Journal of Computer Applications in Archaeology

Comparative Analysis of CNN Architectures and Loss Functions on Age Estimation of Archaeological Artifacts

RESEARCH ARTICLE

SHARON YALOV-HANDZEL

IDO COHEN

YEHUDIT APERSTEIN

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

Automated age estimation of archaeological artifacts is crucial for categorization and dating, yet challenging due to variations in characteristics, degradation, and limited chronological information. This study investigates the performance of Convolutional Neural Network (CNN) architectures and loss functions for accurate age estimation. Using a dataset of about 10,000 labeled images from distinct archaeological sites, spanning 16 periods ranged from the Paleolithic to the Late Islamic periods, our results demonstrate top-5 accuracy above 90%. Notably, our empirical results revealed that InceptionV3, while known for its strong performance in object recognition tasks, outperformed other architectures in this classification task. Additionally, we found that conventional cross-entropy loss functions can, in some architectures, outperform ordinal cross-entropy, challenging conventional wisdom. Our findings not only advance the computational methodologies available for artifact dating but also provide critical insights into the nuanced selection of neural network architectures and loss functions, thereby opening new avenues for research in computational archaeology.

CORRESPONDING AUTHOR:

Dr. Sharon Yalov-Handzel

Afeka, Tel-Aviv Academic College of Engineering, Israel

sharony@afeka.ac.il

KEYWORDS:

Deep Learning; CNN; Image Classification; Ordinal Classification; Archeological Dating

TO CITE THIS ARTICLE:

Yalov-Handzel, S, Cohen, I and Aperstein, Y. 2024. Comparative Analysis of CNN Architectures and Loss Functions on Age Estimation of Archaeological Artifacts. *Journal of Computer Applications in Archaeology,* 7(1): 185–194. DOI: https://doi.org/10.5334/jcaa.136

1. INTRODUCTION

Accurately dating archaeological findings is a challenging task that, beyond its importance for archaeology itself, can have significant impacts on the conclusions drawn by historians. Reliable dating enables archaeologists to construct precise chronologies, which allow historians to better understand migration patterns, cultural changes over time, influences between civilizations, and other anthropological questions. For example, accurately dated artifacts help track the spread of technologies and ideas between groups.

Artifacts often exhibit variations in their characteristics, such as material composition, preservation state, and stylistic changes over time. All these cause difficulty in the establishment of precise chronological relationships. Moreover, artifacts can undergo degradation and alteration over time, further complicating the estimation of their age. The skill of classifying artifacts according to their age requires prior knowledge and expertise in a certain archeological period. Additionally, limited availability of reliable chronological information, particularly for lesser known or poorly documented archaeological sites, adds to the complexity. Methodological limitations also pose a significant hurdle. Traditional dating methods, such as radiocarbon dating, come with their own sets of limitations, including a range of errors and constraints related to the type of material that can be dated. These limitations can make precise dating a challenging endeavor. Furthermore, the complexity of accurate dating often requires a multidisciplinary approach. Incorporating methods from physics, chemistry and geology is often necessary for more accurate results. However, this can be both logistically and financially challenging, requiring specialized equipment and expertise that may not be readily available.

This is where the necessity of computational methods in archaeology becomes evident. Casini et al. 2021 highlight the increasing availability of data in archaeology, emphasizing the growing need for computational methods to deepen our understanding of archaeological issues.

Computer Vision (CV) and Machine Learning (ML) technologies offer a promising avenue for overcoming these challenges by enabling efficient and systematic analysis of large volumes of archaeological data. Specifically, CV algorithms can automatically extract valuable features from artifacts, capturing key details such as shape, texture, and pattern (Itkin et al. 2019, Wu 2021, Zhou 2022). ML models can then be trained using these features to recognize and classify artifacts based on their age. Utilizing this computational approach not only reduces the subjectivity and bias inherent in human interpretation but also provides a more objective data-driven method for age estimation. Moreover, ML models have the ability to identify subtle patterns and correlations within the data that may not be immediately apparent to human observers, thereby enhancing the accuracy and precision of age estimation. The authors (Itkin et al. 2019, Wu 2021, Zhou 2022) highlight the concept of 'barrier of meaning', which refers to the gap between the knowledge in the expert's mind and the knowledge grasped by the machine underscoring the complexities involved in fully automating the interpretation of archaeological data. Moreover, computational methods can integrate diverse types of data, from radiocarbon dating results to stratigraphic information, to produce more reliable age estimates. They can also adapt to new findings, continuously refining their models for greater accuracy (Parisotto et al. 2022, Xu et al. 2023).

Recently, Deep Learning (DL) has shown a particular promise in addressing the complexities of archaeological research. DL algorithms can automatically learn to identify intricate patterns in data, making them well-suited for tasks that require a high level of detail and precision. For instance, in the paper of Parisotto et al. 2022, a Variational Autoencoder (VAE) is employed to cluster Roman potsherds based on visual similarities. This VAE-based approach outperforms other methods in terms of clustering quality, demonstrating the potential of deep learning in archaeological artifact studies. Similarly, research by Reese 2021 has successfully applied neural networks for the dating of residential site occupations, further showcasing the versatility of DL in archaeological analysis. This approach, which predicts annual residential occupation with high accuracy, exemplifies the potential of DL in historical reconstruction and demographic analysis.

