Bachelor Thesis
Artificial Intelligence In The Field Of Archaeology

Philip Gonzalez
University of Gothenburg
The Institution for Historical Studies
Archaeological Bachelor Thesis
Mentor: Christian Horn
Autumn 2023

UNIVERSITY OF GOTHENBURG

I would like to thank Dr Christian Horn and Ashely Green for guiding me through this period. Thank you both for helping me with the technical and archaeological parts of my thesis and pointing me in the right direction.

Abstract

Artificial intelligence is something the world has only just been starting to get a grasp on. The race for who can create the best model is taking place all around the globe. In archaeology, the usage of programs like these is tested every day and through many trials, usable solutions are slowly being born. Artificial intelligence is something the public knows mostly from movies, however, the reality is different from what is often shown. The things within our grasp are currently very different from the ones in fiction. In reality, we have only just begun to excavate the possibilities of what it can do. The purpose of this paper is to highlight what we currently can accomplish with artificial intelligence in both bone mark identification and satellite imagery. The idea is also to ascertain what the current limitations of artificial intelligence are in archaeology and how they might be improved.

1. Introduction

Artificial intelligence has in the last decade emerged as a transformative force in archaeology. Searching for archaeological sites and analyzing ancient bones manually will soon be a thing of the past (Orengo et al. 2020 ; Domínguez-Rodrigo et al. 2020). The algorithms used are particularly suited toward rapid identification and classification of archaeological features and objects. As researchers further the capabilities of artificial intelligence through studies and development we get closer to a more automated future. However, artificial intelligence requires large data sets and a lot of training to function properly. Small data sets introduce problems like potential bias and inflated accuracy numbers (Beck, Jones 1989: 245; Domínguez-Rodrigo et al. 2020: 1). The deep learning models mentioned in this paper show some of these problems. The problem still stands, artificial intelligence can only be as intelligent as its creators and users.

Noteworthy successes underscore the potential of artificial intelligence in archaeology. Researchers have achieved a remarkable 92% accuracy in identifying and classifying marks on ancient bones (Domínguez-Rodrigo et al. 2020: 1) , demonstrating the accuracy and efficiency that artificial intelligence brings to the field. Moreover, the fusion of artificial intelligence with satellite technology has enabled the discovery of previously unknown archaeological sites, further expanding the scope of exploration and knowledge (Orengo et al. 2020: 1).

However, the integration of artificial intelligence in archaeology is not without ethical considerations. Trust in information generated by artificial intelligence remains a complex issue, with concerns about algorithmic bias and the influence of developers' political perspectives on the outcomes. The reliability of artificial intelligence results becomes a critical point of discussion, highlighting the need for vigilance and ethical frameworks as these technologies continue to evolve. Despite the revolutionary impact of artificial intelligence on archaeology, human limitations persist. The effectiveness of artificial intelligence systems is contingent upon the knowledge and dedication of developers, who shape and refine these tools. While artificial intelligence accelerates processes and expands possibilities, it remains a tool created and controlled by humans, subject to the biases and constraints of its creators.

This paper strives to illuminate the current state of artificial intelligence in archaeology, celebrating its achievements while acknowledging the hurdles that accompany this transformative technology. With a focus on the interplay between human ingenuity and technological advancement, it seeks to provide a comprehensive exploration of the evolving landscape where artificial intelligence and archaeology converge.

1.1 Purpose And Questions

The purpose of this thesis is to highlight the great ways that artificial intelligence is used In archaeology. The examples I will discuss are image recognition on bone markings and a similar system used on satellites to find archaeological sites of interest. I also aim to shine a light on the hurdles that we have yet to overcome and the areas where artificial intelligence struggles.

The questions I am asking are as follows:

- How has image recognition been used in archaeology based on two case studies?

- What are the main challenges and limitations of using Artificial intelligence in archaeological research?

I intend to answer these questions through the use of other scientific papers and their results. These questions are what I see as the most enlightening when talking about artificial intelligence in the archaeological field.

1.2 Limitations

As I set out to answer these two formulated questions about the use of artificial intelligence in archaeology some limitations need to be set on the research so it does not deviate from the topic. The sources used must come from those who practice archaeology and have the education. To get a better grasp on how artificial intelligence has changed archaeology I will also give a brief history about how the use of the computer in the field has changed from its inception until today. I will however not give a full length history, only the necessary parts to better emphasize the revolution that we now are a part of.

There are many kinds of different types of artificial intelligence, but this thesis will focus on the ones related to my questions, which are mainly centered around image recognition software. There are also other types that are of interest however as they do not fit into my questions they will not be discussed in this particular paper.

1.3 AI

AI or artificial intelligence can be defined as “the study and design of intelligent agents” where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. It is a computer program that is made to behave intelligently and make decisions on its own based on what information it has been trained on. Image recognition is a branch of AI that enables computers to interpret and understand visual information from images or videos. It involves the use of algorithms and deep learning models to analyze patterns, features, and structures within the visual data, allowing computers to identify and categorize what they see in the given media (Lui et al. 2017: 4).

