

On Artificial Intelligence as Assistive Technology in Archaeological Scholarship

Yixian Wang

Stony Brook School, New York, USA
yixian.wang@sbs.org

Abstract. In archaeological scholarship, the challenge of interpreting large amounts of fragmented evidence can be daunting. This paper will examine how artificial intelligence can assist in this process, mainly through examining visual evidence and classifying material evidence. Drawing from Trevor Paglen's argument of machine vision and Matteo Pasquinelli's history of classification in machines, this paper develops an analysis of two archaeological case studies: archaeological site detection based on satellite imagery in Mesopotamia and deep-learning classification of Tusayan White Ware pottery. The findings of the studies demonstrate that artificial intelligence is capable of expanding preliminary surveys as well as applying typologies consistently. At the same time, they also highlight the importance of human input, coming to the conclusion that artificial intelligence's potential and principal contribution lies in simplifying the archaeological process and directing human attention towards evidence requiring further interpretation, establishing a framework for understanding artificial intelligence as an assistive technology.

Keywords: Artificial Intelligence, Archaeological Scholarship, Human-AI Collaboration

1. Introduction

The rise of artificial intelligence has caused machine learning to become one of the most prominent topics of discussion in recent years. Public discourse surrounding this topic often tends to be prudent, focusing on criticism and controversies surrounding artificial intelligence. While this viewpoint holds value, it also neglects the usefulness of artificial intelligence, particularly in fields of scholarship such as archaeology. Archaeological scholarship depends on large amounts of evidence that is commonly fragmented, such as aerial imagery, field records, digital reconstructions and architectural remains. While these pieces of evidence would traditionally require large amounts of effort by human archaeologists to piece together, artificial intelligence offers new possibilities for archaeological scholarship by simplifying the process before archaeologists can commence the complicated work of interpreting history from material evidence.

The value of artificial intelligence in archaeology is best understood as the ability to locate archaeological sites and classify evidence. Trevor Paglen's *How to See Like a Machine* demonstrates that images can be used by machines as data to derive information from, allowing it to make decisions based on given images [1]. A computer vision system could be used to analyze satellite imagery, discover trends within it and be used to search for archaeological sites. Matteo

Pasquinelli's *The Eye of the Master* discusses artificial intelligence's background in automation, illustrating artificial intelligence's capabilities of classifying and analyzing archaeological evidence [2]. Hence, there are two identified ways for artificial intelligence to assist in archaeology artificial intelligence can improve archaeological research through assisting in the detection of archaeological sites and the consistent classification of archaeological material.

2. Literature review

Research on the potential of artificial intelligence in archaeology raises questions concerning the impact of automation on archaeological scholarship. Dylan S. Davis argues that by requiring explicit definitions and repeatable procedures, machine automation and the development of an artificial intelligence model can expose underlying assumptions within the archaeological field that would otherwise remain implicit in manual analysis [3]. This argument provides a useful starting point for analysis of artificial intelligence as an assistive technology.

Paglen's *How to See like a Machine* offers another theoretical perspective for artificial intelligence in archaeology by going over the changing function of images. His emphasis and studies on the usage of computational images directs attention towards their possible participation within systems of classification. Applied to archaeology, this argument suggests that types of information such as satellite imagery and regular archaeological field photographs can be used as inputs for artificial intelligence models that could assist in the archaeological process [1].

Pasquinelli's *The Eye of the Master* provides complementary arguments supporting the aforementioned approach. By investigating the long history of labor and the automation of labor that artificial intelligence stems from, Pasquinelli argues that machine intelligence relies on and incorporates forms of knowledge created through social activity. When read alongside archaeological processes, this account shows that artificial intelligence might be capable of and suited for assisting in work involved with the construction of surveys, organization of collections as well as defining typologies. While Paglen provides a framework for explaining how artificial intelligence uses computational images, Pasquinelli provides a basis for understanding the functions behind machine learning and how scholarly knowledge can be re-organized [2].