In addition to these advancements, Sakai et al. 2023 also utilized DL, specifically in object detection on aerial photographs, to discover new Nasca geoglyphs, demonstrating the potential of AI in enhancing the efficiency and scope of archaeological surveys. Expanding the scope of CNN applications in archaeology, Lu et al. 2018 utilize a CNN to segment curve structures from depth maps of pottery sherds. This method excels in capturing the nuanced features of the artifacts, thereby aiding in their analysis and potentially their dating (Zhou 2022). Following this, Pawlowicz and Downum 2021, applied CNN to classify digital images of decorated pottery sherds, specifically Tusayan White Ware from the American Southwest. Their study, which achieved accuracy levels comparable to or exceeding that of expert archaeologists, demonstrates the potential of CNN in artifact classification and typology. Such applications of DL are not limited to ceramics; they have also been applied to tasks like periodic discrimination of lithic assemblages and differentiation of bone surface modifications (Siozos et al. 2021). Additionally, DL is making strides in the age estimation of archaeological artifacts. While some efforts have been made to manually extract predefined features for age classification, achieving moderate accuracy

(Cifuentes-Alcobendas and Domínguez-Rodrigo 2019, Domínguez-Rodrigo et al. 2020), the trend is increasingly moving toward automated feature extraction. For example, a study combines Raman spectroscopy with ML algorithms to quantitatively estimate different degrees of thermal alteration on Flint artifacts (Agam et al. 2020). This approach not only automates the feature extraction process but also enhances the accuracy of age estimation, further underscoring the versatility and efficiency of DL methods in archaeological research.

Building on these advancements in automated feature extraction and the growing body of work that leverages deep learning for archaeological applications, our research takes a slightly different yet complementary approach. In the present study, an alternative methodology is employed, which involves the utilization of an advanced, deep CNN. This network performs two functions: it extracts pertinent features from the data and subsequently executes the classification task.

The remainder of this paper is organized as follows: Section 2 outlines the problem and proposed solution. Section 3 summarizes our key contributions. Section 4 describes the methodology including CNN architectures, loss functions, and evaluation metrics used. Section 5 details the dataset. Section 6 presents and discusses results. Finally, Section 7 provides concluding remarks and implications.

2. PROBLEM SCOPE AND OUR SOLUTION APPROACH

This research is grounded in the work of Resler et al. 2021, which initially framed the age estimation of archaeological artifacts as a multi-class image classification problem. Utilizing a CNN, Resler's model was trained on the archeology dataset published by the Israeli Archeology Authority and achieved an accuracy level on par with human experts. Inspired by this work, our study seeks to refine the age estimation process in archaeology. To address the limitations of our dataset, we employ transfer learning techniques across multiple pre-trained CNN architectures. This strategy allows us to capitalize on the feature extraction capabilities of established models, thereby enhancing the robustness and generalization of our own model. In line with the findings of M. Lyons 2021, who emphasized the critical role of selecting the appropriate CNN architecture and hyperparameters in the task of fabric classification of archaeological artifacts, we also experiment with various configurations to optimize our model's performance. By leveraging different CNN architectures and fine-tuning hyperparameters, we aim to offer a more nuanced, data-driven method for age estimation that can adapt to new archaeological findings and continuously refine its accuracy and reliability. To enhance the model's performance, we explore the use of different loss functions, including both regular and ordinal types.

Numerous pre-trained CNNs, each with unique structure, have been developed for general classification tasks. These structural differences lead to varying performance levels. Current research is primarily aimed at determining the most effective network structure and settings for specific classification problems.

In this research we conduct a comparative analysis of three different pre-trained CNN architectures, with two different configurations for each. The goal of this research is to elucidate the influence of CNN architecture and loss function choices on various metrics used to evaluate the quality of classification. It is noteworthy that each metric responds differently to changes in the architecture.

Furthermore, the choice of the loss function, which forms the basis for the backpropagation process, significantly impacts the outcome of the image classification task.

In the context of multi-class classification tasks, the conventional cross-entropy loss function is frequently employed. This function measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label, thus leading the model to make more accurate predictions over time. However, when the class labels have a natural ordering, an ordered or ordinal variant of the cross-entropy loss function may be more appropriate. This function takes into account the order of the classes, and it penalizes predictions based not only on their correctness but also on their distance from the true class in the ordered sequence of classes.

The selection of an appropriate loss function is a critical step in the design of machine learning models and can significantly influence the model's performance. While it may seem intuitive to use an ordered variant of the cross-entropy loss function in such scenarios, this is not always the optimal choice.

The performance of different loss functions can be influenced by various factors, such as the specific characteristics of the data, the complexity of the model, and the training procedure. Therefore, it is often beneficial to experiment with different loss functions and to perform rigorous model validation to determine the most suitable choice for a given task.

