Journal of Archaeological Science: Reports 73 (2026) 105873

Contents lists available at ScienceDirect

Journal of Archaeological Science: Reports

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

Experts against automation? Comparing artificial intelligence and human identifications of multi-cell cereal husk phytoliths

Monica N. Ramsey [a,*], Melanie Pugliese [a], Lachlan Kyle-Robinson [a], Iban Berganzo-Besga [b], Rebecca Roberts [c], Jennifer Bates [d], Francesca D'Agostini [e], Zachary C. Dunseth [f], Thomas C. Hart [g], Emma Jenkins [h], Carolina Jiménez-Arteaga [i], Celine Kerfant [j], Marco Madella [j,k], Sigrid Osborne [h], Robert Power [l], Abel Ruiz-Giralt [j], Philippa Ryan [e]

[a] Ramsey Laboratory for Environmental Archaeology (RLEA), Department of Anthropology, University of Toronto Mississauga, Canada
[b] Barcelona Supercomputing Center (BSC), Spain
[c] McDonald Institute for Archaeological Research, University of Cambridge, UK
[d] Department of Archaeology and Art History, Seoul National University, South Korea
[e] Kew Royal Botanic Gardens, UK
[f] Department of Anthropology, University California San Diego, USA
[g] Department of Anthropology, New Mexico State University, USA
[h] Institute for Modelling Socio-Environmental Transitions, Bournemouth University, UK
[i] Natural Sciences Department, German Archaeological Institute, Germany
[j] CASEs, Department of Humanities, Universitat Pompeu Fabra, Spain
[k] ICREA, Barcelona, Spain
[l] School of Archaeology, University College Dublin, Ireland

ARTICLE INFO

Keywords:
Phytoliths
Multi-cell grass husks
Deep learning algorithm
Manual identification
Environmental archaeology
Paleoethnobotany

ABSTRACT

This study compares the efficacy of a deep learning-based computer vision model against manual expert identification of multi-cell husk phytoliths from wheat (*Triticum boeticum/dicoccoides*), barley (*Hordeum spontaneum*), and oats (*Avena sativa*). Conducted via an online survey in 2024 with twelve respondents (92% completion rate, one respondent did not finish), the survey presented 18 phytolith images selected based on: 1) deep learning performance, 2) manual diagnostic clarity, and 3) random control. The deep learning model achieved 100% classification accuracy, while manual experts averaged 44%, with *Avena* proving the most challenging (26.39% accuracy) compared to *Hordeum* (54.17%) and *Triticum* (48.61%). Statistical analysis confirmed significant accuracy variations ($p = 0.0016$), highlighting *Avena's* genus identification difficulty. Experience level influenced performance, with less than five years post-PhD researchers scoring highest (72%), though completion time did not correlate with accuracy. The algorithm's superior performance, employing and appropriately weighting features like wave pattern and papillae, underscores the potential of artificial intelligence to transform paleoethnobotany by generating reliable, large-scale datasets. This research advocates for the integration of machine learning tools within manual studies in the short-term, to enhance accuracy and efficiency in phytolith analysis, paving the way for broader applications in environmental archaeology in the long-term.

1. Introduction

Artificial intelligence (AI), *the broad field of creating systems or machines that can perform tasks that typically require human intelligence*, is rapidly transforming the landscape of archaeology, with applications ranging from pot sherd analysis (Orengo and Garcia-Molsosa, 2020; Orengo et al., 2021) to phytolith identification (Andriopoulou, et al. 2023; Berganzo-Besga, et al. 2025; Berganzo-Besga, et al. 2022b; Cai and Ge 2017; Díez-Pastor, et al. 2024; Díez-Pastor, et al. 2020; Power, et al. 2015). Machine learning (ML), *a subset of AI where systems learn from data to make predictions or decisions without being explicitly programmed*, and deep-learning (DL), *a specialized subset of ML that uses neural networks with many layers to analyse complex patterns, like image recognition in large datasets*, are becoming an increasingly common component of

* Corresponding author at: University of Toronto Mississauga, 3359 Mississauga Road, Mississauga, ON L5L 1C6, USA.
E-mail address: monica.ramsey@utoronto.ca (M.N. Ramsey).

https://doi.org/10.1016/j.jasrep.2026.105873
Received 26 August 2025; Received in revised form 2 June 2026; Accepted 2 June 2026
Available online 10 June 2026

archaeological methodology. AI-tools are particularly effective for mitigating the time intensive vision-based analyses and categorization tasks that typify most archaeological analyses. Accordingly, computer vision, which uses ML and DL to analyze visual data, is well suited to archaeological research. The rapidly expanding application of DL-based computer vision in archaeology generally falls under artifact/object classification algorithms, such as ceramic typology or artifact type classifications (e.g., Mnasri and D'Andrea 2025; Parisotto, et al. 2022; Pawlowicz and Downum 2021; Yang, et al. 2025) and site/object detection algorithms, such as mound detection, pot sherd detection, or remote sensing artifact detection (e.g., Berganzo-Besga, et al. 2021; Character, et al. 2025; Orengo, et al. 2020; Orengo and Garcia-Molsosa 2020; Orengo, et al. 2021; Turan, et al. 2023). This article contributes to this fast growing body of research by comparing our previously published DL-based computer vision model for phytolith classification (Berganzo-Besga et al. 2022; Berganzo-Besga et al. 2025), with an expert-based identification survey of multi-cell husk phytoliths.

Phytoliths, robust inorganic silica 'casts' of plant-cells, are a critical component to both paleoethnobotanical and paleoenvironmental investigations. Phytolith analysts identify both single-cell phytoliths and silica skeletons (multi-cell phytoliths). The algorithm compared in this paper classifies multi-cell phytoliths. Single-cell and mutli-cell phytoliths present different "problems" in respect to identification. Single-cell identifications are difficult to automate due to the variability in which they can be found on the slide and the intrinsic variability of the morphotypes (3D shape). Whilst silica skeletons tend to present themselves in a "flat" position, easier to image on a transmitted light microscope, they also require more computationally complex DL models to effectively evaluate the range of features that must be considered in rank order to make classifications. Multi-cell cereal husk identification to *genus* is challenging, requiring expertise and experience, but it is an established practice (Rosen 1992) (Fig. 1). These challenges mean that phytolith analysis is very time-consuming and remains subject to inter- and intra-observer bias (i.e. inconsistent identifications or misidentification) (see for example, Out, et al. 2024). This, at its crux, is the reason why our original algorithm targeted multi-cell cereal husks, while they potentially provide high levels of taxonomic specificity (*genus*-level) for plant taxa that have great archaeological interest (e.g., prehistoric cereal foodways and domestication), their morphological complexity continues to make manual identification potentially problematic. At the same time, their morphological complexity provided an exciting challenge for testing the application of DL models.

This project, based in the Ramsey Laboratory for Environmental Archaeology (RLEA) at the University of Toronto Mississauga (UTM), aims to overcome these limitations by using AI to generate "big" paleoethnobotany data, to reconstruct past human environments and transform our understanding of pivotal prehistoric events, such as the origins of agriculture. Phytoliths from the Southern Levant (Israel and Jordan) are only the first step of this research program, as demonstrated in Berganzo-Besga, et al. (2022), where we first presented our algorithm, which is capable of classifying between *Triticum* (wheat), *Hordeum* (barley) and *Avena* (oat) multi-cell husk phytoliths with an accuracy of 93.68%. The broader aim is to develop DL models capable of analyzing phytoliths from diverse regions, and differentiate reliably between phytolith taxa with greater precision, including down to *species*, or identify dynamics like growing environment (e.g., irrigation) or landscape burning. Although the integration of AI into paleoethnobotany and archaeology is still in its early stages, it holds transformative potential for the discipline (Andriopoulou, et al. 2023; Berganzo-Besga, et al. 2021; Bickler 2021; Díez-Pastor, et al. 2024; Díez-Pastor, et al. 2020; Garcia-Molsosa, et al. 2020; Orengo and Garcia-Molsosa 2020; Orengo, et al. 2021). Yet, the potential of AI in archaeology will only be realized if researchers trust the results, accordingly in addition to effective algorithms (e.g., the original algorithm (Berganzo-Besga, et al. 2022a)), we also opened the "black box" of our original algorithm, using visual explainers, like heat maps, to show how the algorithm makes

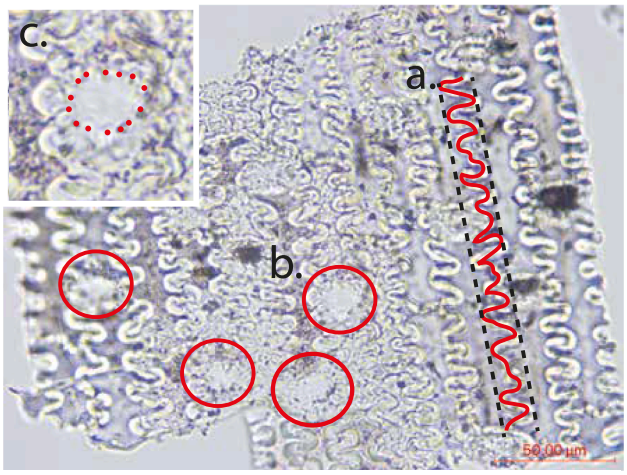

Fig. 1. Summary of main phytolith characteristics used to make manual identifications. (Following Rosen 1992 Table 7.1, page 143). It is important to note that Rosen did not record data for H. spontaneum, the species used in this study. In addition, although Rosen recorded other information including long cell width at different points in the husk (lower and middle husk), these metrics were excluded as this survey demonstrates it tends not to be a regularly employed characteristic.

classifications (Berganzo-Besga, et al. 2025). Finally, it is essential to demonstrate that the proposed algorithm achieves significant improvements over traditional methods to justify the continued development of AI-based approaches.