Recent studies demonstrate the potential of said systems when applied beyond theory. Orengo et al. trained a machine-learning classifier using radar and multispectral satellite imagery on known mounds to detect archaeological mounds in Pakistan's Cholistan desert, illustrating how automated analysis is capable of expanding the spatial scale of archaeological research [4] Casini et al. developed a workflow that uses semantic segmentation and satellite imagery to detect Mesopotamian tells [5]. Reported accuracies of predictions rose to as high as 80 percent during the final phases of testing. During part of a separate test in the Maysan province, the model was able to retrieve 17/20 known archaeological sites in the imagery it was given. Their results support collaborative screening as the presented machine-learning model was able to achieve relatively high accuracies only after experts clarified problems in the original methodology, highlighting the importance of performance measures on machine judgements about visibility, annotation quality and useful predictions [5].

Artifact classification presents a parallel at a different scale. Pawlowicz and Downum trained a convolutional neural network to learn to classify Tusayan White Ware shards. Their results show that the model was able to achieve approximately 82.5 percent agreement with consensus labels of a group of 2,407 shards. The study also introduces the usage of heat maps to visualise influential image regions, as well as a similarity search tool for identifying and comparing fragments that share common traits. These results suggest that artificial intelligence is capable of converting

archaeological labor into a repeatable procedure, supporting both routine classification and comparison [6].

Considered together, the literature builds a basic framework for understanding a relationship between machine learning, computational procedures and archaeological judgement. Paglen and Pasquinelli provide theoretical concepts for analyzing artificial intelligence's capabilities, while the technical studies demonstrate bounded capabilities. This paper brings together these perspectives through a parallel analysis of site detection and ceramic classification. The paper's contribution is in explaining how artificial intelligence can increase the scale, speed and consistency of preliminary analysis in archaeology while preserving room for human interpretation.

3. Analysis

Archaeological scholarship has traditionally depended on human vision. During the early stages of archaeology, human scholars are responsible for noticing differences in soil and distinguishing signs of human activity from the surrounding background. To identify possible excavation sites surveys of large amounts of land must be done and inspected carefully, requiring massive amounts of attention. A slight elevation, an unusual concentration of material or a change in vegetation could all be signs of undiscovered ruins. During the twentieth century, advances in technology allowed for aerial photography, which extended upon what could be done on the ground. Satellite imagery and other forms of remote sensing technologies allow scholars to capture and examine landscapes that cover thousands of square kilometers at a much more efficient pace than going personally on the ground, analyzing traces of human activity that remain after buildings disappear. Nonetheless, gaining access to more imaging does not necessarily equate to easier research - every photograph must still be examined in detail and compared with context such as regional history and environmental factors. A large visual archive therefore produces the new difficulty of creating more material than a human researcher could effectively inspect.

3.1. Site detection

This is where machine vision and Paglen's argument comes into play. Traditionally, the analyzing and understanding of a photograph is dependent on what the human eye can see and interpret. However, with machine-read photographs, the machine systematically measures, classifies, predicts and acts without having to wait for a human viewer to look at the image. For example, a facial-recognition camera does not need to appreciate the view it captures, as it converts visible features into numerical relationships, which it then compares with patterns stored in a dataset. The same principle can be applied to an archaeological context. A computer does not need to understand the history of, say, a settlement in Mesopotamia in order to identify a Mesopotamian mound. Instead, it detects combinations of colors, edges, textures and shapes that resemble examples included in its training data. What appears as a full landscape that would require careful attention for a human scholar becomes, for the machine, a field of measurable relationships [1].

This difference, however, does not mean that machine vision is a superior form of human sight. The two forms of vision are simply meant for performing different tasks. An archaeologist can make the connection between a mound and a nearby canal, to earlier maps and patterns of settlement based on their knowledge of regional history. An artificial intelligence model lacks this complete context and specialized training, but it can apply one narrow detection procedure to thousands of images without ever becoming tired or changing its standard halfway through the search like a human might. This limited (currently) ability can be extremely useful, as the first step in an

archaeological study is often determining points of interest that deserve closer investigation. Once this is established then archaeologists can move on to the process of explaining a site, establishing its chronology and deciding its historical importance. Machine vision serves as a filter between a large visual archive of data and the archaeologist, reducing an enormous landscape to a much smaller set of locations that deserve attention. Its strength lies in the scale, consistency and effectiveness of its searching capabilities.