3. OUR CONTRIBUTION

Our research focuses on the exploration and comparison of various pre-trained CNN architectures and loss functions for automated age estimation of archaeological artifacts. We conduct a comprehensive comparison of three distinct pre-trained CNN architectures: EfficientNetB3, ResNet50, and InceptionV3. Then, we explore the effects of two different loss functions, cross-entropy (CE) and ordinal cross-entropy (OCE), on the performance of the CNN architectures.

Our findings highlight the importance of selecting the appropriate loss function for a given task and demonstrate that the optimal choice can vary depending on the specific architecture used. We provide empirical evidence that, in some cases, conventional CE can outperform the intuitive choice of OCE, contributing to the ongoing discussion on loss function selection in machine learning research.

By leveraging different CNN architectures and fine-tuning hyperparameters, we aim to offer a more nuanced, data-driven method for age estimation that can adapt to new archaeological findings and continuously refine its accuracy and reliability. Our study underscores the importance of careful selection of CNN architecture, its configuration, and the choice of loss functions, providing a framework for further investigation in this domain.

4. METHODOLOGY

In this section, we elaborate on the methodology adopted for our study, specifically focusing on the choice of CNN architectures, loss functions, and evaluation metrics. We utilize three pre-trained CNN architectures: EfficientNetB3, ResNet50, and InceptionV3. These architectures were selected based on their proven efficacy in various image classification tasks. EfficientNetB3 is known for its balance between model size and accuracy, ResNet50 for its deep residual networks that solve the vanishing gradient problem, and InceptionV3 for its inception modules that capture multi-scale features.

The advantage of using pre-trained networks lies in the application of transfer learning, a technique that allows us to capitalize on the feature extraction capabilities of models initially trained on large datasets like ImageNet (Chatterjee et al. 2021). This is particularly beneficial for our specific task of archaeological age estimation, where the available training data is limited (Janković Babić 2023).

Our training process incorporates two distinct loss functions: CE and OCE. These loss functions were chosen to examine their impact on the model's performance in multi-class classification tasks.

Finally, we discuss the evaluation metrics employed to assess the effectiveness of our models in accurately estimating the age of archaeological artifacts. These metrics serve as a quantitative measure of our models' performance and guide future refinements.

4.1. CNN ARCHITECTURES

ResNet, or Residual Network, is a type of CNN that was introduced in He et al. 2016A. The key innovation of ResNet is the introduction of "skip connections" or "shortcut connections", which allow the gradient to be directly backpropagated to earlier layers. Its advantage is the fast training relative to other CNNs. It uses residual blocks which allow the propagation of the gradients effectively through very deep networks He et al. 2016B.

The specific architecture we use in this study is ResNet50, a 50-layer CNN. It starts with an initial convolutional layer and a max pooling layer. The core of the network consists of four sets of convolutional blocks, each containing multiple layers. Each block starts with a "bottleneck" layer that reduces and then expands the dimensionality of the input, helping to reduce computational complexity. The number of layers in each block is 3, 4, 6, and 3, respectively, totaling 16 blocks or 48 layers. Each block also includes a shortcut connection, which helps mitigate the problem of vanishing gradients during training. After the convolutional blocks, the network includes an average pooling layer followed by a fully connected layer, which outputs the final classification results. Due to its depth, sometimes it tends to be overfitted. We used the pre-trained model as was implemented in the Keras library (Rosebrock 2017).

The second model is based on the EfficientNetB3 [16] as is implemented in the Keras library. This CNN architecture uniformly scales depth, width and resolution dimensions. This CNN consists of 29 layers where 27 of them are convolutional, one is fully connected layer and the last one is a classification layer. It is considered to achieve improved accuracy, but its training takes a long time. EfficientNetB3 is newer than ResNet50, more efficient and achieves competitive accuracy in image classification tasks. It uses mobile inverted bottleneck convolutions (MBConv), an efficient version of traditional convolutional layers, and squeeze-and-excitation (SE) blocks, which recalibrate feature maps to focus on informative features. The 'B3' signifies the level of scaling applied to the base model, encompassing depth (number of layers), width (number of neurons), and resolution of the input image. This compound scaling method, unique to the EfficientNet family, enhances performance.

InceptionV3 (Szegedy et al. 2016, Agrawal 2021) is a CNN from the Inception family, known for its efficiency and performance in image classification. It was designed by Google researchers to improve accuracy and performance of image recognition. The 'V3' indicates it is the third version of CNN architecture, with several enhancements over its predecessors. InceptionV3 uses 'inception modules' with parallel branches of different operations, enabling the network to learn a wider variety of features. It also employs factorized convolutions, batch normalization, and label smoothing for improved computational efficiency and model generalization. In the context of our study, the ability of InceptionV3 to learn a broad range of features and its computational efficiency position it as a good option for the task of archaeological artifacts' age estimation. While InceptionV3 was designed to be highly versatile and efficient, it also makes the architecture more complex and potentially harder to interpret and optimize and less intuitive choice for our task.

InceptionV3 uses 48 layers that are a combination of convolutional, pooling and fully connected layers. Also, it uses 'inception modules', which are a series of layers that perform multiple convolutions of varied sizes and pooling operations in parallel. This model has approximately 23 million parameters and it was trained on the ImageNet dataset. The network has multiple auxiliary classifiers. The advantage of the InceptionV3 is that it is not sensitive to image scaling.