1.4 Keywords and their explanation used in this paper:

Kernel - A kernel refers to a small matrix used for various operations such as filtering and convolution. Kernels are crucial in feature extraction for tasks such as object detection and recognition in computer vision.
Tensor- A tensor is a mathematical representation of an image, often used to represent pixel values.
Convolutional layer - A convolutional layer is a fundamental building block of convolutional neural networks. The main job of this layer is the convolution operation. It involves sliding a small window, called a kernel, over the input data. The kernel extracts patterns or features from the chosen data.
Overfitting - Overfitting is when the trained model learns the training data too well and fails to adapt to new data.
Max pooling - Max pooling layers that are effective in capturing the most significant features within a local region and discarding less important information
Stride - The stride is the number of pixels by which the filter is shifted during each step.
Feature map - In the context of image recognition, a feature map can be thought of as a spatial map of features that the network has learned to detect.
Data augmentation - Data augmentation is a technique used in computer vision to artificially increase the size and/or variation of a training dataset by applying various transformations to the existing data.
Transfer learning - Transfer learning consists of using a model to a specific problem that was trained for a different problem
Panchromatic - A panchromatic image refers to an image that is sensitive to and records light across the entire visible spectrum, typically from blue to red. Even though they record all visible forms of light the images are grayscale. This together with the recorded light gives higher clarity.
Spatial resolution - Spatial resolution is a measure of the smallest object that can be seen by the sensor of the camera.
Downsampling - By decreasing the sample rate or resolution of something, downsampling allows for reduced data storage requirements. This is beneficial when dealing with large datasets.
Upsampling - Upsampling is when you are increasing the spatial resolution while keeping the 2D representation of an image. It is typically used for zooming in on a small region of an image.
Cross validation - By dividing the dataset into two parts, one for training and other for testing, you can then train the model on the train set and validate the results on the model set. It is meant to prevent overfitting.
Rectified Linear Unit - The Rectified Linear Unit is the most commonly used activation function in deep learning models. The function returns 0 if it receives any negative input, but for any positive value x it returns that value back.
Fully Connected Layers - The purpose of the fully connected layer in a convolutional neural network is to detect certain features in an image. Most of the time they are placed in the end of the architecture of a CNN.
Dropout Layer - The dropout layer functions by randomly deactivating a portion of input units during each training update

1.5 Previous Research

From the inception of the computer, mankind has strove for its perfection and for newer and greater ways to use its capabilities. George Cowgill wrote in 1967 that the number of computer related tasks taking place in archaeology was no more than 20 in the entire world (Cowgill 1967: 17). These were the earliest days of the computer and since then it has evolved significantly. There have been several studies and scientific achievements using artificial intelligence in the field of archaeology.

Turing and the birth of artificial intelligence

The first to introduce the idea of artificial intelligence was Alan Turing, a British mathematician. His idea of artificial intelligence came in the form of the imitation game. He posed the question, whether machines can think, in 1950 (Turing 1950: 433). The imitation game refers to a test where a human judge interacts with two unseen participants, one human and one machine, via a computer interface. The goal is for the judge to determine which is which based on the responses received during the interaction. If the judge cannot reliably distinguish between the human and the machine based on their responses, then the machine is said to have passed the Turing test.The Turing test is designed to evaluate a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. The test does not focus on the machine's internal processes but rather on its external behavior in natural language conversation. Turing has been influential in the history of artificial intelligence and was the first to develop a real world approach to the subject.

Early use of artificial intelligence in archaeology

One of the earliest usages of artificial intelligence in archaeology was in 1989 by Vanda Vitali who has a Ph.D in materials science (Vitali 1989). The study done had two phases. The first phase was to analyze the archaeological interpretations made by an unaided archaeologist who participated in the study. The interpretations took the form of a text that used the results of chemical and data analyses as the basis for interpreting the origin of certain ceramic wares (Vitali 1989: 385). The results of this first phase revealed flaws in both argumentation and a lack of understanding of the origin of the items given the information. Having seen these problems that the archaeologist faced, the next phase was to formulate the reasoning and rules that would describe and guide the interpretation of ceramic origin for groups of ceramics based on the determination of their chemical composition. For this they developed a “prototype expert system”, something similar to artificial intelligence (Vitali 1989: 386). The information given to the artificial intelligence was archaeological and technical. The archaeological information gives the location of where the ceramics were found, the time period of which they were assigned and their type. The percentage of each type of ceramic was also given. The technical information given was based on the chemical composition of ceramics in terms of their elemental concentration for 15 minor and trace elements. Then they characterized the different ceramic groups with the basis of a comparison of their chemical composition (Vitali 1989: 386). The artificial intelligence they built, called VANDAL, was built using SNARK as the programming language.

SNARK and VANDAL is described in Vitali’s paper with the following quote:

> SNARK uses declarative schemes of a firsft-order logic to represent facts or data in the system. Specifically, facts are expressed in the form a R b where a designates an entity, R designates a relationship between a and b, and b designates a value…The database information in VANDAL is grouped in blocks and organized into a hierarchical fact base containing some 380 lines. This fact base was used to develop VANDAL and to initially test the rule base. A second data set describing additional ceramic groups was then used to test the whole system (Vitali 1989: 387).

This study is one of the first uses of artificial intelligence in archaeology and it has since paved the way for even more complex and impressive systems. This model that Vitali created, in spite of the technical limitations of its time, represents useful techniques for aiding archaeologists in their work. The use of artificial intelligence in archaeology has since then evolved over time, and even though it is challenging to pinpoint a single individual or group as the first to use artificial intelligence in the field, AI applications in archaeology have been growing in popularity in recent decades. Some researchers and archaeologists (Orengo et al. 2020; Sharafi et al. 2016) have incorporated machine learning algorithms and computer vision to analyze and interpret archaeological data.

How has image recognition been used in archaeology based on two case studies?