The southern Mesopotamian floodplain provides quite the useful setting for testing this possible usage of artificial intelligence. The region is central to the study of ancient civilizations, cities, irrigation and long term settlement, yet it is quite difficult to survey consistently. Many remaining traces of ancient settlements appear in the form of "tells", mounds created through the accumulation of debris from repeated construction and deconstruction. While some are visually distinct hills, others rise only slightly above the surrounding landscape and can be obscured and damaged by irrigated agriculture, flooding and urban development. Earlier research done by archaeologists such as Robert McCormick Adams Jr. used aerial photographs and systematic ground survey to record settlements and ancient canals across the region [7]. The results provided a foundation for interpreting, in the select area, the relationship between settlement and water management in ancient Mesopotamian settlements, but reproducing the work across the entire floodplain through manual image inspection would take unreasonable amounts of time. Thus, Luca Casini and his collaborators set up an experimental system in an attempt to make the preliminary survey more manageable [5].

Casini et al. began with accumulated knowledge from the University of Bologna's Floodplains Project, which had assembled a georeferenced outline from published surveys across approximately 66,000 square kilometers of land in the floodplains [8]. This dataset contained 4,934 shapes representing archaeological sites that were confirmed through ground inspection and the study of surface artifact scatters [5]. The outlines provided what is called ground truth in machine learning: the dataset that a model derives its information and standards from, and against which predictions are evaluated.

However, the original survey records could not be directly transferred to a neural network, as they were purposed to document archaeology, not to teach a machine a specific scholarly visual subcategory. Casini et al. removed around two hundred of the largest shapes because they included more than just a singular mound, and the size and complexity of any single one would be nearly enough to fill up a training image on its own [5]. They also removed 684 records that were either too small to reliably represent a tell or were marked by archaeologists as destroyed. These exclusions determined which parts of the archaeological record would define the model's idea of a mound. The team then added 1,155 counterexamples that included no target site, such as highly urbanized districts, intensive agricultural zones, flooded areas and rocky hills so that the model would see landscapes that might otherwise resemble archaeological features. By the end of the preparation phase, the collection included 5,025 images.

The model uses semantic segmentation, a method that uses the classification of individual pixels rather than assigning a label to an entire image. In this experiment, the model would classify each pixel to one of only two categories: part of a tell or background. During training, the artificial intelligence model would examine a satellite image and attempt to reproduce the associated archaeological mask by proposing an outline constructed from pixels that had a high probability to be assigned as part of a mound. Semantic segmentation allowed the model to produce outlines that could be used as visual references or displayed in geographic softwares for archaeologists to analyze [5].

Casini et al. also tried to prevent the model from picking up random patterns during its training process. As such, they randomly cropped, rotated and adjusted the brightness of the photographs, meaning that sites would not always appear in the photographs under the same condition. They also found that a larger field of vision improved the model's performance as the surrounding environment provided the model with more context for distinguishing a mound from an isolated anomaly such as a patch of color. Images containing sites were combined with the counterexamples and then were divided into training, validation and testing sets. The testing, or holdout, images were not used to teach the model, allowing its prediction to be validated against images that it had not encountered during training, testing whether the system had learned a transferable visual pattern rather than simply memorizing its inputs [5].

The best model was initially reported with a detection accuracy of approximately 60 to 63 percent. Upon human review of issues causing the large inaccuracy, it was revealed that some apparent mistakes were the result of problems in the archaeological annotations that were provided rather than failures of the model. Casini et al. found that several sites that were not excluded were no longer visible in modern satellite imagery, so a model could not detect them based on its input yet they were present in the archaeological records. In other cases, the model had highlighted a visible mound near the vicinity of the annotation or produced a low-probability outline that the model itself was unsure of. After reviewing and making adjustments to avoid the aforementioned issues, accuracy rose to roughly 80 percent [5].