4.2. LOSS FUNCTIONS

In this study, we explore the effects of two different loss functions on the performance of the CNN architectures: Cross Entropy (CE) and Ordinal Cross Entropy (OCE). The choice of loss function is a critical step in the design of machine learning models and can significantly influence the model's performance.

The CE loss function is commonly used in multi-class classification tasks. It measures the performance of a classification model whose output is a probability distribution across the classes. The CE loss increases as the predicted probability distribution diverges from the actual distribution, thus leading the model to make more accurate predictions over time. Formally, for a single data point, the CE loss is defined as:

$$L_{CE}(y,p)=\sum_{k=1}^{K} y_k \log p_k$$

where y_k is the true label for class k (1 if the data point belongs to class k, 0 otherwise), p_k is the predicted probability for class k, and the sum is over all *K* classes.

The OCE loss function is an extension of CE that is used when the class labels have a natural ordering. This function considers the order of the classes and penalizes predictions based on their correctness and distance from the true class in the ordered sequence of classes. Formally, the OCE loss is defined as:

$$\mathrm{OCE}=\sum_{k=1}^{K} \frac{|y_k - p_k|}{K-1} y_k \log p_k$$

The selection of an appropriate loss function is a critical step in the design of machine learning models and can significantly influence the model's performance. While it may seem intuitive to use an ordered variant of the cross-entropy loss function in such scenarios, this is not always the optimal choice. The performance of different loss functions can be influenced by various factors, such as the specific characteristics of the data, the complexity of the model, and the training procedure. Therefore, it is often beneficial to experiment with different loss functions and to perform rigorous model validation to determine the most suitable choice for a given task.

4.3. EVALUATION METRICS

In this research, we leveraged four key metrics to evaluate our models' performance in artifact age estimation: Accuracy, Precision, Entropy, Top-3 Accuracy, and Top-5 Accuracy (Nagda 2019). Accuracy and Precision provided us with a measure of overall correctness and exactness of our predictions. Accuracy is defined as the proportion of correct predictions made by the model out of all predictions. Precision is a metric that considers the number of true positives (i.e., the number of items correctly identified as belonging to the positive class) in relation to the number of all positive predictions made. Given that the exact age of an artifact can often be ambiguous, experts typically consider a range of possible ages. Recognizing this complexity and uncertainty, we also utilized Top-3 and Top-5 Accuracy. These metrics, often used in this field, provide a shortlist of most likely ages, aligning with the expert practice of considering a range of possible ages, and offer insight into the model's confidence in its predictions. Additionally, they offer insight into the model's confidence in its predictions, which can be beneficial even when the exact class is not correctly predicted. To further understand the model's predictive behavior, we also examined the entropy of the predictions, which quantifies the uncertainty and variability in the model's output, providing a deeper understanding of its decision-making process under varying conditions.

5. DATA

The goal of this study is to train a CNN on images dataset of archeological artifacts in order to classify them according to their age. The dataset is publicly accessible on the Israel Antiquities Authority (IAA) website (http://www.antiquities.org.il/t/default_en.aspx). It contains about 10,000 images of archeological findings from 120 different Israeli sites categorized into 53 historical time periods from the Levantine hominin history. The attribution of periods was provided by archaeologists working for or within the Israel Antiquities Authority (IAA) and available at their website.

The resolution of the images is 600 × 600 pixels. The data is labeled with the archeological site and age, so that there are 1,262 different classes. In this work the site label was ignored, and the archaeological periods were categorized into 16 rough classes according to IAA and Israeli Institute of Archaeology (https://www.israeliarchaeology.org) definition as described in Table 1.

The histogram in Figure 1 represents the distribution of the training images among these time periods. It is clearly seen that the data is imbalanced, which might cause bias in the training process.

The data was divided into training and test sets in a ratio of 75% and 25%, respectively, so that the distribution of the data between classes in both sets is similar. In Figure 2, there are sample images from the archeologic repository.

CLASS LABEL	ARCHAEOLOGICAL PERIOD	TIME RANGE	CLASS LABEL	ARCHAEOLOGICAL PERIOD	TIME RANGE
1	Paleolithic	1,400,000–24,000 BP	9	Iron	1200–586 BCE
2	Epi-Paleolithic	24,000–11,800 BP	10	Persian	586–333 BCE
3	Pre-Pottery Neolithic	8500–5500 BCE	11	Hellenistic	333–63 BCE
4	Pottery Neolithic	5500–5000 BCE	12	Roman	63 BCE-330 CE
5	Chalcolithic	5000–3500 BCE	13	Byzantine	330–636 CE
6	Early Bronze	3500–2200 BCE	14	Early Islamic	636–1099 CE
7	Middle Bronze	2000–1550 BCE	15	Crusader	1099–1260 CE
8	Late Bronze	1550–1200 BCE	16	Late Islamic	1260–1918 CE

Table 1 Archaeological Periods.