Manuel Domínguez-Rodrigo is probably one of the names that you will encounter when you are researching this subject. Together with other researchers around the world he and his team has been developing and trying out artificial intelligence in different image recognition softwares, such as VGG16 and Jason1, in order to identify markings on bones (Domínguez-Rodrigo et al. 2020: 2). The goal is to be able to see when meat eating first emerged in human society, which is something that has long been debated. The way I differentiated my research from the authors was by comparing the different papers to each other and drawing empirical conclusions from them.The research question that drives the work in this thesis is not present in any used paper and it is fairly simple in nature.

The research that previously has been done for the second case study was in order to more easily be able to use satellites to archaeologists' advantage. Image recognition software is something they also use in these types of research, just as in the research about bones. The difference in this scenario is that satellites have been utilized in all three examples I intend to make use of (Orengo et al. 2020; Sharafi et al. 2016; Soroush et al. 2020). The procedure is similar throughout all of the examples I use. Their main idea was to use satellite imagery of known archaeological sites in the Middle East located in barren landscapes. This thesis shows both how they were able to find new archaeological sites of interest and what the fundamental differences between them are. In the following, I discuss how the act of recognition was made possible and advance this research by showing how this can be applicable in other parts of the world by considering the methodological premises.

What are the main challenges and limitations of using Artificial intelligence in archaeological research?

This question was answered, not just with archaeological literature, but with literature about AI and its challenges from other relevant sources. Artificial intelligence is something that takes great work to be usable and it requires a lot of data to give reliable results. The European Commission had a report made on the impact of artificial intelligence and in this report, they used the term "datavore". A datavore, in this case, is a program that needs to devour enormous amounts of data to be trained (Tuomi 2018:3). This term that they used is something I myself use in this thesis, describing one of the biggest challenges we face in archaeology when using artificial intelligence. The archaeologist Christian Horn has written a few papers on the challenges of artificial intelligence in archaeology. In one of his papers that I will be using he mentions the term "bias" when writing about artificial intelligence and its challenges (Horn et al. 2022: 1218-1219). This is something that not just archaeologists write about but researchers in general when researching artificial intelligence. In all data based programs, there will always be some form of algorithmic bias. This type of bias exists in artificial intelligence due to a number of factors which I have written about. I differentiate myself in this particular question by giving a broader scope on what struggles exist in artificial intelligence for archaeologists but also why they exist and possible solutions for these. Like the examples of the studies on bone mark identification and satellite imagery demonstrate how important advances have been made. However, challenges persist, including the need for large labeled datasets, the existence of potential biases, and the interpretability of AI-driven results, highlighting the ongoing quest to harness AI's full potential in archaeological research.

1.6 Theoretical View

The theoretical part of my bachelor thesis discusses different theoretical approaches and word procedures.

The theoretical viewpoints that this paper leans on are "cyber archaeology" and “diffusion of innovation”. Cyber archaeology focuses on the bridge between computer science and archaeology. It is the exploration of how digitalization has revolutionized data collection, analysis, visualization, and dissemination of archaeological knowledge (Forte, Danelon 2019). Diffusion of innovation is the idea of something (AI in this case) gaining momentum and over time spreading into different areas of a population or field, originally formulated by Everett Rogers in 1962 (1962; see also Wayne 2022). The spread of artificial intelligence is not something archaeologists are responsible for as the advancements of this technology are being made by the large tech giants of the world. The archaeological field is simply affected by diffusion of innovation in the sense that the spread is inevitable and it is just the natural outcome of the strive for improvement.

It has been argued that archaeologists should not fear artificial intelligence, on the contrary, they should welcome it and adapt it to their research. Artificial intelligence is not something that will take away any jobs due to the special nature of our field (Horn, Green 2021:32). Horn and his team incorporates cyber archaeology in their way of thinking and reasoning. They emphasize that we should focus on semi automated data interpretation. The artificial intelligence we currently use and will develop can not and should not be able to do everything. There needs to be a human hand in the analysis and results that Artificial intelligence will bring. The big question is whether we can rely upon the interpretation the

Artificial intelligence will make without human intervention. It is in this question that Horn criticizes the idea to rely fully on automated models.

Diffusion of innovation is something that you might encounter in archaeology when artificial intelligence is involved in the sense that this work that is being done will significantly revolutionize the approaches that archaeologists may take in the future, the work being done will make AI spread more and more. "One way or another, the results will be of paramount importance for human evolutionary studies" (Domínguez-Rodrigo et al. 2020: 7). This quote from one of the papers used shows the diffusion of innovation theory (Wayne 2022) that lingers throughout almost all the research in artificial intelligence. Being able to use image recognition in order to identify and categorize bone surface modifications will shape the practice of archaeology and how future post excavation work will be conducted. It is also here that the theoretical term cyber archaeology can be applied. Just as mentioned above, Horn and his team stresses the idea of semi automated artificial intelligence, where both parties have a hand on the wheel, as a stable bridge between man and machine for more reliable and controlled results. The idea of this joint collaboration between archaeologists and machines gives rise to the theoretical term cyber archaeology. By exploring the ongoing revolution of the digitalization of the field and how the data is being worked, we get closer to truly understanding the true nature of both theoretical ideas presented here and their respective role in artificial intelligence and archaeology.