In this paper however, we do not claim that AI-tools are ready to overtake manual experts in the identification of archaeological phytoliths, indeed our original algorithm is not "archaeology ready," it has not been trained on appropriate "dirty" (i.e. imperfect) archaeology data. Our original algorithm is trained entirely on "clean" modern comparative plant materials. Accordingly, here, we compare the results of our original algorithm to an online survey, composed of images taken from the original testing data, completed by phytolith experts. Our aim is simple; to evaluate how effective our original DL-based model is at categorizing visually complex, "clean" modern phytoliths, compared to human experts.

2. Methods

2.1. Training data and original model

The survey images used in this study were selected from the original multi-cell phytolith dataset, which includes 378 microscope images (original dataset via RLEA Zenodo: https://zenodo.org/records/15692103 (Berganzo-Besga, et al. 2022): 121 images of *Avena sativa* L. (32.01% of the total), 90 of *Hordeum spontaneum* C. Koch (23.81% of the total), 142 of *Triticum boeticum* Boiss. (37.57% of the total) and 25 of *Triticum dicoccoides* (Körn. Ex Asch. & Graebn) (6.61% of the total). The comparative material was sourced from the RLEA plant collections,

including the Mount Scopus Collection, collected over four seasons by trained botanists in Israel, and the Hillman Collection, sampled from UCL's comparative material originally collected by Prof. Gordon Hillman in Turkey.

The comparative material was prepared following established protocols: plant material was first cleaned using a lab wash bottle with distilled water to remove dust. Once dry, the samples were sorted by anatomical part (leaf, culm, husk, awn). To isolate phytoliths, organic

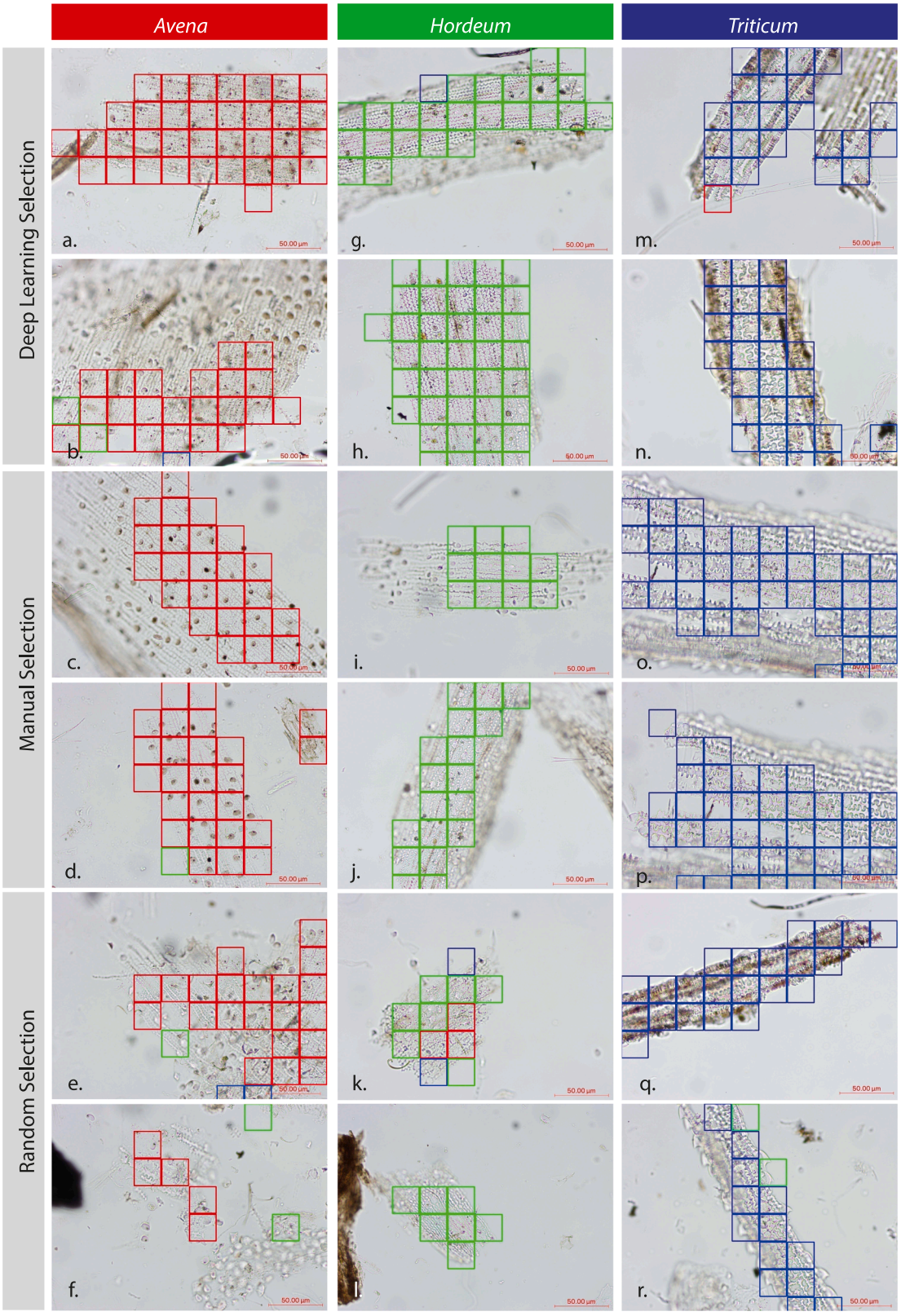

Fig. 2a. Algorithm results, arranged horizontally according to *Genus* and vertically according to how the survey image was selected. The algorithm's classifications are visually represented through colour coded tiles: red for *Avena*, green for *Hordeum,* and blue for *Triticum*.

material was burned off in a muffle furnace at 500 °C for 3 h. The resulting ash was then directly mounted onto slides using Entellan TM and covered with a cover slip (protocol outlined in, Weisskopf and Lee 2016). To ensure taxonomic and anatomical variability was captured within each category, images of *Avena*, *Hordeum* and *Triticum* were taken from multiple accessions (multiple slides of different plants from the same taxa) (*Avena*: 2 (lab ID: 1843, 1847; *Hordeum*: 3 (lab ID: 1841, 1840, 1846); *Triticum boeticum*: 2 (Lab ID: 1842, 1839); *Triticum*