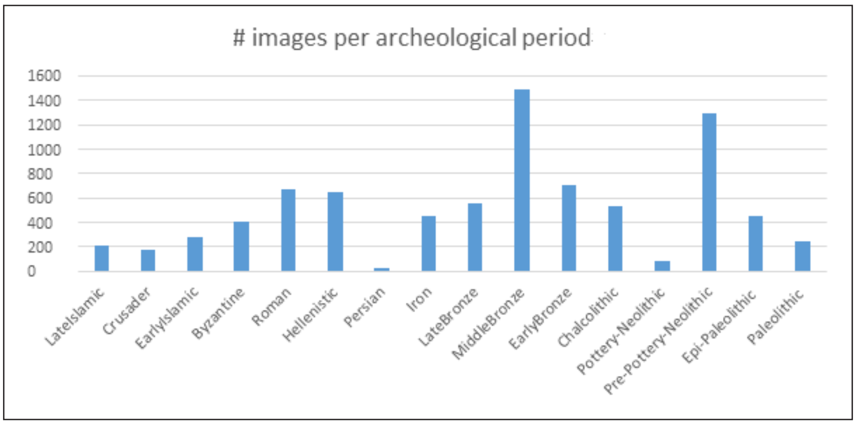

Figure 1 Number of training images in each time category.

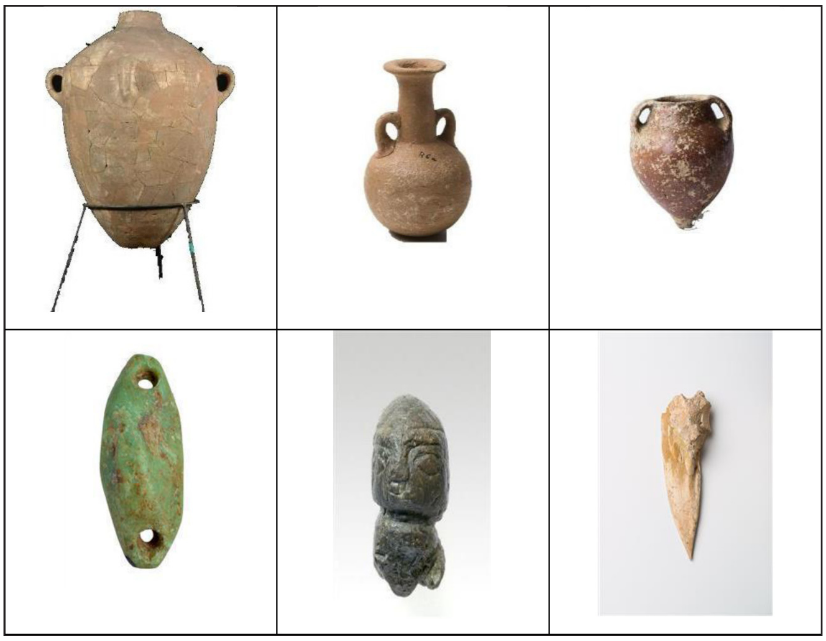

Figure 2 Sample archeological images.

In the next step, three different architectures of CNN were trained on this classified dataset. The architectures are: ResNet50, EfficientB3 and InceptionV3. Each CNN was trained twice with two different loss functions: log and ordinal log. The trained CNNs were applied to the test set to evaluate and compare the performance of the different architectures and loss functions. The Python code implementing the three CNN architectures and applying them to the archaeological artifact image data is available in the ArchImgClassifier repository on GitHub: https://github.com/robilbiu/ArchImgClassifier.

6. RESULTS

Each CNN was trained twice, with two different loss functions. The first one is CE that takes into consideration the likelihood of the sample, which should be classified into its proper class. Any misclassification has the same weight, no matter whether the erroneous class is close in time to the correct class. The second function that was applied to the model training is OCE, which gives larger penalty to misclassification of 'far' categories, in term of archeological age.

The performance of each combination of a model and a loss function was measured by the following metrics mentioned above. Table 2 summarizes the results.

To provide a more detailed analysis of each model's performance across different classes, Figure 3 presents the classification confusion matrix resulted from each trained model. This matrix offers an in-depth view of the model's accuracy in classifying each archaeological period. The correspondence of class labels and archaeological periods as presented in Table 1.

MODEL	ACCURACY	PRECISION	TOP-3	TOP-5	ENTROPY
ResNet50 with CE	0.717	0.721	0.871	0.918	1.709
ResNet50 with OCE	0.689	0.692	0.869	0.928	1.747
InceptionV3 with CE	**0.721**	**0.723**	0.877	0.926	2.969
InceptionV3 with OCE	0.708	0.714	**0.881**	**0.929**	**1.618**
EfficientNetB3 with CE	0.632	0.654	0.817	0.893	1.995
EfficientNetB3 with OCE	0.694	0.696	0.862	0.912	1.738

Table 2 Accuracy, precision, top-3 and top-5 measurements.