The most revealing test took place in a 104-square kilometer part of the Maysan province in southeastern Iraq that was not included in the surveyed dataset. Twenty alleged sites were first identified via remote sensing, and the artificial intelligence model was able to retrieve 17 out of the 20 locations. The model also produced roughly twenty additional shapes, most of which were false positives. Despite having reached quite a high level of accuracy, it is still far from consistent, producing nearly as much false information as useful ones. However, most of the false positives produced by the model could be quickly rejected by an expert due to how their size, location or surrounding environment made little archaeological sense. The model was able to save massive amounts of labor by performing a broad search and narrowing down sites of interest that deserve further attention, allowing archaeologists to spend more effort on in-depth analysis [5].

The model's predictions could also be applied in a less rigid form. Instead of converting every result above a specific threshold that deemed whether a result was probable into a fixed outline, Casini et al. also constructed a heat map based on probability values. Hotter areas on the heat map indicated where the model found stronger visual resemblance to known mounds, giving an overall prediction of areas of high recurrence. This gave archaeologists control over the level of uncertainty produced by the model, as the archaeologist has the freedom to compare this heat map with contextual evidence such as historical aerial photographs and topographic maps. Paglen's argument is particularly strong at this point, as this experiment demonstrates that the image no longer operates as just a visual representation to be viewed - it becomes machine-readable data that can be used to produce a probability layer that reorganizes where and how a person looks [5].

At this point in the research, the distinction between detection and discovery is important to understand. If one were to call every predicted outline a "newly discovered site" then the artificial intelligence model would be granted more authority than necessary. The model's prediction simply marks a statistical resemblance that can act as a research lead, an early detection of a site. Discovery, on the other hand, requires deeper inquiry that goes beyond simply finding a possible point of interest. It is unethical and unsafe to credit the model for "discoveries" when it is not actually conducting the process of discovery: comparison with historical materials, consideration of

landscape formation, inspection of surface material and conducting field surveys. However, the model is valuable precisely because of its ability to operate before these costly steps, helping speed up and reduce the costs of the overall archaeological process. In this sense, machine vision changes the order of archaeological attention, as archaeologists can now begin with machine generated candidates and then widen or narrow the search based on contextual evidence. Artificial intelligence models do not replace discovery; they make the search leading to discovery easier and allow for potentially more extensive discovery.

At the same time, the process that an artificial intelligence model undergoes highlights that machine vision remains dependent on human archaeological judgement. Specialists helped assemble the dataset that the model relied on, and also participated in deciding how a prediction should be evaluated. Human expertise was also required to interpret the model's output in a realistic situation - a high-probability contour could just be a modern disturbance whilst a low-probability area could still be of significance when compared with older imagery. Each round of expert review could produce corrected outlines that would improve a later version of the model, making the research process an exchange between accumulated human knowledge and machine prediction [5].

The model's failure to smoothly transfer to other locations further highlights this system's dependence on context and human input. Casini et al. attempted to apply the Mesopotamian model to mound sites in Samarkand, Uzbekistan. Although the target features were generally similar to those in Mesopotamia, the model was only able to accurately detect 25 to 30 percent of them. The region contained vegetation and patterns of urbanization that differed from Mesopotamia, and the methodology used to annotate the Uzbek surveys also differed slightly from those of the Mesopotamian surveys. A model trained to recognize tells in one specific environment could not be moved to another and still be expected to perform on the same level [5]. This limitation only represents the present state of artificial intelligence technology.

This limitation clarifies the first claim rather than defeating it. Artificial intelligence improves the process of site detection when it is used as a method of searching through masses of visual evidence. The model can process large collections of images, outline statistical similarities and then list possible locations for review at a more efficient pace compared to humans. Casini et al. produced a measurable expansion of this preliminary vision as the artificial intelligence model was able to retrieve sites with around 80 percent accuracy in the Maysan test. However, it did not establish the historical statistics and cultural importance of the sites, which requires human expertise to explain.

3.2. Pottery classification

Locating an archaeological site is just the beginning of research. Once archaeologists survey and excavate a location, they must record and classify the material recovered from the site. A single excavation could produce thousands of pottery fragments, bones, tools, pieces of glass and many other types of remains. When considered individually, each piece may not reveal much, but when putting into consideration their distributions and patterns, they can reveal trade relationships, technological advances and parts of daily life. Classification turns fragmented material into evidence that can be used to explain the past. However, the process of classification is arduous and time consuming, and demands a high level of expertise. Specialists may have to spend years learning to distinguish small differences in material, shape and manufacture, and even the most experienced researchers cannot always accurately classify an object to the correct category. The rise of artificial intelligence, however, means that a more efficient alternative may be available to archaeological scholars.