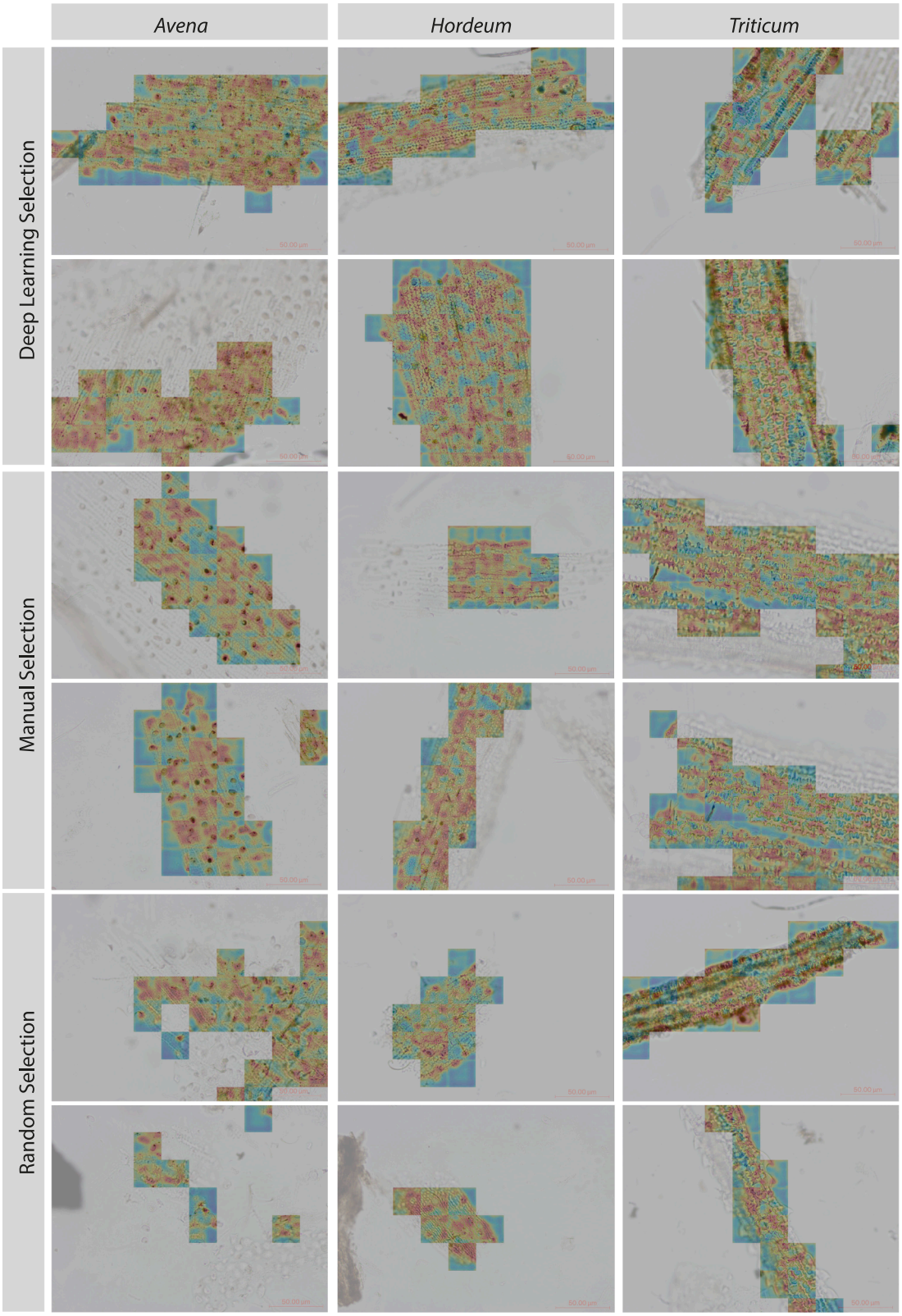

Fig. 2b. Heat Maps (visual explainer) of algorithm results, arranged horizontally according to *Genus* and vertically according to how the survey image was selected. The heat map images demonstrate what characteristics the algorithm emphasized (red) or deemphasized (blue) when classifying the phytoliths.

dicoccoides: 1 (Lab ID: 1844) (Lab ID refers to unique number assigned in RLEA Sample Tracker Database. Database includes all collector information)). This process ensured that our results are not biased by the unique characteristics of one plant or slide preparation. All images were captured using a Leica DM500 (magnification: 400x) and a GXCAM-U3-5 digital camera with ToupeLite software, producing images with a resolution of 2560 x 1922-pixels.

The DL model used in this study is based on a pretrained VGG19 convolutional neural network architecture (for further model and training details see, Berganzo-Besga et al., 2022). It is designed to classify or identify between wheat (*Triticum*), barley (*Hordeum*) and oats (*Avena*). The classification process begins by dividing each original image into 256 x 256-pixel RGB image tiles (see Fig. 2a). Each tile is then individually classified, and the overall identification of the entire original image is determined based on the highest number of tiles (or sub-images) identified relative to the class (e.g., *Triticum*, *Hordeum* and *Avena*).

The trained and validated model was tested with 190 microscope images (50.26% of the total dataset), which were excluded from the training phase. This test set included 64 images of *Avena sativa* (33.68% of the total), 62 of *Hordeum spontaneum* (32.64% of the total) and 64 of *Triticum boeticum/dicoccoides (*33.68% of the total). The model achieved an overall classification accuracy of 93.68%. Amusingly, the model achieved an average classification confidence greater than 98% for all images and classes included in the analysis. For further details concerning the training dataset, image preprocessing and classification procedures, please refer to the original paper (Berganzo-Besga, et al. 2022a) and original dataset via RLEA Zenodo: https://zenodo.org/records/15692103.

2.2. Survey Design

The annonymised survey was designed using SurveyMonkey (https://www.surveymonkey.com) an online platform that facilitates the development of customized surveys aligned with specific research objectives. The survey was structured to evaluate expert manual identification of multi-cell husk phytoliths using a targeted sample of 18 images comprising six each of wheat (*Triticum dicoccoides/boeticum*), barley (*Hordeum spontaneum*) and oats (*Avena sativa*). These images were selected from the original algorithm's testing dataset (Berganzo-Besga, et al. 2022). Image selection for the survey followed three criteria, with two images chosen per category: 1) *Manual Selection*—images with clear visibility of diagnostic characteristics used to make manual identifications, such as wave pattern and papillae (see Fig. 1); 2) *Deep-Learning Selection*—phytolith images whose heat map outputs display lots of red regions, indicating areas and features the DL model emphasized during classification (see Fig. 2b); and 3) *Random Selection*—images selected randomly to serve as a control. The images were categorized and selected by M.N Ramsey. Manual selection images were assessed based on M.N. Ramsey's experience manually identifying phytoliths. Deep-learning selection images were assessed by M.N. Ramsey by evaluating the heat-map images. Random selection images were selected by M.N. Ramsey manually scrolling and selecting images from their file names (no visual cues). All data used in this paper, including raw survey data, are available on the RLEA Zenodo (https://zenodo.org/records/15626668) (RLEA 2025b) and GitHub site (links to these repositories are provided under the heading, supplementary materials).

Each multi-cell phytolith image belonging to either wheat, barley or oats was presented on a separate page, resulting in 18 individual survey pages. Manual expert respondents were asked to identify the plant taxon represented in each image and respond to accompanying questions designed to elicit insights into their identification process. To facilitate detailed visual inspection, each page contained a link to the RLEA dropbox which allowed individual images to be downloaded if needed. However, it should be noted, most phytolith researchers tend to identify phytoliths directly under the microscope, making use of zoom and focus adjustments to improve precision. Working from images gives the DL algorithm an admitted advantage.

The first question on each page employed a sliding confidence scale bar, ranging from 0 to 100, with higher values indicaiting a higher degree of confidence that the correct identification had been made. This measure was included to assess the relationsip between expert confidence and classification accuracy, particularly in images that may be difficult to identify, involving morphologically ambiguous or diagnostically challenging phytoliths.

Following the confidence scale, respondents were presented with a multiple-choice question asking them to note which morphological characteristics they used in making their identification. This question was based on criteria first outlined by Rosen (1992), and included: 1) wave pattern (Fig. 1a), the negative space between dendritic long-cells; 2) papilla size, shape, and regularity (Fig. 1b) (a characteristic used to differentiate between wild vs. domestic taxa), and; 3) the number of pits around the papilla (Fig. 1c). The inclusion of this question aimed to document the decision making process of manual experts by identifiying both *how* and *what* features are considered during identifcation. Specifically, it sought to determine which diagnostic traits were consistently referenced, underutilized or overlooked and whether there were commonalities in the features that were used or not used across respondents. This information offers a valuable point of comparison for evaluating differences in diagnostic feature prioritization between experts and AI-based identification.

Finally, a comment box was included to allow respondents to provide any additional information they deemed relevant to their identifcation process. This open-ended response option served as a supplementary component to the previous question, offering deeper insight into *why* certain morphological features are potentially utilized over others.

The survey was also designed to record the total completion time, beginning with the first question and running until completion. This metric aimed to account for possible delays during the identification process and to capture the inherently human element, distraction, which may result from a variety of external factors – helping to demonstrate a critical benefit of the AI-based approach, speed.

To minimize order effects, the pages of the survey were randomized for each respondent. This was implemented by randomizing the sequence in which page numbers appeared—for example, one respondent may start on page 18 followed by page 6 whereas a second respondent may start on page 14 followed by page 8. For consistency in analysis, the data was recorded and interpreted according to the original numerical sequence of pages.