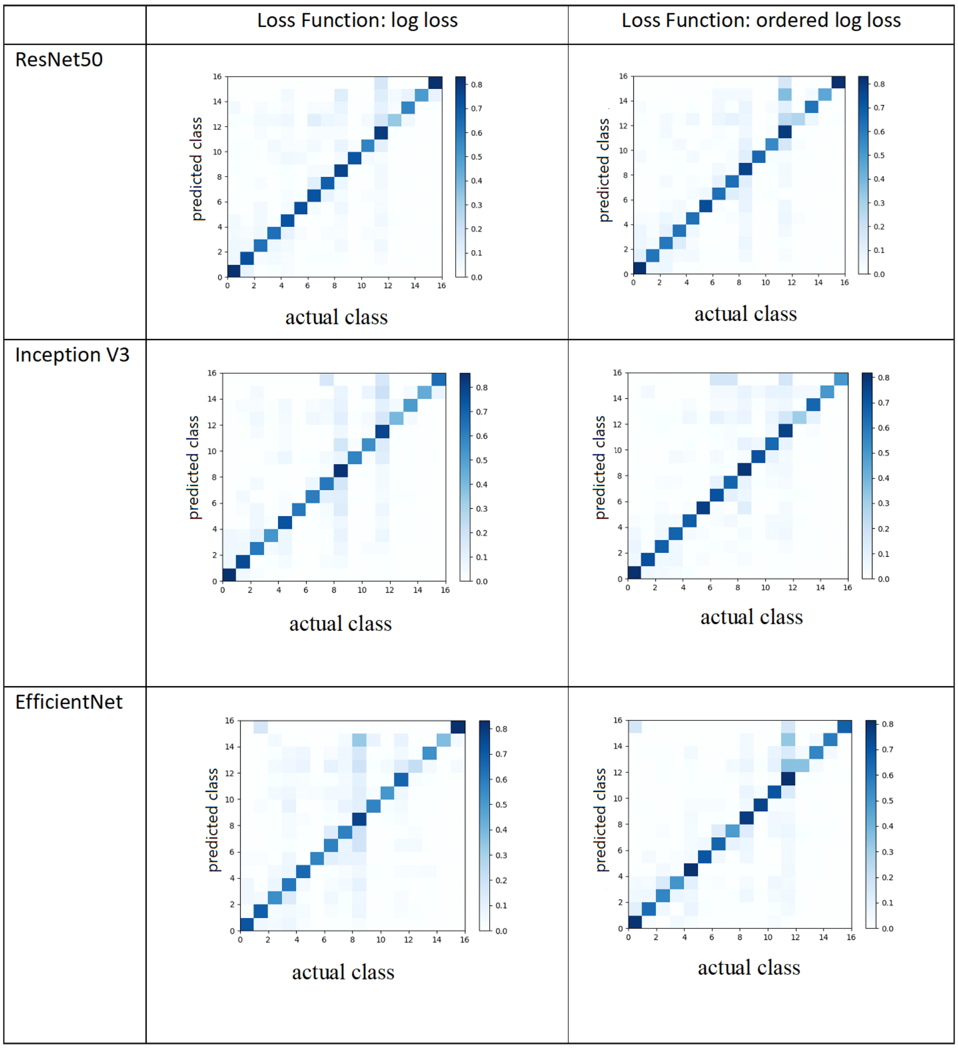

Figure 3 Confusion matrices of the 6 trained models.

Analysis of the confusion matrices reveals differences in classification patterns between the CE and OCE loss functions depending on the CNN architecture. For the EfficientNetB3 and InceptionV3 models, the OCE loss results in less dispersion of predicted ages across the actual age classes compared to the CE loss (for instance, see Iron age, class 9, classification results). This indicates OCE better preserves the ordinal relationships between nearby age ranges. However, ResNet50 displays no significant difference in confusion spread between the CE and OCE losses.

This analysis also indicates artifacts from the Paleolithic and Roman periods were classified with the highest accuracy across all models. In contrast, most misclassified artifacts across all six models were wrongly classified as belonging to the Iron Age and Roman periods. A potential explanation is that artifacts from certain ages may have more distinct stylistic features that enable easier discrimination by the CNN models, while artifacts from periods with more subtle differences in styles are more easily confused.

InceptionV3 with CE loss function achieved the best results in terms of accuracy and precision, while grading as the worst model in terms of entropy. This result indicates that when the model correctly classifies an artifact's age, it tends to predict the exact correct class with high confidence. However, when the models misclassify an artifact's age, the errors are dispersed across multiple incorrect classes rather than being concentrated systematically into one wrong class.

InceptionV3 with OCE loss function achieved the best results in both top-3, top-5 and entropy measures. This means that when the model misclassifies an artifact, it confuses it with certain classes.

7. DISCUSSION AND CONCLUSIONS

In this research, we tackled the complex task of automating archaeological artifact age estimation using pre-trained CNN architectures, addressing challenges like artifact quality, period duration discrepancies, and dating uncertainties. We compared three CNN architectures: EfficientNetB3, ResNet50, and InceptionV3, and found that InceptionV3, typically used for object recognition, outperformed others in classification tasks. This emphasizes the need for empirical testing in architecture selection. Additionally, we evaluated the impact of cross entropy (CE) and ordinal cross entropy (OCE) loss functions, discovering that CE often matched or exceeded OCE's performance, thereby contributing to the broader discourse on loss function effectiveness in machine learning.

