error period regarding morphometrics and automatic identification in phytolith studies. We need to increase our common experience on using such methods so that criteria can be hierarchized according to their capability to train the algorithms and produce the best results.

Automatization processes and image processing techniques will substantially reduce time-consuming tasks such as standard microscopy analysis. At the same time, they will limit the subjective bias stemming from different researchers due to differences in background, training level, and regional experience (i.e., phytolith assemblages in the American tropics differ from the Near Eastern Neolithic). Standardization will produce data that is comparable worldwide.

The study presented in this paper constitutes the first step in developing a tool that will assist in the identification and

classification process. Even though several researchers have attempted to create automatic tools for the identification of archaeobotanical remains, none of the attempts has produced a tool that is accessible online or as a downloadable app. We are sharing the code used in our research (which is accessible at https://github.com/alvarag/AutomaticPhytolithClassification) to stimulate other researchers to join in the effort to build a real and functional tool that can be trained online, increasing its accuracy.

Future Research Lines

The development of new features and the application of feature selection techniques are some of the research avenues we are planning to explore. It would be important to determine which attributes are the most representative ones. In particular, the use of irrelevant or redundant attributes usually induces a lower classifier performance and higher execution times. Therefore, the application of different types of feature selection strategies to this problem would probably result in the improvement of the general performance of most of classifiers (Guyon & Elisseeff, 2003).

Other potentially beneficial research lines would include (a) analysis of the impact of the sample size on the phytolith classification success rate, (b) application to new morphotypes and descriptors and sub-classification of morphotypes according to more detailed features, (c) isolation of taxonomically diagnostic morphotypes, and (d) automatic isolation and digitization of the outline of a phytolith from a photomicrograph in order to achieve a completely automatic phytolith identification system.

References

Arráiz H, Barbarin N, Pasturel M, Beaufort L, Domínguez-Rodrigo M & Barboni D (2016). Starch granules identification and automatic classification based on an extended set of morphometric and optical measurements. *J Archaeol Sci Rep* **7**, 169–179.

Ball TB, Davis AL, Evett RR, Ladwig JL, Tromp M, Out WA & Portillo M (2016). Morphometric analysis of phytoliths: Recommendations towards standardization from the international committee for phytolith morphometrics. *J Archaeol Sci* **68**, 106–111.

Barboni D & Bremond L (2009). Phytoliths of east African grasses: An assessment of their environmental and taxonomic significance based on floristic data. *Rev Palaeobot Palynol* **158**, 29–41.

Boser B, Marchant R, de Garidel-Thoron T, Tetard M, Barboni D, Gally Y & Beaufort L (2020). Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains. *Comput Geosci* **140**, 104498.

Boser BE, Guyon IM & Vapnik VN (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152. New York, NY: Association for Computing Machinery.

Breiman L (2017). *Classification and Regression Trees*. Routledge. https://books.google.es/books/about/Classification_and_Regression_Trees.html?id=gLs6DwAAQBAJ&redir_esc=y.

Cai Z & Ge S (2017). Machine learning algorithms improve the power of phytolith analysis: A case study of the tribe Oryzeae (Poaceae). *J Syst Evol* **55**, 377–384.

Chen X, Kundu K, Zhu Y, Ma H, Fidler S & Urtasun R (2018). 3D object proposals using stereo imagery for accurate object class detection. *IEEE Trans Pattern Anal Mach Intell* **40**, 1259–1272.

Cover T & Hart P (1967). Nearest neighbor pattern classification. *IEEE Trans Inf Theory* **13**, 21–27.

Crane G, Babeu A & Bamman D (2007). eScience and the humanities. *Int J Digit Libr* **7**, 117–122.

Davis DS (2019). Object-based image analysis: A review of developments and future directions of automated feature detection in landscape archaeology. *Archaeol Prospect* **26**, 155–163.

Dunn ME (1983). Phytolith analysis in archaeology. *Midcont J Archaeol* **8**, 287–297.

Evett RR & Cuthrell RQ (2016). A conceptual framework for a computer-assisted, morphometric-based phytolith analysis and classification system. *J Archaeol Sci* **68**, 70–78.