Pasquinelli's discussion on the history of artificial intelligence provides a framework for understanding how evidence classification can be automated. Pasquinelli challenges the conventional idea that artificial intelligence is attempting to recreate the human brain in a machine. He instead connects artificial intelligence to a history in which machines absorb and reorganize forms of knowledge people use in their work. Calculating engines have existed for over three centuries, and despite them being a very different technology, both calculating engines and artificial intelligence rely on making tasks measurable and manageable. Automation begins when a portion of labor can be made into a procedure rather than requiring complete human understanding. Pasquinelli's argument directs attention from the idea of an independent machine to the social knowledge built into a system that makes it function. In archaeology, classification of evidence fits into this pattern of the organization of knowledge as generations of scholars have already constructed typologies, categorized information and undergone the labor of comparison and analysis [2].

Before proceeding, it is important to understand the meaning of an archaeological typology. It is a system through which objects that might have had no individual meaning become comparable. Pottery, for instance, can be divided into categories according to their material, temper, surface treatment, shape, design or method of production. When a style is associated with a known period, the classification of a fragment could help archaeologists understand the age and date of the deposit it was found in. When pottery of the same category is found in distant sites, it can contribute to research about exchange, movement or shared cultural practices. However, these categories often create uncertainty, especially when an artifact possesses traits that can be associated with multiple different styles. As such, much of the routine work following the establishment of a typology consists of comparing new fragments with previously established examples.

A convolutional neural network, or a CNN, could possibly automate this part of the process. The model works by learning from images that possess labels. It will repeatedly predict labels, measuring the difference between its prediction and the assigned answer each time, and then adjusts its internal numerical values and relationships so that similar errors become less likely to happen. Theoretically, the model should be able to learn the connections between specific visual features and the corresponding categories, and apply that knowledge in categorizing new pieces of evidence.

Leszek M. Pawlowicz and Christian E. Downum explored this possibility by testing whether a convolutional neural network could classify shards of Tusayan White Ware, a type of painted pottery from northeastern Arizona. The pottery was produced by communities of the Kayenta branch of the ancestral Pueblo culture and dates from between approximately 825 to 1300 CE. Archaeologists studying this type of pottery have divided it into subcategories including Kana'a, Black Mesa, Sosi, Flagstaff, Dogoszhi, Kayenta and Tusayan. Each type is associated with a different date range and decorative features. For instance, Kana'a may include fine lines and small ticking marks while Black Mesa is characterized by its usage of heavier geometric elements. These distinctions can help archaeologists date and compare sites, but the distinctions are very subtle when applied to individual shards of pottery. A sherd, or a pottery shard, may only preserve a small portion of its original design, and many types are stylistically similar making the identification process painstakingly difficult [6].

The researchers began with a group of 3,067 photographs of Tusayan White Ware sherds that were clear enough and contained enough detail for classification. Four specialists with at least 30 years of experience had independently examined each photograph, and they could choose to mark a photograph as either part of an established type or indeterminate, and they also had the option to select an intermediate category if a sherd contained combined characteristics. The results showed

that even experts could not avoid disagreements, as the pairwise agreement on the full collection ranged from 52.4 to 62.6 percent. This result does not mean that the analysts lacked skill, but rather that some sherds were genuinely ambiguous. This ambiguity meant that several different answers could be defensible, showing that conventional classification requires more judgement and variation than a final chart may suggest [6].

To construct a usable training set, Pawlowicz and Downum identified images that had sufficient consensus from the human experts, reducing the 3,067 photographs from the original set down to 2,407. These remaining photos then each received a consensus label that would function as the answer that both the human experts and the convolutional neural network would be measured against. In this selected dataset, the four experts matched the consensus labels at different rates, ranging from the lowest at 73.6 percent accuracy to the highest at 86.9 percent accuracy, while pairwise agreement between the researchers remained between 60.4 and 72.7 percent. The consensus set itself was full of uncertainties, as although more than one expert supported every retained label, the labels themselves still came from combining human judgements. Therefore, the judgements in this experiment were made based on this defensible culmination of human thought rather than an archaeological truth, as it is quite difficult to tell at this point [6].