2.3. Survey respondents

Survey respondents were solicited by first inviting established lab directors and professors, who have published papers featuring cereal husk identifications and therefore demonstrated the appropriate expertise to participate in this survey, those researchers then forwarded the survey invitation to their appropriately trained lab members. This resulted in twelve researchers, at various stages of their careers, ultimately participating in the survey, although one individual did not finish. Respondents experience level was arbitrarily catigorized based on their years of experience post PhD, resulting in 3 experience categories (PhD Candidates (n = 2); "Less than 5 years" post-PhD (n = 2); "More than 5 years" post-PhD (n = 7). Whilst this is not a large survey, the group of researchers with appropriate "expertise" is limited. Accordingly, it was decided that we had an adequate survey size. But, this lack of available phytolith experts, a reality any site director can attest to, only emphasizes why AI tools holds such transformative potential, not only for improving the speed, volume and accuracy of phytolith analyses, but also for improving access to phytolith analyses.

3. Results

The online phytolith identification survey was conducted during the summer and fall of 2024, with twelve respondents and a completion rate of 92% (1 respondent did not finish) (see supplementary materials, available via the RLEA Zenodo (https://zenodo.org/records/15626668) and GitHub sites). On average, respondents completed the survey in fifty-two minutes. The survey presented 18 multi-cell husk phytolith images — six each from wheat (*Triticum boeticum/dicoccoides*, barley (*Hordeum spontaneum*) and oats (*Avena sativa*). These images had previously been analyzed (but not used for training) with our original DL model. The complete anonymised survey dataset, including respondent comments are available via the RLEA Zenodo (RLEA 2025b) and Github repositories (links provided below under the heading supplementary materials). The DL model correctly classified all survey images and did so instantaneously. Fig. 2a (tiles) and 2b (heat maps), illustrate the

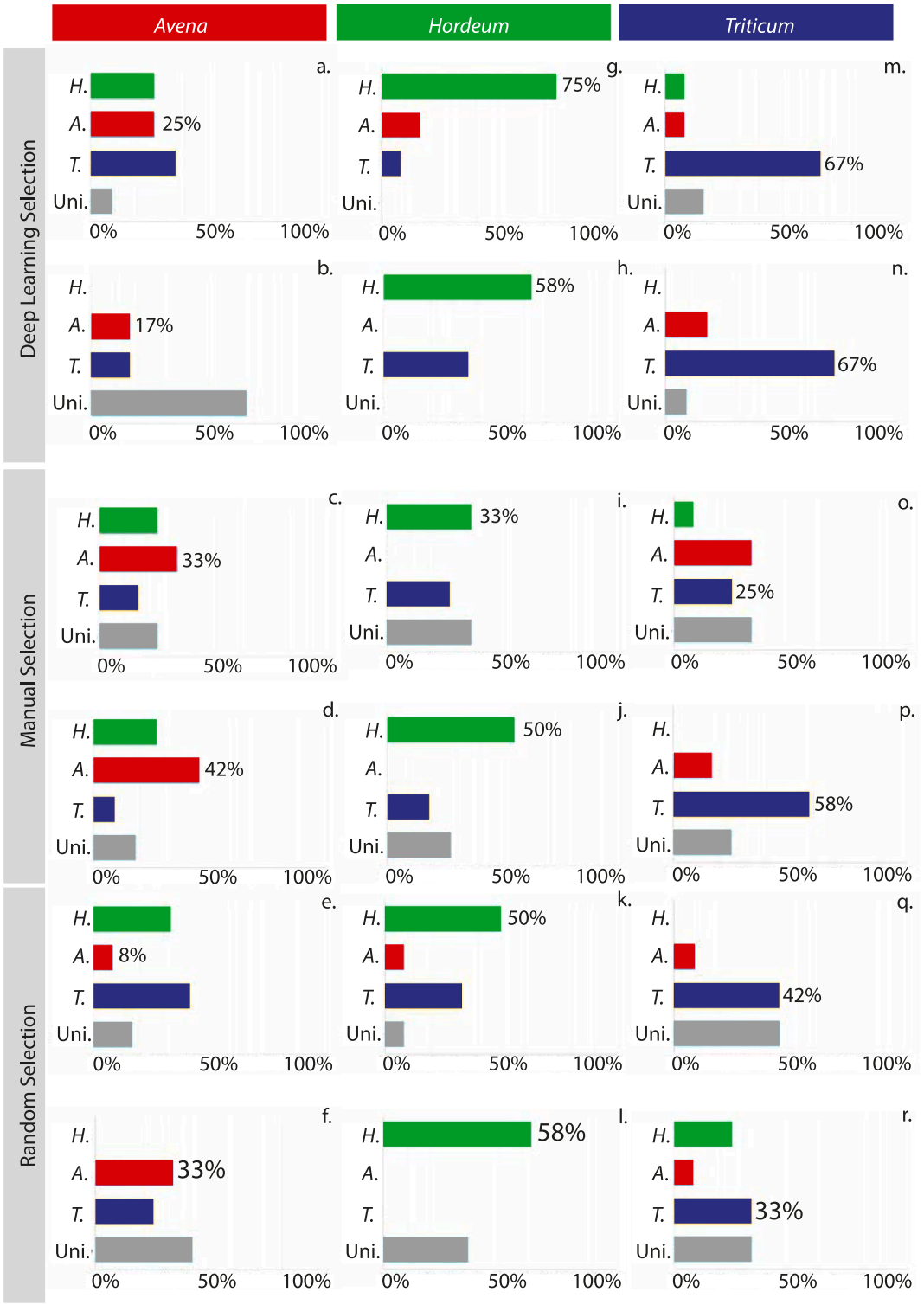

Fig. 2c. Breakdown of Survey Respondent Results, arranged horizontally according to *Genus* and vertically according to how the survey image was selected. The number listed at end of the histogram column reflects the percentage of respondents who correctly identified the phytolith.

identification process employed by the algorithm. As shown in figure 2a, each image is divided into smaller tiles, which are then individually categorized; the final classification of the image is determined by the category with the highest number of identified tiles. For example, in Fig. 2a, image k, the algorithm correctly identified the *Hordeum* (barley) multi-cell phytolith. However, the tiling process reveals that individual tiles within the image were variably classified as *Avena* (oat), *Triticum* (wheat) and *Hordeum* (barley) as indicated by red, blue and green tiles —red for *Avena*, blue for *Triticum* and green for *Hordeum*. This example demonstrates how the algorithm integrates information across the image, even when some tiles are classified incorrectly.

As part of a recent publication (Berganzo-Begsa, et al. 2025) visual explainers such Guided Grad-CAM were employed to clarify the internal decision making process of the DL model and address the so-called 'black box' problem. One of the methods employed, Grad-CAM, involved the use of heat maps as shown in Fig. 2b. These maps highlight the areas of the phytolith image that the algorithm emphasizes (red) or deemphasizes (blue) during the classification process. The application of visual explainers not only enhances transparency but also facilitates refinement of the training dataset. By revealing which morphological features may be confusing or misleading to the model, we can then strategically curate and augment training data to improve classification accuracy. More broadly, the integration of explainable AI methods is essential for demonstrating the integrity of DL approaches, thereby fostering wider acceptance of these models and tools within paleoethnobotany and archaeology.

The survey respondents only achieved an average accuracy of 44% (the average accuracy was calculated by adding up the number of correct responses from all respondents for each image and dividing that by the number of respondents, this value was then converted into a percentage). These results are broken down generally in Fig. 2c, and in more detail in Figs. 3-7. An analysis of the survey results (Fig. 3) reveals notable differences in identification accuracy among the three taxa: *Avena, Hordeum, and Triticum. Avena* exhibited the lowest correctness rate at 26.39%. In contrast, *Hordeum* had the highest correct identification rate (correctness) at 54.17%, followed closely by *Triticum* at 48.61% (correctness was calculated by dividing the number of correct answers by the total number of answers).

Statistical analysis confirmed that these differences were significant. An ANOVA test indicated a meaningful variation in identification accuracy between groups ($p = 0.0016$), prompting further pairwise comparisons (a one-way ANOVA was prepared using the means displayed in figure 3). Independent t-tests indicated that both *Hordeum* and *Triticum* were identified significantly more accurately than *Avena.* However, the difference in correctness between *Hordeum* and *Triticum* was not statistically significant, indicating comparable levels of accuracy in identification for these taxa.

Notably, in addition to correctness, we can also consider consistency. *Triticum* O and P (Fig. 2a) are the same multi-cell phytolith cropped differently, they are 'twin' images. Both images were selected for their visible manual identification characteristics. Whilst this selection occurred in error, they now provide an illuminating example of identification consistency or inconsistency. Seven out of twelve (7/12) of the respondents (58.3%) provided the same identification for both images, but only four respondents (33.3%, 4/12) correctly identified both as *Triticum*. Two respondents identified both as unidentifiable. One respondent identified both images as *Avena*. Five respondents (41.7%) provided different identifications for the 'twin' images (RLEA 2025b). *Triticum* O (Fig. 2c) has a correctness rate of only 25% and was more commonly identified as *Avena*, while *Triticum* P (Fig. 2c) has a correctness rate of 58%.