Fix E & Hodges Jr. JL (1951) Discriminatory analysis-nonparametric discrimination: Consistency properties. Tech. rep., California Univ Berkeley.

France I, Duller AWG, Duller GAT & Lamb HF (2000). A new approach to automated pollen analysis. *Quat Sci Rev* **19**, 537–546.

Friedman JH (2001). Greedy function approximation: A gradient boosting machine. *Ann Stat* **29**, 1189–1232.

Gallaher TJ, Akbar SZ, Klahs P, Marvet CR, Senske AM, Clark LG, Strömberg C (2020). 3D shape analysis of grass silica short cell phytoliths (GSSCP): A new method for fossil classification and analysis of shape evolution. *New Phytol* **228**, 376–392.

Guyon I & Elisseeff A (2003). An introduction to variable and feature selection. *J Mach Learn Res* **3**, 1157–1182.

Hart TC (2016). Issues and directions in phytolith analysis. *J Archaeol Sci* **68**, 24–31.

Hein I, Rojas-Domínguez A, Ornelas M, D'Ercole G & Peloschek L (2018). Automated classification of archaeological ceramic materials by means of texture measures. *J Archaeol Sci Rep* **21**, 921–928.

Hockey S (1994) Electronic texts in the humanities: A coming of age. In *Literary Texts in an Electronic Age: Scholarly Implications and Library Services* [1994 Clinic on Library Applications of Data Processing].

Kruskal WH & Wallis WA (1952). Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* **47**, 583–621.

Kuhl FP & Giardina CR (1982). Elliptic fourier features of a closed contour. *Comput Graphics Image Process* **18**, 236–258.

Leighton I, Hiemstra JF & Weidemann CT (2013). Recognition of micro-scale deformation structures in glacial sediments-pattern perception, observer bias and the influence of experience. *Boreas* **42**, 463–469.

Li P, Treloar WJ, Flenley JR & Empson L (2004). Towards automation of palynology 2: The use of texture measures and neural network analysis for automated identification of optical images of pollen grains. *J Quat Sci* **19**, 755–762.

Li S, Song W, Fang L, Chen Y, Ghamisi P & Benediktsson JA (2019). Deep learning for hyperspectral image classification: An overview. *IEEE Trans Geosci Remote Sens* **57**, 6690–6709.

Lombardo U, Ruiz-Pérez J & Madella M (2016). Sonication improves the efficiency, efficacy and safety of phytolith extraction. *Rev Palaeobot Palynol* **235**, 1–5.

Lombardo U, Ruiz-Pérez J, Rodrigues L, Mestrot A, Mayle F, Madella M, Szidat S & Veit H (2019). Holocene land cover change in south-western Amazonia inferred from paleoflood archives. *Glob Planet Change* **174**, 105–114.

Madella M, Alexandre A & Ball T (2005). International Code for Phytolith Nomenclature 1.0. *Ann Bot* **96**, 253–260.

Madella M, Powers-Jones AH & Jones MK (1998). A simple method of extraction of opal phytoliths from sediments using a non-toxic heavy liquid. *J Archaeol Sci* **25**, 801–803.

Mayle FE & Iriarte J (2014). Integrated palaeoecology and archaeology: A powerful approach for understanding pre-Columbian Amazonia. *J Archaeol Sci* **51**, 54–64.

Miehe G, Miehe S, Bohner J, Kaiser K, Hensen I, Madsen D, Liu JQ & Opgenoorth L (2014). How old is the human footprint in the world's largest alpine ecosystem? A review of multiproxy records from the tibetan plateau from the ecologists' viewpoint. *Quat Sci Rev* **86**, 190–209.

Mihlbachler MC, Beatty BL, Caldera-Siu A, Chan D & Lee R (2012). Error rates and observer bias in dental microwear analysis using light microscopy. *Palaeontol Electron* **15**, 1–22.

Neumann K, Stromberg CAE, Ball T, Albert RM, Vrydaghs L & Cummings LS (2019). International Code for Phytolith Nomenclature (ICPN) 2.0. *Ann Bot* **124**, 189–199.