This distinction is quite important in order to understand what was regarded as a correct prediction later in the experiment. The CNN was expected to be able to make a prediction based on the categories defined by archaeologists, and when it matched the consensus label its prediction could be counted as correct. In the meantime, the researchers split the 2,407 photographs into five separate groups for teaching the model. Four groups were used for training and the fifth group was withheld for testing, a process which was repeated until every group had served as unseen test material. This process allowed for each individual photograph to be evaluated as if the model had never seen it before, mimicking what field testing would likely result in [6].

The team trained modified versions of two established models, VGG16 and ResNet50, combining their outputs. During training, the researchers rotated the sherd images at random angles, varied the image scales and converted all images to greyscale so that differences in lighting, angle, weathering and white balance would not become misleading clues for the model, making the experiment more consistent [6].

The final combined results matched the consensus label for 1,986 of the 2,407 sherds, resulting in about 82.5 percent accuracy across the board. The model's performance fell within the range of the four experts, being more accurate than two experts, comparable to the third, and falling short of the best expert with 86.9 percent agreement with the consensus labels. This result demonstrates that an artificial intelligence model could reproduce a typological system at expert-level accuracy, supporting the claim that artificial intelligence is capable of acting as an assistive technology in artifact classification [6].

The consistency of the model provided another advantage beyond the accuracy score. A trained convolutional neural network will always produce the same answer if it is given the same image under the same conditions. What this means is that the model will never change its judgement over time unless told to do so, while the same cannot be said for humans. Human analysts tend to change their judgements as they gain experience, become tired or experience other factors that might influence human decision. While these changes in judgement may sometimes reflect valuable learning, they may also increase the difficulty of comparing separate collections due to possible inconsistencies caused by the aforementioned. An artificial intelligence model, on the other hand, offers a repeatable procedure across the entire collection. In addition, most of the errors made by the model in the experiment were caused by types that were stylistically close. For instance, Black Mesa

fragments were often confused with Sosi fragments due to similar layouts in their design. This pattern also shows that the model was able to learn meaningful features even when it fails to identify the exact consensus label [6].

Pawlowicz and Downum also made attempts at making the model's decisions more predictable and interpretable. Convolutional neural networks are often referred to by experts as "black boxes" because of their tendency to produce a prediction without explaining the reasons behind the prediction. To resolve this, Pawlowicz and Downum created activation heat maps showing which parts of the photograph contributed the most to the model's prediction process and influenced the outcome the most. If highlighted areas of the map mostly followed diagnostic design elements, it would show that the model was actually attending to relevant evidence rather than unrelated parts of the photograph. These visualisations offered a way to inspect how the model analyzed the relationship between consensus labels and fragment decorations [6].

The convolutional neural network was also able to accomplish tasks beyond categorizing evidence. Pawlowicz and Downum additionally developed an image matching tool called the "sherdc engine," which was capable of comparing a given fragment with thousands of visual images and retrieving the most similar examples for human review. This method could be used by researchers to identify pieces of the same vessel, compare different designs and draw relationships across collections [6].

This experiment gives Pasquinelli's theory a concrete archaeological form. The model's performance was dependent on nearly a century of human labor in the form of archaeological scholarship. Archaeologists defined Tusayan White Ware types, museum and field labor was involved in the creation and preservation of the photographs used, and four specialists supplied the judgements upon which the training consensus dataset was built. Even the disagreements amongst the experts were part of the system because those disagreements helped define the certainty of the consensus labels. The convolutional neural network compressed this accumulation of human knowledge and labor into a repeatable classification procedure, and as such its intelligence is inseparable from the human labor organized inside the dataset [2, 6].

This relationship between human and machine also explains why the system is not universally applicable. Tusayan White Ware is a strong test case for the model because of the pottery's dark painted designs which contrast strongly against a light background, making it much easier for the model to interpret. The same method used in this experiment is highly unlikely to be applicable for cases regarding undecorated plainware, for instance. Human input is also required in forming the original typology to which a model is trained to recognize. The model trained on Tusayan White Ware should not be expected to be able to identify unfamiliar pottery traditions, as its competence and knowledge was bounded by the labels and visual examples that it learned. A pixel input limit of 224 by 224 also limited the fine details available for analysis, and photographs often omit tactile and contextual evidence that may matter, such as texture and temper [6].