1.7 Method and Materials

This thesis was conducted as a literary study based on already published results. When writing an examination of artificial intelligence in archaeology, the choice of materials is paramount to ensure the accuracy of the research. In this endeavor, the primary focus has been on utilizing existing papers and research publications authored by experts in both archaeology and artificial intelligence. Several compelling reasons underpin this selection. First and foremost, leveraging the work of other researchers provides a solid foundation of established knowledge and insights. Archaeology is a multidisciplinary field and by drawing from a diverse array of papers this study aims to capture a broader spectrum of perspectives, methodologies, and findings. This approach gives a nuanced understanding of the complex relationship between artificial intelligence and archaeology as it allows the incorporation of various viewpoints and methodologies. The literary study sets up its own limitations in that no new AI training was done to compare to the already published papers that have been used. The preferred method would have been to train AI models to conduct my own test, however this was not feasible in the time frame given to this work.

In summary, the decision to rely on other people's written papers and research in this investigation stems from the desire to construct a well informed, credible, and nuanced exploration of artificial intelligence in archaeology. By engaging with the already existing research this study aims to contribute to the collective understanding of the complex relationship between artificial intelligence and archaeology.

Are there any drawbacks to the chosen approach?

While utilizing other people's written papers and research offers numerous advantages, there are some potential drawbacks that should be acknowledged and addressed in order to enhance the robustness of the study. One notable challenge is the risk of bias inherent in the selected literature. Academic publications may reflect certain theoretical frameworks, methodological approaches, or cultural perspectives that could introduce bias into the study. To mitigate this, I will try to seek a more diverse range of sources where it is possible. By incorporating a broader spectrum of perspectives, this study can strive for a more comprehensive and balanced understanding of artificial intelligence in archaeology. Another limitation is the potential for outdated information. Given the rapid evolution of both artificial intelligence and archaeological methodologies, relying solely on literature that is older than a few years may lead to an incomplete representation of the current state of the field. The older AI models used in Domínguez-Rodrigo´s study (Domínguez-Rodrigo et al. 2020: 8-9) about bone mark identification might become obsolete within a few years due to the rapid development of artificial intelligence. To address this, it is crucial that I supplement the literature with the most recent and relevant sources available, where possible. Incorporating recent studies and advancements ensures that the research reflects the latest developments and innovations in artificial intelligence applications within the archaeological field.

2. Artificial intelligence, Pattern and image recognition.

In the strive to find a working and successful artificial intelligence that can identify and categorize bone marks, there have been two major studies. Manuel Domínguez-Rodrigo has been part of both of them with one succeeding the other. The goal now is to briefly present the two, their methods, and findings to better get a grasp on the situation and what has been achieved.

2.1 First study on bone marks

Still, to this day researchers debate the emergence of meat-eating in the hominin species. Meat-eating has been linked to large changes in structural behavior as well as encephalization (the increase in brain size) and the emergence of stone tools. There are even theories that meat-eating started more than a million years before our current evidence of encephalization (Byeon et.al 2019:36). If that is true then it will change the current theories on the emergence of meat-eating in the hominin species. This is where the importance of correctly being able to identify markings on bones shows itself. Those bones come from a dig in Ethiopia, which McPherron and his team dated to 3,4 million years old (McPherron 2010: 857). Being able to properly identify bone marks is also of great relevance to archaeologists due to the fact that it may show behavioral traits in early hominins. Behavioral traits such as hunting, scavenging, and gathering. It may even show that hominids became apex predators at an even earlier stage than previously thought. Previous analysis of cut marks on fossils has shown us signs of different butchering techniques and the carcass acquisition strategies used by early hominins during the Pleistocene era in the early stages of evolution. Due to all this, there is a great need to accurately identify cut marks and to distinguish them from other modifications on the bones such as trampling or sediment abrasion (Byeon et.al 2019:36). The need for correct identifications is not only applicable to the earliest hominids but also the ones of homo sapiens. Homo sapiens is believed to have first arrived on the American continent some 20 000 years ago. However recent finds and claims have been made that suggest that homo sapiens had a presence there even as far back as 30 000 years. There have been findings of cut marks on fossilized bones that set the date back 10 000 years more than what was previously thought. In Uruguay, at the Dikika and Quranwala sites, there have been findings of what seems to be trampling and/or sedimentary abrasions (Byeon et.al 2019:37).

There has long been a disagreement between taphonomists on how cut marks could be properly identified. It is suggested by Domínguez-Rodrigo et al. that a scanning electron microscope is not practical for identifying the impact of butchery and trampling marks in complete bone assemblages (Domínguez-Rodrigo et al. 2009: 1). Through experimentation, a list of criteria was developed, but then another disagreement emerged. How should these experimental results be interpreted by individual researchers? There has been a lack of objective methods and the identification of bone marks and their classification depend solely on the knowledge and subjective expertise of the researcher doing it. The proposed solution to this problem is to introduce artificial intelligence such as a Convolutional Neural Network (CNN) and a Support Vector Machine (Byeon et.al 2019:37). Both methods will be explained in more detail below. However, both methods rely on machine learning algorithms to classify close up images of bone marks, they have even been able to match and sometimes supersede expert classifications.