The survey results reveal variation in both performance and completion time based on respondent experience level (Fig. 4). PhD Candidates (green n = 2) completed the survey the fastest with a mean completion time of 32 min. However, their average score (44.5%) was lower than that of the "Less than 5 years" of experience group. Participants in the 5 years post-PhD "Less than 5 years" (orange N = 2) performed best overall with an average score of 72%, but their completion time was significantly longer (262.5 min on average). Notably, the highest individual score, exceeding 80%, was achieved by a respondent in this group, suggesting that recent training or more active engagement, may contribute to improved performance. More than 5 years of experience, classed as those 5-years or more post-PhD (blue N = 7) had the lowest average score (37.3%) and moderate completion time (89.3 min),

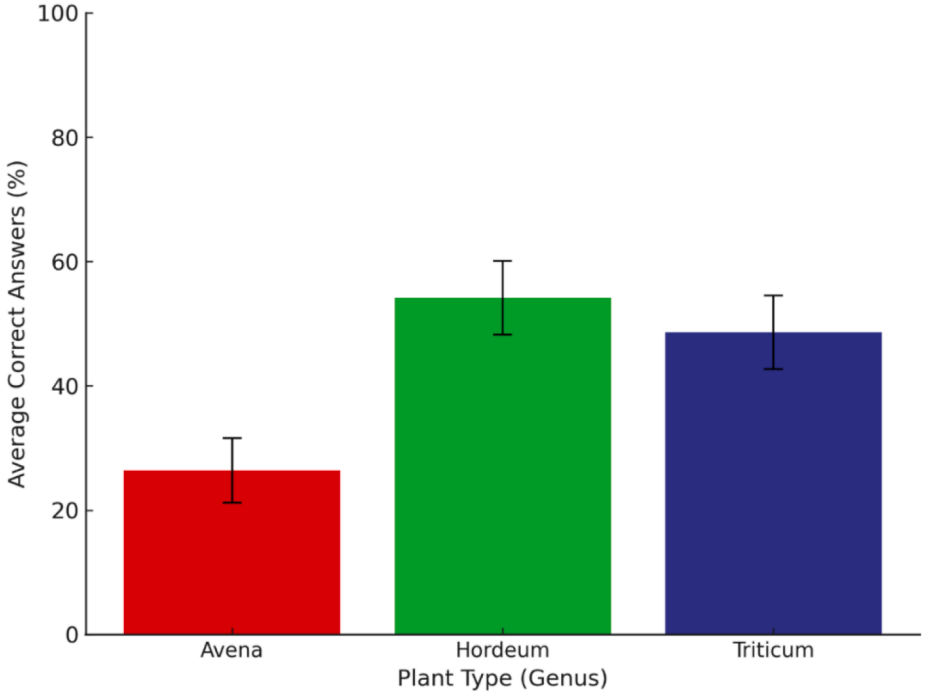

Fig. 3. Comparison of Correct Survey Respondent Answers Based on *Genus* The figure illustrates the mean correctness rates for each taxon included in the survey: *Avena sativa* (26.39%), *Hordeum spontaneum* (54.17%), and *Triticum dicoccoides/boeoticum* (48.61%).

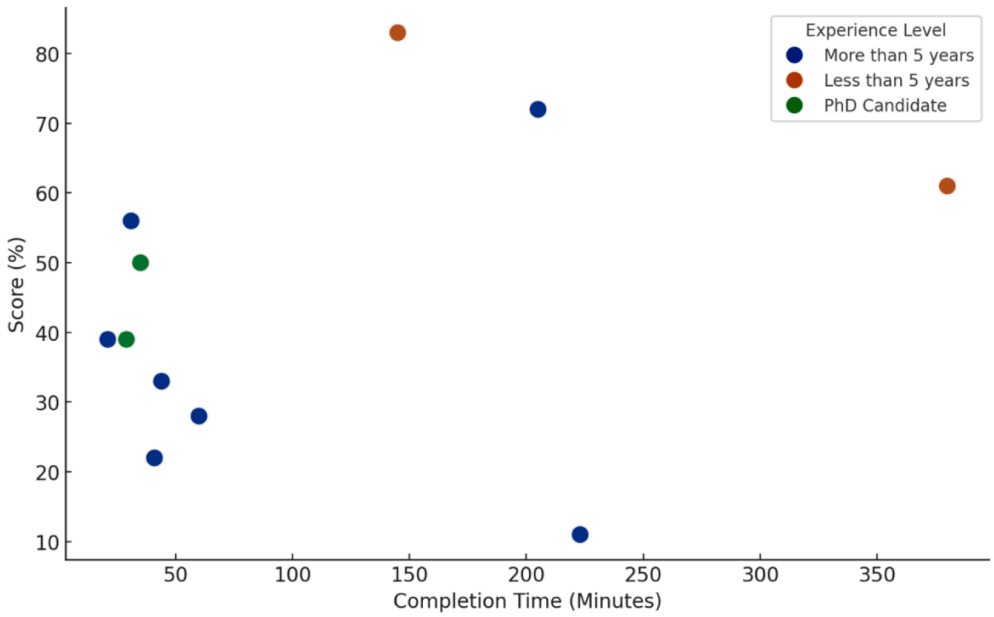

Fig. 4. Survey Performance and Completion Time by Experience Level The figure shows the average identification accuracy (score) and mean completion time for three respondent groups: PhD Candidates (green, n = 2), researchers with less than 5 years of post PhD experience (orange, n = 2), and researchers with 5 or more years of experience post-PhD (blue, n = 7).

demonstrating that more experience does not necessarily correspond to improved performance. However, it must be noted that because our survey size is limited, the number of individuals in certain experience categories is very limited (e.g., only 2 PhD Candidates). Accordingly, outliers (particularly strong performers or particularly poor performers) may have a disproportionate impact on our results.

Statistical analysis however found no significant correlation between completion time and score, suggesting that time spent on the task was not strongly associated with identification accuracy. Additionally, although differences in mean scores were observed across experience levels, these were not statistically significant, indicating that performance did not vary meaningfully by experience group within this sample.

Fig. 5 presents the frequency of unidentifiable phytolith identifications across genera, categorized by image selection type: DL, Manual, and Random. Within the survey, an 'unidentifiable' classification refers to cases in which respondents indicated that a phytolith lacked discernible diagnostic features (as outlined in Fig. 1).

Among the three genera, *Avena* exhibited the highest overall unidentifiable rate. Specifically, DL selected images of *Avena* had a relatively high unidentifiable rate of 39.13%, manual selection images had the lowest unidentifiable rate of 21.74%, and randomly selected images showed an intermediate rate of 29.17%. For *Hordeum*, DL selected image had a 0% unidentifiable rate, while manual and random images had

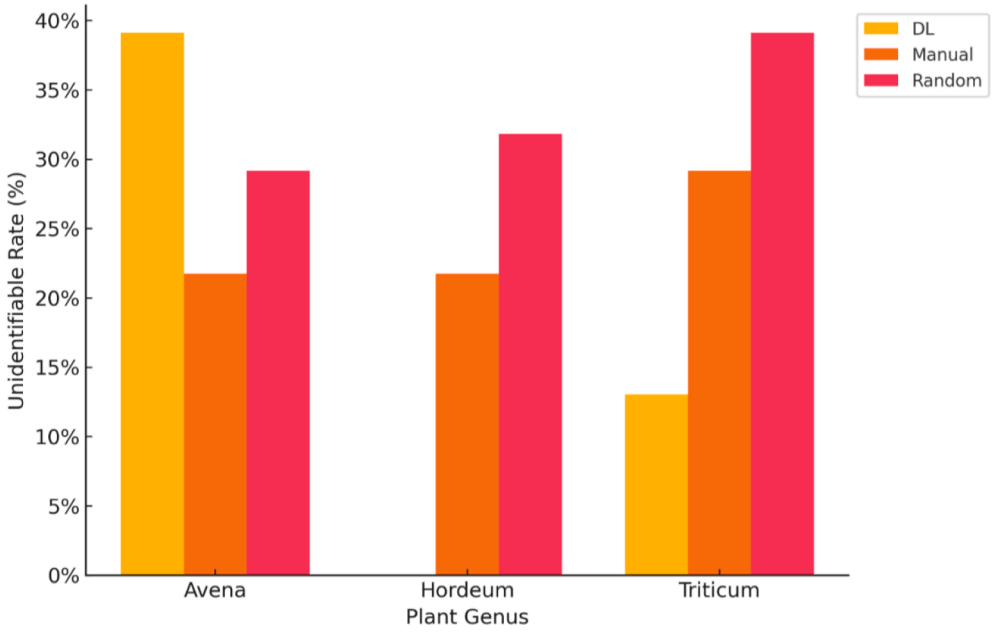

Fig. 5. Frequency of Unidentifiable Phytolith Identifications by Genus and Image Selection Type. Three image selection types are displayed (DL, Manual, Random). Unidentifiable responses refer to cases where respondents indicated that diagnostic characteristics were not present. Data is shown for *Avena*, *Hordeum* and *Triticum* across all three image selection categories.

rates of 21.74% and 31.82% respectively. In the case of *Triticum*, DL selected images had a mean unidentifiable rate of 13.04%, manual selections had a mean unidentifiable rate of 29.17%, and randomly selected images had the highest unidentifiable rate at 39.13%. Across all responses, *Avena* accounted for 21 unidentifiable instances (30%), *Hordeum* for 12 instances (17.65%) and *Triticum* for 19 instances (27.14%).