Out WA & Madella M (2016). Morphometric distinction between bilobate phytoliths from *Panicum miliaceum* and *Setaria italica* leaves. *Archaeol Anthropol Sci* **8**, 505–521.

Pearsall DM (2015). *Paleoethnobotany: A Handbook of Procedures*. Walnut Creek, California: Left Coast Press.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M & Duchesnay E (2011). Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12**, 2825–2830.

Peperzak L (2010). An objective procedure to remove observer-bias from phytoplankton time-series. *J Sea Res* **63**, 152–156.

Piperno DR (2014). *Phytolyth Analysis: An Archaeological and Geological Perspective*. Saint Louis: Elsevier.

Piperno DR, McMichael C & Bush MB (2015). Amazonia and the Anthropocene: What was the spatial extent and intensity of human landscape modification in the Amazon Basin at the end of prehistory? *Holocene* **25**, 1588–1597.

Portillo M, Ball TB, Wallace M, Murphy C, Pérez-Díaz S, Ruiz-Alonso M, Aceituno FJ & López-Sáez JA (2019) Advances in morphometrics in archaeobotany. *Environmental Archaeology*.

Rosenblatt F (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* **65**, 386.

Rovner I & Russ J (1992). *Darwin and Design in Phytolith Systematics: Morphometric Methods for Mitigating Redundancy*. Boston: Springer.

Seidl M, Wieser E & Alexander C (2015). Automated classification of petroglyphs. *Digit Appl Archaeol Cult Heritage* **2**, 196–212.

Selvikvåg-Lundervold A & Lundervold A (2019). An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys* **29**, 102–127.

Shillito LM (2018). *Phytolith Analysis*, pp. 1–3.The Encyclopedia of Archaeological Sciences.

Strömberg CAE (2009). Methodological concerns for analysis of phytolith assemblages: Does count size matter? *Quat Int* **193**, 124–140.

Sun Y, Wang X & Tang X (2013) Hybrid deep learning for face verification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ticay-Rivas JR, del Pozo-Baños M, Travieso CM, Arroyo-Hernández J, Pérez ST, Alonso JB & Mora-Mora F (2011. Pollen classification based on geometrical, descriptors and colour features using decorrelation stretching method. In *Artificial Intelligence Applications and Innovations*, Iliadis L, Maglogiannis I & Papadopoulos H (Eds.), EANN 2011, AIAI 2011. IFIP Advances in Information and Communication Technology, vol. **364**. pp. 342–349. Berlin, Heidelberg: Springer.

Traub MC, Ossenbruggen JV & Hardman L (2015) Impact analysis of OCR quality on research tasks in digital archives. *International Conference on Theory and Practice of Digital Libraries*, pp. 252–263.

Treloar WJ, Taylor GE & Flenley JR (2004). Towards automation of palynology 1: Analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning electron microscope images. *J Quat Sci* **19**, 745–754.

Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T & Vapnik V (2001). Feature selection for SVMs. In *Advances in Neural Information Processing Systems*, Leen TK, Dietterich TG & Tresp V (Eds.), pp. 668–674.MIT Press.

Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometr Bull* **1**, 80–83.

Wilson J, Hardy K, Allen R, Copeland L, Wrangham R & Collins M (2010). Automated classification of starch granules using supervised pattern recognition of morphological properties. *J Archaeol Sci* **37**, 594–604.

Wolpert DH & Macready WG (1997). No free lunch theorems for optimization. *IEEE Trans Evolut Comput* **1**, 67–82.

Wright PJ (2010. Methodological Issues in Paleoethnobotany: A Consideration of Issues, Methods, and Cases. In *Integrating Zooarchaeology and Paleoethnobotany*, VanDerwarker A & Peres T (Eds.), pp. 37–64. New York, NY: Springer.

Zurro D (2018). One, two, three phytoliths: Assessing the minimum phytolith sum for archaeological studies. *Archaeol Anthropol Sci* **10**, 1673–1691.

Zurro D, García-Granero JJ, Lancelotti C & Madella M (2016). Directions in current and future phytolith research. *J Archaeol Sci* **68**, 112–117.