In this study they used a sample size of 79 bone marks, 42 of which were trampling marks and the remaining 37 were cut marks. This quote from the author explains the small number of marks used for the experiment: "…the sample size (79 marks) was kept intentionally short in order to maximize human expert scores (the larger the sample the higher the identification failure rates by humans) and to minimize the computer's accuracy (larger sample sizes lead to better training and higher identification rates)" (Byeon et.al 2019:37) . All bone marks in this experiment were man made in the present time just for this task. The trampling marks were all made by using four different sizes of sand as well as gravel and clay. Domínguez-Rodrigo had previously done similar tests with bones and sediment to artificially produce trampling marks resembling those that would occur in a natural setting. The cut marks for this experiment were made with quartzite flakes which was the most common stone type used during the Pleistocene era, and through a process of elimination they came out with bone cut marks being statistically indifferent from those made by natural rock flakes due to the edges being irregular and extremely similar to those made by ancient hominids. Both sets of these bone marks were then mixed together to discriminate between the two. The 79 subjects were then photographed with a microscope at 30x zoom and later converted to grayscale to get better vision over the markings (Byeon et.al 2019:37). The reason why the images were converted to grayscale is probably because when doing this it simplifies the algorithm and reduces computational requirements (Canan, Cottrell 2012: 1), however this is not stated anywhere in the study.

The three classification methods used in this test for classifying and identifying the samples are: Convolutional Neural Network (CNN), a Support Vector Machine and human experts.

Convolutional Neural Network

CNN are potent deep learning based methods for classifying images (Byeon et.al 2019:38). The strength of this method is that a CNN can analyze hundreds of images in the time it takes a human to analyze a single one. The CNN takes an image and transforms it via several layers. Each layer includes convolutional, Rectified Linear Unit, Max-Pooling, and Fully-Connected layers. A CNN uses neurons in each layer to interpret images and analyze them. In this study, the CNN was trained by using adaptive moment estimation (Adam).

What Adam does great is being efficient at stochastic optimization. Stochastic optimization is about finding the best solution to a problem when the problem itself or the data involved has some randomness or uncertainty (Kingma & Lei Ba 2015:1). The training was done on 59 images of both types of marks (Byeon et.al 2019:39).To make this test easier and speed up the process they resized all the images to a fixed scale at 180 x 520 pixels / 0,09 megapixels. Other than this they also used the histogram equalization algorithm.

> Histogram equalization is a method to process images in order to adjust the contrast of an image by modifying the intensity distribution of the histogram. The objective of this technique is to give a linear trend to the cumulative probability function associated to the image (Coste 2018).

Another explanation for this is that they use this algorithm to make the dark parts brighter on each image for easier detection of the marks by the CNN.

Support Vector Machine

In this test, they classified the images by training a binary Support Vector Machine classifier. The first thing they did was to create visual words matching the images using what they refer to as the Bag of Words technique. In simple terms: the BoW technique ties words to different features in images. Then when an image is analyzed by the classifier or Support Vector Machine in this case, it will give results tying the words classified to similar findings in the object analyzed. The objective of the SVM algorithm is to find a hyperplane that to the best degree possible can separate data points of one class from those of another class when analyzing a specific image (Byeon et.al 2019:38).

Human experts

In this method, they used three experts in taphonomy with experience ranging from seven to twenty years. In their test, they had to analyze and categorize both modern and fossilized bones that had various different markings on them. After each had separately completed their analysis they had to compare their result among themselves and then to the ones that both computer methods had achieved (Byeon et.al 2019:38).

Data set, evaluation and results

As previously mentioned the dataset consisted of 79 different bones with individual marks on them, all of whom were made specifically for this test. The 79 bones were photographed and converted to grayscale. The test gave each computer program 10 images from each category (trampling and cut), the rest were used for training. For both computer programs, the bone mark identification was averaged by a little over 60 training models that all occurred at random, and from that they drew their results. The three tests all had different mean accuracy as expected. The CNN (Convolutional Neural Network) managed to get a mean accuracy of 91% when identifying marks. The best case mean accuracy for the Support Vector Machine came out at 83%. Both the SVM and CNN could pinpoint more marks overall on the bones than the three expert taphonomists in this study. The experts all shared similar numbers of found marks and the identification percentage was almost equal among the three. The total average success rate for identifying bone marks among the three experts was as low as 63%. The difference in success between the CNN and the human experts was 44.4444% in favor of the CNN, and 31.746% in favor of the SVM compared to the experts (Byeon et.al 2019:39). This study has shown that human experts with decades of experience in the field are only moderately accurate (63%) at successfully identifying bone marks compared to artificial intelligence when dealing with pictures like the ones used in this test. That is an insufficient and unreliable success rate when depending on correct analysis and evaluation of bone marks. Many archaeologists depend on experts like these to make correct assessments, especially at the Dikika and the Quranwala sites that could change the homo sapiens history as we know it. This study has shown that artificial intelligence can mimic human capacity and even greatly surpass it when dealing with bone mark identification. However, all this depends on the training that each artificial intelligence receives. The training set has to be diverse enough to successfully be able to identify a number of different marks. Future training and analysis should include more diverse marks and different techniques in making them. Different materials should also be used. The larger the data set used for training the larger the identification success rate may be (Byeon et.al 2019:41). Ultimately, the researchers concluded that "This work showcases the capabilities of deep learning algorithms to resolve the highly-controversial issue of BSM identification in taphonomy" (Byeon et.al 2019:41). They used the most likely and similar type of materials that can be linked to old fossils in order to make the bone trampling marks as authentic as they could.