Fig. 6 presents the confidence scores associated with correct and incorrect identifications across all three plant types: *Avena, Hordeum, Triticum*. Across all responses, the average confidence score was 45.6, while correct answers had a higher average confidence score of 52.8. Whilst the difference in confidence scores is statistically significant ($p = 0.049$), there is clear overlap between the two distributions, which suggests confidence is not a good indicator for correctness. It is important to note that this analysis includes only instances where an identification was made and excludes responses marked as unidentifiable.

For *Avena*, the average confidence score for correct answers is 49.2, compared to 48.4 for incorrect responses. The nominal difference between the correct and incorrect scores suggest that participants' confidence levels remained relatively stable, regardless of accuracy. For *Hordeum*, the pattern was reversed: the average confidence score for correct answers is 49.5, whereas incorrect answers had a higher average confidence score of 56. In contrast for *Triticum*, the average confidence score for correct answers was 58.6, while incorrect responses averaged just 29. As noted previously, the model achieved an average classification confidence score greater than 98% for all images and classes included in the analysis (see, Berganzo-Besga, et al. (2022) and original dataset via RLEA Zenodo: https://zenodo.org/records/15692103).

The data displayed in Fig. 7 summarizes the correlation between number of characteristics employed to make a correct or incorrect identification. While the results on average weakly support the commonly held assertion that using more characteristics produces more secure or correct identifications for *Hordeum (correct 1.6* vs. *incorrect 1.4)* and *Triticum* (correct 1.3 vs. incorrect 1.2), the results for *Avena* (correct 1.7 vs. incorrect 1.8) indicate that correct identifications were generally associated with the use of fewer diagnostic characteristics.

In reference to the histogram, these results demonstrate that whether identified correctly or incorrectly, manual experts tend to employ one or two characteristics (wave pattern and papillae shape) for *Avena* and *Hordeum*, while they tend to favor employing only one characteristic for *Triticum* (wave pattern). This likely reflects the fact that *Triticum* has a more distinct wave pattern. Overall, and perhaps surprisingly, these results do not strongly support the commonly held assertion that using more characteristics provides a more secure identification.

All respondent comments demonstrated careful consideration and justification in the identification process. Wave Pattern was the most consistently prioritized diagnostic feature across all three genera (Table 1). Respondents frequently described wave pattern using terminology aligned with Rosen's (1992) typology such as "squarish" for *Hordeum*, "lobed" or "crenelated" for *Triticum*, and "thin" or "pointed" for *Avena*.

Papilla shape and number of pits was also commonly referenced, though these features were often noted as difficult to evaluate due to limitations in image focus, magnification, or visual obstruction. When visible, pit counts were compared to reference materials (e.g., Rosen 1992), with specific counts (e.g., ~10 for *Triticum,* ~12–18 for *Hordeum*) contributing to identification decisions.

Many respondents noted challenges with image focus, magnification, or overlapping silica skeletons, leading to unidentifiable classifications or low confidence. Several expressed a desire for microscope-like control (zoom, focus adjustment). Related to this, respondents often stated they would not identify phytoliths in archaeological samples without clearer images or additional features (e.g., morphometrics, stomata), reflecting a cautious approach to *Genus*-level identifications. These comments however also demonstrate that manual identification requires more time, and can only be effectively employed on a small subset of images because it relies on more features (that all need to be clearly displayed), whereas automatic identification, even with blurred or partially out-of-focus images, still achieves high efficiency.

The respondent comments provide critical insight into the realities of manual phytolith identification. Hordeum K (Fig. 2a) was commented on by one respondent as follows "*Unclear / part of a husk where the dendritic are v narrow – very tentatively barley, but almost said non-ID.*" This comment reveals the respondent's awareness of the phytolith's specific anatomical context (part of a husk with narrow dendritics), which influences identification. The tentative identification paired with near-unidentifiable status shows a cautious approach, highlighting the challenge of manual identification and the reality that many multi-cells are considered unidentifiable by manual experts. When identifying Avena E (Fig. 2a), one responded noted, *"I was unsure about this one. I*

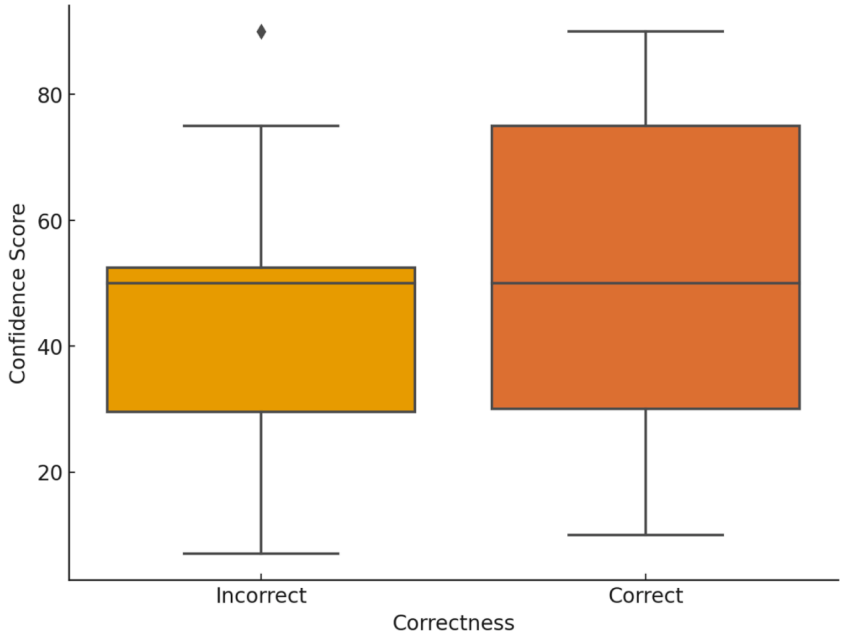

Fig. 6. Confidence Score Based on Correctness. Box plots showing the distribution of confidence scores for correct and incorrect responses.

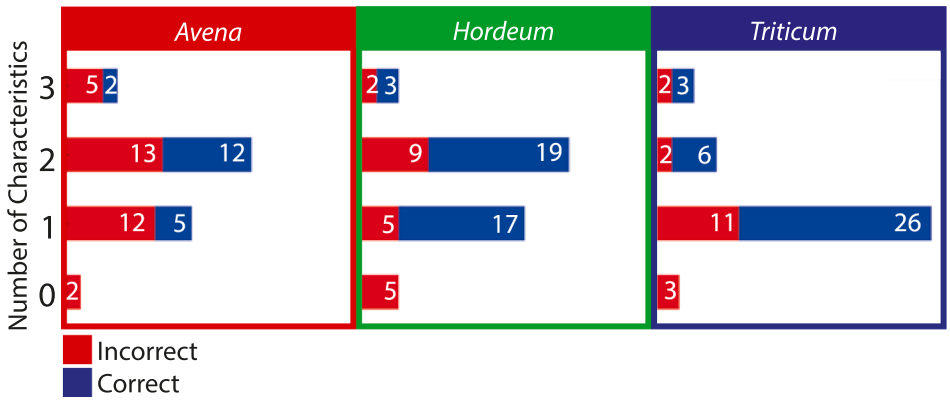

Fig. 7. Comparison of the number of characteristics used for correct (blue) and incorrect (red) responses. Characteristics included were Wave Pattern, Papilla Shape and Pits around Papilla.