Furthermore, the consistency of the model should not be confused with complete objectivity. While the model is capable of repeated classification at the same standards, the procedure itself is a reflection of human judgement. If the typology contains uncertain boundaries, the model is only capable of reproducing those boundaries without being able to resolve the underlying disagreement. However, this limitation is also useful in a way, as it could help experts identify fragments that challenge existing definitions. Specialists can then proceed from this to determine whether an object is transitional between multiple categories or whether a typology is poorly defined. As such, artificial intelligence can assist in processing routine cases whilst directing human attention towards areas where the classification system is least secure [6].

Another benefit of trained artificial intelligence models comes in the form of the redistribution of archaeological time. Large excavations and museum collections often require extensive amounts of logistical labor in the forms of sorting and counting before evidence can be interpreted. Trained models such as the ones used in the Tusayan White Ware study could handle substantial portions of excavated sherds, leaving archaeologists with more time to investigate issues such as why a specific distribution exists, whether a typology remains useful or what certain pieces of evidence show about a site's chronology. Pawlowicz and Downum also describe continued roles for archaeologists in the management of the model, such as evaluating uncertainty and improving labels, helping accelerate comparison in an established archaeological framework. This study further demonstrates the feasibility of artificial intelligence acting as assistive technology in the field of archaeological scholarship [6].

4. Conclusion

Paglen and Pasquinelli provide two interconnected ways of understanding artificial intelligence. Paglen shows that an image, once converted to machine readable data, can be used to generate predictions that can reorganize human attention. Pasquinelli explains how the operation that resulted in the prediction depends on labor and knowledge that are already expressed through datasets. Together, their arguments describe artificial intelligence as a technology that can extend human abilities in fields such as archaeology [1, 2].

The Mesopotamian mound study and Tusayan White Ware study both demonstrate this argument in practice. In the Mesopotamian mound study, the model was capable of achieving approximately 80 percent site detection accuracy through machine vision, while in the Tusayan White Ware study the convolutional neural network was able to achieve the same level of accuracy in its categorization. Both are capable of expanding the scale at which archaeological scholarship can be organized [5, 6].

Artificial intelligence can therefore act as an assistive technology in archaeological scholarship through two clear contributions. It can analyze visual information to identify possible sites across vast landscapes, and can classify material evidence at a speed and consistency difficult to achieve through human analysis alone. The works of Trevor Paglen and Matteo Pasquinelli establish the theoretical frameworks for the aforementioned contributions, while the Mesopotamian mound and Tusayan White Ware studies show that these frameworks can be realistically applied to archaeological studies.

References

- [1] Paglen, T. (2026). *How to see like a machine: Images after AI*. London: Verso.
- [2] Pasquinelli, M. (2023). *The eye of the master: A social history of artificial intelligence*. London: Verso.
- [3] Davis, D. S. (2020). Defining what we study: The contribution of machine automation in archaeological research. *Digital Applications in Archaeology and Cultural Heritage*, 18, e00152.
- [4] Orengo, H. A., Conesa, F. C., Garcia-Molsosa, A., Lobo, A., Green, A. S., Madella, M., & Petrie, C. A. (2020). Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *Proceedings of the National Academy of Sciences*, 117(31), 18240–18250.
- [5] Casini, L., Marchetti, N., Montanucci, A., Orrù, V., & Roccetti, M. (2023). A human–AI collaboration workflow for archaeological sites detection. *Scientific Reports*, 13, 8699.
- [6] 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. *Journal of Archaeological Science*, 130, 105375.

- [7] Adams, R. McC. (1981). *Heartland of cities: Surveys of ancient settlement and land use on the central floodplain of the Euphrates*. Chicago: University of Chicago Press.
- [8] University of Bologna. (n.d.). FloodPlains. *OrientLab*. Retrieved August 12, 2026, from <https://floodplains.orientlab.net/>