2.2 Second study on bone marks

In the last two decades, the analysis of bone marks has become more sophisticated and archaeologists have started to use both 2D and 3D computing techniques in different methods. However, these methods are not without their flaws and they all share similar ones. They have shown to be too inconsistent when analyzing subsamples of an already analyzed larger sample. Some have had the unfortunate over exaggeration of their success even though they only used small sample sizes with little differences. One example is the previous study where they did exactly that. Although they justified their choice because they intentionally wanted to give the human experts a higher success rate (Byeon et.al 2019:37). Most recently the use of artificial intelligence has started to make a name for itself through the use of computer vision-based machine learning techniques. Domínguez-Rodrigo and other researchers have tried a total of seven different deep learning models in this study, in their efforts to successfully identify markings on bones. In this test, the most reliable one turned out to be a model named VGG16 (Domínguez-Rodrigo et al. 2020: 1-2). The name comes from “Visual Geometry Group”, the team that originally made it, and the number 16 refers to its number of layers used (Simonyan & Zisserman 2015: 6). VGG16 is a deep learning vision based artificial intelligence originally designed with the goal of furthering the development of CNN´s. The makers of VGG wanted to increase the depth of the layers used and this was only possible through the use of small (3 × 3) convolution filters in all layers (Simonyan & Zisserman 2015: 1). VGG16 has also been previously trained to be able to identify 1000 different image categories of different animals and objects, with over one million pre-trained images. Using this large set of pictures together with their own training set resulted in the best possible accuracy. Domínguez-Rodrigo and his colleagues also tried training the model from scratch on nothing but their own pictures of the artificially created bone marks. They do not however state if the training was annotated (Domínguez-Rodrigo et al. 2020: 3-4). Only using their own pictures as a training set resulted in VGG16 getting significantly worse results at only 76% mean accuracy (Domínguez-Rodrigo et al. 2020: 3-4). The theory is that a model trained on more than just bone marks will give significantly better results. Something that they proved in this test. Domínguez-Rodrigo hints at his previous work at the beginning of his text. He states that artificial intelligence's success at correctly identifying bone marks now exceeds human experts by almost 50%. The problem human experts face in challenges like these is important to highlight as it limits their rate of success. The previous study showed that the human taphonomist experts only had a success rate of 63% when identifying bone marks. This can be because they did not have access to the bones themselves, but only the two dimensional gray-scale pictures that were taken with a microscope. It is mentioned in the later article, that when presented with more variables, the expert's rate of success ranged from 68 to 80 % (Domínguez-Rodrigo et al. 2020: 2). That is something that also changes the rate of success dramatically depending on the expert's experience.

The different models used in this test are Alexnet, ResNet50, VGG16 and Inception V3 who are all winners of the "Imagenet Large Scale Visual Recognition Challenge". A competition that tests image recognition programs on 1000 different categories to find the best ones. In this study, these four models were also put up against two similar models called Jason1 and 2, as well as another one called Densenet 201. They did this in order to try and show that a high accuracy in identifying bone marks can be achieved not only with complex architectures but also with simple ones. The goal through this study is to develop and be able to use a more objective method of analyzing and categorizing bone marks as the experts can only be subjective in their assessment of bone marks.

Data set and approach

The way Domínguez-Rodrigo takes on the challenge of bone mark identification is similar to the previous study but with one addition, tooth marks. He also mentions that it previously has been argued that tooth marks are easily discerned from cut marks, given their widely divergent microscopic features. Therefore the addition of tooth marks in this study was necessary to better train the artificial intelligence and enhance the research. The test consisted of two different types of tooth marks, lion and wolf. The cut marks used in this test were made with stone flakes (no type specified) and the trampling marks were made with sand abrasions in a similar fashion as the previous study. The two different tooth marks as well as the other bone marks are chosen to serve as a way to address the accuracy in identifying structurally similar marks. The lion tooth marks were samples obtained from four long bones that had been fed to semi captive lions at the Cabárceno nature reserve in Spain. The bones with the tooth marks made by wolves also came from a nature reserve called El Hosquillo. Just like with the lions these bones had been used in feeding the wolves. Both sets of bones were properly cleaned and dried out to make it easier to analyze. Originally they had a total of 418 tooth marks from all the bones but these were sorted out due to distortion and unclarity, resulting in 106 being accepted for analysis. The tooth marks used in the study were all tooth scores, the other marks such as tooth pits were all documented but not used. The 488 cut marks were all made on cow bones, with 22 different non retouched flakes of flint. The marks were created in a special butchering practice using stone tools instead of modern equipment. This was decided to get the most authentic marks possible to how it would have been made in ancient times. These bones were also properly cleaned and dried after the butchering was done. The trampling marks were made from deer bones using two different methods to simulate different types of terrain. The first method consisted of using different grains of sand and having three individuals with different weights trampling on them with special shoes made from esparto grass. The other method used had small gravel instead of sand and a similar approach as the first method. From these two methods, images were taken and sorted to get the clearest and most visible marks (Domínguez-Rodrigo et al. 2020: 7). They argue in the study that by combining wolf and lion tooth marks you get the marks made by both types of carnivores, lions who solely eat meat and wolves who are durophagous carnivores. The different types of sediments were selected to best represent the types of archaeological sites where Pleistocene remains can be found. The same can be said about the cut marks, which were also made to represent Pleistocene butchering techniques (Domínguez-Rodrigo et al. 2020: 8). The images for the experiment were taken with a microscope at 30x zoom, all at the same angle and light to get the same results. The images were later cropped so only the marks themselves were visible, all to ensure fair testing. The test consisted of 488 cut marks, 106 tooth marks, and 63 trampling marks. Just as in the first study, this time they also converted the images to gray scale and resized them to all be 80 x 400 pixels (0.032 megapixels) (Domínguez-Rodrigo et al. 2020: 7).