Table 1
Characteristics used by Survey Respondents to Make Identifications Based on Survey Comments and Multiple-Choice Questions. Compiled using a Large Language Model (LLM). LLM was provided with Excel file, 'Multi-Cell Near East Husk Identification Survey', and prompted as follows: "This excel file summarises the results of a phytolith identification survey. Can you pull out the survey respondents comments for why they made certain identifications?" (RLEA 2025b).

Characteristic	Description	Number of Mentions*
Avena		
Wave Pattern	thin and pointed	27 (60)
Papilla Shape	ovoid and irregular in size	12 (36)
Long Cell Shape	thin	8
Pit Count	High number of pits around papillae	5 (16)
Hordeum		
Wave Pattern	Squarish, flattened, even and low aplitude	42 (59)
Papilla Shape	Pointy ('Hershey kiss')	16 (33)
Pit Count	Number of pits around papillae	10 (12)
Triticum		
Wave Pattern	Lobed, rounded ('crenelated'), wiggly	46 (62)
Long Cell Shape,	Thick, rounded margins, highly dendritic	17
Papilla Shape	Ovoid, small	12 (17)
Pit Count	Number of pits around papillae	7 (8)

*The first value represents the number of mentions in the survey text box, while the second value, in parenthesis, represents the number of times that characteristic was selected in the multiple-choice question asking them to note which morphological characteristics they used in making their identification.

thought the wave and cork cell shape looked like barley but the pits around the papillae were more in line with wheat." This comment highlights the wide range of characteristics that manual experts employ, beyond those outlined in Rosen (1992), but also highlights conflict arranging diagnostic features into an effective categorization hierarchy (barley-like wave and cork shape vs. wheat-like pit count), showcasing the complexity of phytolith identification. In contrast, another responded cited "gut feeling" in their assessment of Triticum Q (Fig. 2a), although their gut, unfortunately, was wrong (*Avena*). This candid admission, acknowledging intuition, is honest, and reflects a feeling that often comes with experience, perhaps one that experts should embrace cautiously. In contrast, another respondent when identifying Hordeum H (Fig. 2a) noted, *"I suggest Hordeum, but in my opinion, Hordeum is very difficult to ID reliably."* The explicit scepticism about the reliability of *Hordeum* identification suggests a broader critique concerning the reliability of manual phytolith identification to the *Genus* level. A scepticism this survey demonstrates is warranted.

4. Discussion

The findings of this paper challenge some of the commonly held norms and expectations in manual phytolith analyses, including: the idea that the husks of grass *Genera* and particularly cereal *Genera* can be accurately and consistently identified through manual microscopy; that analyst experience correlates with better results; that taking your time with identifications should produce more accurate results; and that using multiple diagnostic characteristics leads to more correct and secure identifications.

First, the variation in identification accuracy observed across the three cereal taxa reveals important insights into the challenges of manual phytolith classification. Among the taxa evaluated, *Avena* was consistently the most difficult for participants to identify with confidence, showing the lowest overall accuracy. This pattern likely reflects a combination of factors, including limited exposure to *Avena* in comparative collections and more ambiguous or variable morphological traits that complicate manual identification. In contrast *Hordeum* and *Triticum* were identified with greater accuracy, and at statistically similar levels. This increase in accuracy suggests that respondents may be more familiar with these taxa or that their diagnostic features are more readily recognized. Unfortunately, despite this increase in accuracy, overall identification rates for these taxa remain relatively low suggesting that even for more familiar or morphologically distinct cereals, manual identification continues to pose challenges. Regarding consistency, our results suggest that even when faced with the same *Triticum* image twice, only seven out of twelve (7/12) of the respondents (58.3%) provided the same identification for both images, and only 33.3% (4/12) correctly identified both as *Triticum*. These results highlight the limitations of manual identification. Importantly, the statistical significant difference in identification accuracy between *Avena* and the other two taxa emphasizes the need for targeted methodological improvements. For more challenging genera such as *Avena*, enhanced training resources, expanded comparative collections, and refined classification criteria may be necessary to support more consistent and reliable manual identification.

Second, the survey results reveal that identification accuracy and time to completion vary according to respondent experience, but not in a straightforward or predictable manner. First, it is important to emphasize caution when interpreting trends within this limited dataset (PhD Candidates (n = 2); "Less than 5 years" post-PhD (n = 2); "More than 5 years" post-PhD (n = 7)). In spite of this, while one might expect that greater experience would correspond to higher accuracy, the data suggests a more nuanced relationship. Participants with less than five years of post-PhD experience achieved the highest identification scores overall, despite requiring the most time to complete the task. This group also produced the single highest score, indicating that recent training or more active engagement in analytical tasks may be a more important

factor than years of experience alone. PhD candidates, in contrast, completed the survey in the shortest time on average, yet their performance in terms of accuracy was more moderate. Respondents with more than five years of experience post PhD tended to score the lowest on average, suggesting that extended time away from 'hands-on' phytolith analysis, or reliance on more general expertise, may not support accuracy in tasks requiring fine morphological distinctions. The presence of outliers also reveals important behavioral patterns. Some individuals completed the survey very quickly and performed poorly—raising the possibility of inattentive or rushed responses in certain cases – whilst others, took substantially more time than their peers, but still did not achieve higher accuracy. Statistical analysis reinforces these qualitative observations. There was no significant correlation between completion time and accuracy, indicating that spending more time on the task did not necessarily improve performance. Likewise, observed differences in average scores across experience groups was not statistically significant. Together, these findings suggest that neither speed nor experience level can reliably predict success in manual phytolith identification. Instead, effective performance may hinge on factors such as recent 'hands-on' experience, familiarity with specific morphological traits, and sustained attention to detail, emphasizing the value of continual methodological practice and targeted preparation for those engaging in paleoethnobotanical analysis.

Third, the findings of this paper also challenge the common assumption that using multiple diagnostic features inherently leads to more accurate or secure identifications. While the use of more identifiers appears beneficial for *Hordeum* and *Triticum,* for the more difficult to identify *Avena*, it does not yield the same advantage. In fact, for *Avena* greater use of characteristics was associated with lower accuracy, suggesting that additional features may introduce confusion or lead to overinterpretation. This underscores the importance of not only the number but also the relevance and clarity of the features selected during the identification process.

Critically, this paper demonstrates that the DL model consistently and reliably detects diagnostic characteristics leadings to correct identifications, including images that pose a challenge for human identification. As outlined in the original publication of this model (Berganzo-Besga, et al. 2022), the VGG19 architecture features a deep stack of convolutional filters which automatically learns and detects low-level features (e.g., edges, textures) in early layers and progresses to higher-level, composite patterns in deeper layers. For multi-cell grass husk phytoliths, the most recent publication (Berganzo-Besga et al. 2025) shows that the model excels at capturing the diagnostic morphological characteristics originally defined by Rosen (1992) (see figure 1, and figure 2a), such as wave pattern, but also demonstrated unexpectedly, that different characteristics (papillae, dendritic long-cell shape, and wave pattern) may be emphasized for different *Genera*. For example, in *Avena* classifications, papillae are used in 94% of the images compared to 84% of those using the wave-pattern. This is a notable result because unlike the DL model, manual experts, as this survey demonstrates, *always* prioritize wave pattern. Berganzo-Besga et al. (2025) also found that the algorithm employed dendritic long-cell shape, as a category on its' own (i.e. rather than being an aspect of wave pattern). The subtle, repetitive, and interconnected structural details of multi-cell husk phytoliths are challenging for human analysts due to similarity across genera and form complexity. The DL algorithm's ability to 'see' and prioritize or weight different characteristics is simply superior to manual experts.

It should be noted that such low accuracy among experts (44%) does not indicate that this is common in their own research. Specialists may have invested less time and effort than they would have done in their own research; there were limitations due to specialists being given an image of the forms rather than viewing them under a microscope where the focus could be altered and different characteristics more clearly observed; some specialists may use a non-permanent mounting medium in their own analysis allowing phytoliths to be reangled aiding identification; and for difficult to identify forms specialists may use higher-powered microscopy allowing a much clearer and detailed view. Additionally, experts might choose not to identify many examples within an archaeological assemblage if they are deemed unsuitable (e.g., obscured or lack of features, unclear). For this reason, along with the others mentioned, the use of machine learning in phytolith identification becomes more relevant, as AI will always dedicate the same amount of time and effort to reliably generate its classifications.