Among the various CNN architectures and loss functions compared, the combination of InceptionV3 model paired with the CE loss emerged as the most effective for the age estimation of archaeological artifacts, based on accuracy and precision evaluation metrics. Specifically, the InceptionV3 + CE model attained an accuracy rate of 72.1% and a precision rate of 72.3%.

While the InceptionV3 with CE loss model achieved the highest accuracy, it is notable that precision slightly exceeded accuracy across all six evaluated models with a margin ranging from 0.2% to 2.2%. This indicates some artifacts were misclassified, but models successfully categorized artifacts into the correct age range. The minor accuracy-precision gap highlights room for improvement by modifying the classification threshold, balancing training data, utilizing ensembles, and other techniques to further increase precision without sacrificing accuracy. Still, strong precision demonstrates feasibility of using CNNs to broadly categorize artifacts into age ranges, although refinements are needed for year-level estimation.

For Top-3 and Top-5 metrics, the InceptionV3 model paired with OCE loss performed best, attaining 88.1% Top-3 and 92.9% Top-5 accuracy. Inception V3's multi-scale processing likely enabled flexibility in identifying age-indicative visual patterns for ranking predictions. Additionally, Top-5 exceeded Top-3 accuracy by 4.7–7.6% across models, and both substantially outperformed overall accuracy and precision. This shows CNN reliability for categorizing artifacts into general age groups, if not precisely estimating age in years. Results indicate age estimation within a tolerance would achieve greater success.

A potential explanation for the impact of the loss function on the entropy of the model is that ResNet's residual learning approach may internally represent age in a continuous manner, reducing the benefits of enforcing inter-class ordinal relationships. Overall, choice of loss function significantly impacts some architectures' abilities to discriminate between neighboring age classes without confusion, while others show slight difference. Further investigation into model internals could provide more insight into these architectural dependencies in confusion patterns.

The choice of loss function depended on model architecture and complexity. InceptionV3 benefited more from OCE loss, as their multi-scale processing better captured relative age relationships between classes rewarded by OCE. In contrast, EfficientNetB3 achieved top performance with CE loss, likely because its scaling approach and parameter efficiency provided strong baseline accuracy that CE loss further enhanced by heavily penalizing each age error.

InceptionV3 slightly outperformed ResNet50 and EfficientNetB3, potentially due to Inception's multi-scale processing identifying subtle age differences, its dimensional reduction and concatenation enabling specialized feature extraction, and its balance of depth, width and computational cost. However, performance differences were not highly significant, suggesting process of selecting appropriate architectures and hyperparameters is important for maximizing results, but not crucial for assessing application potential.

The observed algorithm behaviors prompt questions on the influence of dataset characteristics. Follow-on work could investigate whether modifications to the dataset composition and structure, like more balanced chronological distribution or larger image sizes, can reduce these effects and further improve classification accuracy.

In summary, this research advances our understanding of the application of machine learning techniques in computational archaeology. It highlights the importance of careful selection and empirical testing of both the architecture and the loss function in deep learning models. Furthermore, it underscores the potential of these techniques in providing practical tools for archaeologists and opens up avenues for future research in this field. Despite the challenges encountered, our approach and findings provide a promising direction for further exploration and application of deep learning in the field of archaeology.

ACKNOWLEDGEMENTS

We would like to extend our thanks to Dr. Philipe Natalio from the Weizmann Institute and to Avi Resler from Tel-Aviv University for their significant contribution, the inspiration of their work, and the time invested on their part in the completion of this work.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Dr. Sharon Yalov-Handzel orcid.org/0000-0002-6958-4388
Afeka, Tel-Aviv Academic College of Engineering, Israel

Ido Cohen
Afeka, Tel-Aviv Academic College of Engineering, Israel

Yehudit Aperstein orcid.org/0000-0001-6390-9463
Afeka, Tel-Aviv Academic College of Engineering, Israel

REFERENCES

Agam, A, Azuri, I, Pinkas, I, Gopher, A and **Natalio, F.** 2020. Publisher Correction: Estimating temperatures of heated Lower Palaeolithic flint artefacts. *Nature Human Behaviour*, 4(12): 1322–1322. DOI: https://doi.org/10.1038/s41562-020-01017-0

Agrawal, S. 2021. Metrics to Evaluate your Classification Model to take the right decisions.

Casini, L, Roccetti, M, Delnevo, G, Marchetti, N and **Orrù, V.** 2021. The Barrier of meaning in archaeological data science. *arXiv preprint* arXiv:2102.06022.

Chatterjee, R, Chatterjee, A and **Halder, R.** 2021. Impact of deep learning on arts and archaeology: An image classification point of view. In *Proceedings of International Conference on Machine Intelligence and Data Science Applications: MIDAS 2020* (pp. 801–810). Springer Singapore. DOI: https://doi.org/10.1007/978-981-33-4087-9_65

Cifuentes-Alcobendas, G and **Domínguez-Rodrigo, M.** 2019. Deep learning and taphonomy: high accuracy in the classification of cut marks made on fleshed and defleshed bones using convolutional neural networks. *Scientific reports*, 9(1): 18933. DOI: https://doi.org/10.1038/s41598-019-55439-6

Domínguez-Rodrigo, M, Cifuentes-Alcobendas, G, Jiménez-García, B, Abellán, N, Pizarro-Monzo, M, Organista, E and **Baquedano, E.** 2020. Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. *Scientific Reports*, 10(1): 18862. DOI: https://doi.org/10.1038/s41598-020-75994-7

He, K, Zhang, X, Ren, S and **Sun, J.** 2016A. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). DOI: https://doi.org/10.1109/CVPR.2016.90

He, K, Zhang, X, Ren, S and **Sun, J.** 2016B. Identity mappings in deep residual networks. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14* (pp. 630–645). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-46493-0_38

Itkin, B, Wolf, L and **Dershowitz, N.** 2019. Computational ceramicology. *arXiv preprint* arXiv:1911.09960.