The different models used and their methods

Unlike the previous study, this time they only used Convolutional Neural Networks in the study (Domínguez-Rodrigo et al. 2020: 7). No human experts were used as comparisons against the different models, probably due to their low success rate in the previous study when given the same images. All seven CNN models used were made with the Keras platform, which functions as the interface for the different programs. All models used had the TensorFlow as its backend (Domínguez-Rodrigo et al. 2020: 8). Backend refers to the parts of the program that can not be seen but it is what makes the entire thing work, like the gears in a clock. TensorFlow is a library of tools and tasks available to its users that allows them to build AI models more easily with already pre programmed functions. Both of these were used due to them being open source programs that anyone can use. TensorFlow may have been the chosen backend option due to Keras only supporting that particular one for a time around the same time this article was written (Keras 2023). All seven models used in this test were trained on 70% of the total number of pictures with the remaining 30% used for the classification test. They used data augmentation to alter the training images used, this was to get a larger variation and prevent overfitting. VGG16, ResNet50, DenseNet 201 and InceptionV3, were all used via transfer learning in this study (Domínguez-Rodrigo et al. 2020: 7-8).

Alexnet - This model is the oldest one used for this test as it was the winner of the "Imagenet Large Scale Visual Recognition Challenge" in 2012 (Domínguez-Rodrigo et al. 2020: 8). Alexnet is a model consisting of eight layers, five convolutional and three fully connected. The description used in this study varies from the one found in Alexnet's creator's own paper, therefore I will use the original creator's description of how it works. AlexNet consists of five convolutional layers. The first convolutional layer has 96 filters with a kernel size of 11x11x3 and a stride of 4 pixels (distance between the receptive field centers of neighboring neurons). The subsequent convolutional layers use smaller filter sizes. AlexNet uses max pooling layers. Pooling layers are typically inserted between convolutional layers to progressively reduce the spatial dimensions of the input volume. In AlexNet, max pooling layers are inserted between convolutional layers 1-2, 2-3 and after the 5th layer. To prevent overfitting, dropout is applied to the fully connected layers during training. Dropout randomly drops out a number of neurons during each training iteration, forcing the network to learn more (Krizhevsky et al. 2012: 1-7).

Jason1 - This model was inspired by the architecture of VGG, the same as VGG16 is made from. This Model has eight layers in total, 4 CNN layers and 4 max pooling layers. The first CNN layer had 32 filters with 3 layers following it with 64 and 128, 128 respectively. Between these four layers, they placed the four max pooling layers to filter the relevant information (Domínguez-Rodrigo et al. 2020: 8).

Jason2 - This model builds upon Jason1 but incorporates the workings of VGG16 which is also used in this test. "...the model consists of a series of three blocks, each of them containing 3×3 kernel double layers of 32, 64 and 128 neurons respectively. In between each block, there are max-pooling (2×2 kernel) layers" (Domínguez-Rodrigo et al. 2020: 8). Dropout layers were also used in this model as it was of a more complex build.

VGG16 - This model turned out to be the best one in this study. Like some of the other models, this has also won the "Imagenet Large Scale Visual Recognition Challenge", this time in 2014. The model originally contained 16 layers but an extension to it was made from VGG16, which added 19 more layers, however, it is not stated if they used this extension or not (Domínguez-Rodrigo et al. 2020: 8). For being the best model used, the authors of this paper gave a very vague description of how it works, the rest is from the original creator's description. The core building blocks of VGG16 are simple 3x3 convolutional layers stacked on top of each other. The creators believed that using a small kernel size (3x3) with a stride of 1 was more effective than using larger kernels. The total layers of the model are 22. 13 of which are convolutional, 5 max pooling, 3 fully connected and 1 softmax (Simonyan et al. 2015: 2-4).

ResNet50 - This model is also a previous winner of the competition but in 2015. The total layers of this model are an astonishing 50. ResNet50 uses something called "skip connections", allowing inputs to "skip" some convolutional layers. This model takes after and expands on VGG layered blocks. ResNet50 consists of sixteen blocks, each containing three layers with one layer at the beginning and one at the end, making a total of fifty layers. In between these layers, the model allows skipping, which enables the training of models with many layers to direct the flow of information from one layer to a later layer. This ensures a significant reduction in training time and improved accuracy in image analysis (Domínguez-Rodrigo et al. 2020: 9).

InceptionV3 - This model was originally called GoogleLeNet but since then, it has had updates to its functions as well as its name. The layers in this model are a total of 42. The model has Inception modules that use parallel convolutional layers of different filter sizes to capture features at various scales. The model also uses factorized convolutions, batch normalization, and global average pooling for efficiency. InceptionV3 includes auxiliary classifiers for intermediate supervision during the training. Pre-trained weights are used for transfer learning on specific tasks (Domínguez-Rodrigo et al. 2020: 9).

Densenet 201 - This was the biggest model used with a total of 201 layers. Each layer in this architecture gets the feature map from the previous layer as input. This method enables the detection of a wider diversity of features in images compared to the other CNNs used. Densenet 201 is organized into several dense blocks (as the name suggests), transition blocks, and a final classification layer. The number of dense blocks and their sizes contribute to the overall layer count. The total number of blocks is only four but these four blocks contain a large number of layers (Domínguez-Rodrigo et al. 2020: 9).

Results

Below are the accuracy and loss for each model. The closer the accuracy number is to 1, the more successful the model was in its identification. The closer the loss number is to 1, the worse the model's prediction was from the actual value. However, prediction accuracy in relation to 0 as loss can be well over one, a model can still predict somewhat accurately even though its loss is high. Loss is a mathematical function that quantifies the difference between predicted and actual values in the model (Domínguez-Rodrigo et al. 2020: 1-2).