Taken together, these findings emphasize the need for greater investment in both manual training and methodological refinement. In addition, the findings of this paper make clear that supporting analysts with better tools must include continued investment in the development of AI-assisted approaches. We advocate investing first in developing AI-assisted approaches for those morphotypes that manual experts already identify to genus, particularly those with complex 3D morphologies (e.g., Banana (*Musa* sp.*);* Enset (*Ensete* sp.). While a complete DL identification workflow for archaeological phytoliths is still some ways from being operational, developing targeted AI identification tools, such as the Berganzo-Besga et al. (2022) DL algorithm for the identification of multi-cell cereal husks, could in the short-term help reduce identification errors and improve the reliability of phytolith analyses. This work is currently being undertaken in the Ramsey Laboratory for Environmental Archaeology (RLEA), at the University of Toronto (Mississauga), and involves expansion of the training data set (e.g. new samples, multiple accessions), mass imaging of higher quality images (e.g. automatic z-stack imaging workflow development) and algorithm development (e.g. employing ensemble or stacking methods (Díez-Pastor, et al. 2024)).

5. Conclusion

The comparative analysis of AI versus manual expertise in phytolith identification presented in this paper, underscores a transformative shift in paleoethnobotanical research. The DL algorithm's flawless 100% result in classifying multi-cell husk phytoliths from wheat, barley, and oats starkly contrasts with the 44% average accuracy achieved by manual experts in the same survey, revealing the limitations of traditional manual methods. Notably, the significant difficulty in identifying *Avena* (26.39% accuracy) compared to *Hordeum* (54.17%) and *Triticum* (48.61%) highlights inconsistencies in human performance, particularly for less familiar taxa. The lack of correlation between experience, completion time, and accuracy further challenges assumptions about the reliability of manual expertise. By employing diagnostic features like wave patterns and papillae with superior precision, the DL algorithm not only outperforms human analysts but also offers, in the long-term, scalability for generating robust and accurate datasets critical to reconstructing past human environments. While manual expertise is still necessary for the application of phytolith analyses to archaeological contexts, this study advocates for the integration of AI-assisted tools to enhance accuracy and efficiency, signalling that automation, far from opposing expertise, is poised to redefine it.

Funding

This research was supported by a SSHRC Doctoral Award (M.P.) and a SSHRC MA Award (L.K.R.). During M.N.R.'s time as a Leverhulme Early Career Fellow (EFC-2020–318) M.N.R. and was awarded a D M McDonald Research Grant from the McDonald Institute for Archaeological Research (Deep Origins: AI Deep Learning ID of Plant Phytoliths for the Origins of Agriculture) which funded I.B.B's original analysis and investigation. I.B.B. acknowledges his AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR.

CRediT authorship contribution statement

Monica N. Ramsey: Writing – original draft, Investigation, Conceptualization. **Melanie Pugliese:** Writing – original draft, Investigation. **Lachlan Kyle-Robinson:** Writing – original draft, Investigation. **Iban Berganzo-Besga:** Writing – review & editing, Software, Investigation. **Rebecca Roberts:** Writing – review & editing, Investigation. **Jennifer Bates:** Writing – review & editing, Investigation. **Francesca D'Agostini:** Writing – review & editing, Investigation. **Zachary C. Dunseth:** Writing – review & editing, Investigation. **Thomas C. Hart:** Writing – review & editing, Investigation. **Emma Jenkins:** Writing – review & editing, Investigation. **Carolina Jiménez-Arteaga:** Writing – review & editing, Investigation. **Celine Kerfant:** Investigation. **Marco Madella:** Writing – review & editing, Investigation. **Sigrid Osborne:** Investigation. **Robert Power:** Writing – review & editing, Investigation. **Abel Ruiz-Giralt:** Investigation. **Philippa Ryan:** Writing – review & editing, Investigation.

Supplemental files

The supplementary materials include: grok_report.pdf; individual question summaries.docx; Survey Monkey – comments per question.xlsx; Survey Monkey – Overall accuracy.xlsx; Survey Monkey – responses per question.xlxs; Survey Monkey – Full survey preview.pdf. These materials are available online at 10.5281/zenodo.15626237 (RLEA 2025b). These data can also be found in GitHub: https://github.com/Ramsey-Environmental-Archaeology-Lab/Survey-Paper-. Original survey data available online at 10.5281/zenodo.15692102 (RLEA 2025a).

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

I have shared the repository URLs above.

References

Andriopoulou, N.C., Petrakis, G., Partsinevelos, P., 2023. Twenty thousand leagues under plant biominerals: a deep learning implementation for automatic phytolith classification. Earth Sci. Inf. 16 (2), 1551–1562.

Berganzo-Besga, I., et al. 2022a Automated detection and classification of multi-cell Phytoliths using Deep Learning-Based Algorithms. Journal of Archaeological Science 148.

Berganzo-Besga, I., et al., 2025. Deep learning black box and pattern recognition analysis using Guided Grad-CAM for phytolith identification. Ann. Bot. 136 (2), 355–366.

Berganzo-Besga, I., et al. 2021 Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia. Remote Sens. 13.

Berganzo-Besga, I., et al., 2022. Automated detection and classification of multi-cell Phytoliths using Deep Learning-based Algorithms. J. Archaeol. Sci. 148, 105654.

Bickler, S.H., 2021. Machine Learning Arrives in Archaeology. Adv. Archaeol. Pract. 9 (2), 186–191.

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 (4), 377–384.

Character, Leila , et al. 2025 Broadscale Deep Learning Model for Archaeological Feature Detection Across the Maya Area. Journal of Archaeological Science (Available at SSRN: https://ssrn.com/abstract=4744069 or DOI: 10.2139/ssrn.4744069).

Díez-Pastor, J.-F., et al., 2024. Towards automatic phytolith classification using feature extraction and combination strategies. Prog. Artif. Intell. 13 (3), 217–244.

Díez-Pastor, J.-F., et al., 2020. "You are not my Type": an Evaluation of Classification Methods for Automatic Phytolith Identification. Microsc. Microanal. 26 (6), 1158–1167.

Garcia-Molsosa, A., et al., 2020. Potential of deep learning segmentation for the extraction of archaeological features from historical map series. Archaeol. Prospect.

Mnasri, Z., D'Andrea, A., 2025. Automatic inventory of archaeological artifacts based on object detection and classification using deep and transfer learning. Digital Appl. Archaeol. Cult. Heritage 39, e00458.

Orengo, H.A., et al., 2020. Automated detection of archaeological mounds using machine learning classification of multi-sensor and multi-temporal satellite data. PNAS 117 (31), 18240–18250.

Orengo, H.A., Garcia-Molsosa, A., 2020. A brave new world for archaeological survey: automated machine learning-based potsherd detection using high-resolution drone imagery. J. Archaeol. Sci. 112, 105013.

Orengo, H.A., et al., 2021. New developments in drone-based automatic surface survey: Towards a functional and effective survey system. Archaeol. Prospect. 1–8.

Out, W.A., et al., 2024. Inter- and intra-observer variation in phytolith morphometry. Ann. Bot. 135 (5), 851–866.

Parisotto, S., et al., 2022. Unsupervised clustering of Roman potsherds via Variational Autoencoders. J. Archaeol. Sci. 142, 105598.

Pawlowicz, L.M., Downum, C.E., 2021. Applications of deep learning to decorated ceramic typology and classification: a case study using Tusayan White Ware from Northeast Arizona. J. Archaeol. Sci. 130, 105375.

Power, R.C., et al., 2015. Dental calculus evidence of Taï Forest Chimpanzee plant consumption and life history transitions. Sci. Rep. 5 (1), 15161.

Rlea, 2025. Automated detection and classification of multi-cell Phytoliths using Deep Learning-based Algorithms. Zenodo. https://doi.org/10.5281/zenodo.15692102; https://zenodo.org/records/15692103.

RLEA 2025b Experts Against Automation? Comparing Artificial Intelligence and Human Identification of Phytoliths. Zenodo, DOI: 10.5281/zenodo.15626237; https://zenodo.org/records/15633124.

Rosen, A.M., 1992. Preliminary identfication of silica skeletons from Near Eastern archaeological sites: an anatomical approach. In: Rapp, G.J., Mulholland, S.C. (Eds.), Phytolith Systematics, Emerging Issues. Plenum Press, New York, pp. 129–147.

Turan, D. E., et al. 2023 Interpreting Hyperspectral Remote Sensing Image Classification Methods Via Explainable Artificial Intelligence. International Geoscience and Remote Sensing Symposium (IGARSS), 2023. Vol. 2023-July, pp. 5950-5953.

Weisskopf, A.R., Lee, G.-A., 2016. Phytolith identification criteria for foxtail and broomcorn millets: a new approach to calculating crop ratios. Archaeol. Anthropol. Sci. 8 (1), 29–42.

Yang, LiHua, WenBo Zhou, and WangRen Qiu 2025 Chronological classification of Ming and Qing dynasty ceramics images based on an enhanced ResNet50 model. STAR: Science & Technology of Archaeological Research 11(1):e2498260.