Janković Babić, R. 2023. A comparison of methods for image classification of cultural heritage using transfer learning for feature extraction. *Neural Computing and Applications*, 1–11. DOI: https://doi.org/10.1007/s00521-023-08764-x

Lu, Y, Zhou, J, Wang, J, Chen, J, Smith, K, Wilder, C and **Wang, S.** 2018, April. Curve-structure segmentation from depth maps: A cnn-based approach and its application to exploring cultural heritage objects. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1). DOI: https://doi.org/10.1609/aaai.v32i1.12306

Lyons, M. 2021. Ceramic Fabric Classification of Petrographic Thin Sections with Deep Learning. *Journal of Computer Applications in Archaeology*, 4(1). DOI: https://doi.org/10.5334/jcaa.75

Nagda, R. 2019. Evaluating models using the Top N accuracy metrics. *Dostupné z:* https://medium.com/nanonets/evaluating-models-usingthe-top-n-accuracy-metrics-c0355b36f91b.

Parisotto, S, Leone, N, Schönlieb, CB and **Launaro, A.** 2022. Unsupervised clustering of Roman potsherds via

Variational Autoencoders. *Journal of Archaeological Science*, 142: 105598. DOI: https://doi.org/10.1016/j.jas.2022.105598

Pawlowicz, LM and **Downum, CE.** 2021. Applications of deep learning to decorated ceramic typology and classification: A case study using Tusayan White Ware from Northeast Arizona. *Journal of Archaeological Science*, 130: 105375. DOI: https://doi.org/10.1016/j.jas.2021.105375

Reese, KM. 2021. Deep learning artificial neural networks for non-destructive archaeological site dating. *Journal of Archaeological Science*, 132: 105413. DOI: https://doi.org/10.1016/j.jas.2021.105413

Resler, A, Yeshurun, R, Natalio, F and **Giryes, R.** 2021. A deep-learning model for predictive archaeology and archaeological community detection. *Humanities and Social Sciences Communications*, 8(1). DOI: https://doi.org/10.1057/s41599-021-00970-z

Rosebrock, A. 2017. Imagenet: Vggnet, resnet, inception, and xception with keras. *Mars*.

Sakai, M, Lai, Y, Canales, JO, Hayashi, M and **Nomura, K.** 2023. Accelerating the discovery of new Nasca geoglyphs using deep learning. *Journal of Archaeological Science*, 105777. DOI: https://doi.org/10.1016/j.jas.2023.105777

Siozos, P, Hausmann, N, Holst, M and **Anglos, D.** 2021. Application of laser-induced breakdown spectroscopy and neural networks on archaeological human bones for the discrimination of distinct individuals. *Journal of Archaeological Science: Reports*, 35: 102769. DOI: https://doi.org/10.1016/j.jasrep.2020.102769

Szegedy, C, Vanhoucke, V, Ioffe, S, Shlens, J and **Wojna, Z.** 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826). DOI: https://doi.org/10.1109/CVPR.2016.308

Wu, H. 2021. Texture Image Classification Method of Porcelain Fragments Based on Convolutional Neural Network. *Computational Intelligence and Neuroscience, 2021*. DOI: https://doi.org/10.1155/2021/1823930

Xu, J, Guo, J, Zimmer-Dauphinee, J, Liu, Q, Shi, Y, Asad, Z, Wilkes, DM, VanValkenburgh, P, Wernke, SA and **Huo, Y.** 2023. Semi-supervised contrastive learning for remote sensing: identifying ancient urbanization in the south-central Andes. *International Journal of Remote Sensing*, 44(6), 1922–1938. DOI: https://doi.org/10.1080/01431161.2023.2192879

Zhou, J. 2022. *Identifying and Discovering Curve Pattern Designs from Fragments of Pottery* (Doctoral dissertation, University of South Carolina).

http://www.antiquities.org.il/t/default_en.aspx.

https://www.israeliarchaeology.org.

TO CITE THIS ARTICLE:
Yalov-Handzel, S, Cohen, I and Aperstein, Y. 2024. Comparative Analysis of CNN Architectures and Loss Functions on Age Estimation of Archaeological Artifacts. *Journal of Computer Applications in Archaeology,* 7(1): 185–194. DOI: https://doi.org/10.5334/jcaa.136

Submitted: 08 September 2023 **Accepted:** 07 February 2024 **Published:** 06 March 2024

Journal of Computer Applications in Archaeology is a peer-reviewed open access journal published by Ubiquity Press.