Model	Accuracy	Loss
Alexnet	0.78	1.69
Jason1	0.88	0.36
Jason2	0.86	0.57
VGG16	0.92	0.36
ResNet50	0.74	0.96
InceptionV3	0.74	1.13
Densenet 201	0.76	0.76

(Domínguez-Rodrigo et al. 2020: 2).

As it was previously mentioned, VGG16 was the overall best performer, both when it came to accuracy and loss. The closest model to VGG16 was Jason1. Just like the authors wanted to try and show that a high accuracy in identifying bone marks can be achieved not only with complex architectures but also with simple ones. The success of Jason1 shows just what they wanted to prove (Domínguez-Rodrigo et al. 2020: 2). The low loss value for both Jason1 and VGG16 gives a higher probability of classifying individual marks. However, the loss for both Alexnet and InceptionV3 together with their lower accuracy makes them inappropriate for usage in the identification of bone marks.

The conclusion from both studies

As the first study only used 79 images for their study, and those images did not include tooth marks, they were able to achieve an accuracy of 91%, with no loss mentioned (Byeon et.al 2019:36-37). This high accuracy was the result of a small number of pictures and with little variation. However, in the later study, they used a total of 197 images for their identification process with 460 images used for training (Domínguez-Rodrigo et al. 2020: 7-8). This resulted in both the training size and sample size being much larger than the other study, giving more room for error. The addition of tooth marks also gave the CNN models another entire category to sort the analyzed pictures in. Another addition to this study is the usage of gravel in the making of the trampling marks, giving way to much larger markings on the bones than what was previously made from sand alone. VGG16 with its high accuracy of 92% is something that had never before been achieved on so many samples of different types of marks and with such a large test set. Going back to my original question - Can Artificial intelligence help in pattern recognition and the identification of markings on bones? The development and implementation of artificial intelligence in the form of image recognition software in archaeology and taphonomy has already shown to be both equal and most of the time, it surpasses that of human experts. It is safe to say that artificial intelligence is the next step of evolution for this type of research and we have yet to see its full capabilities.

3. The discovery of new archaeological sites through the usage of artificial intelligence

In the last decade researchers have been studying artificial intelligence and image recognition in bone mark classification. However, there have also been other studies with image recognition software and the effort of trying to find and identify new, never before seen archaeological sites. The archaeological field has seen great change since the time of Howard Carter. From painstaking manual search and excavation through jungles and deserts in the hopes of finding treasure, to a computer, scanning the earth for new and exciting discoveries. At the core of this revolutionary approach lies the utilization of artificial intelligence. Archaeologists are now able to use computers, satellites and artificial intelligence image recognition software to search and find new archaeological sites of interest. The speed and efficiency that artificial intelligence brings to the table have significantly increased the speed of which archaeologists can get their work done. In this section of the thesis I will explore and discuss how artificial intelligence can help archaeologists when searching for new discoveries. The goal is to compile different and fully understand the significance in image recognition software and all its benefits.

3.1 *Using image recognition to find new archaeological sites - First study*

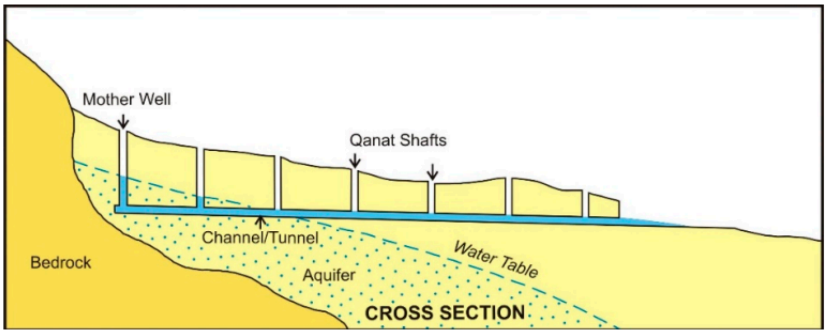
Over the last decade, archaeologists have started to look at computer based image recognition software in order to help them advance their work. Some archaeologists like Domínguez-Rodrigo and Byeon have used image recognition software to identify bone marks (Domínguez-Rodrigo et al. 2020; Byeon et.al 2019). However, there are others that make use of satellite imagery in order to find new and undiscovered archaeological sites of interest (Orengo et al. 2020; Sharafi et al. 2016; Soroush et al. 2020).

Automated Qunat shaft detection

In this paper Mehrnoush Soroush and his team made use of a CNN based deep learning model, they refer to this model as 2D FCN (Soroush et al. 2020: 8). The model was used together with cold war era CORONA satellite imagery in order to detect qanat shafts in Iraq. ‘ The author also states that this team are the first ones to use automated techniques on historic satellite imagery (Soroush et al. 2020: 8). It is also noted that the volume of space and air-born imagery is ever increasing. This has great potential for the field and archaeologists can easily make use of and compare old and new imagery. The reality, however, is much different from what it could be. This large quantity of imagery that is available to archaeologists is often still handled by traditional methods, such as manual visual inspection and markings of potential points of interest. Researchers can not handle the ever increasing quantity of available imagery and still rely on traditional methods. They risk missing important information due to solely relying on their own visual and subjective judgment as a result of their individual experience. The best way to address this is to automate parts of the process when searching for archaeological sites in imagery (Soroush et al. 2020: 8-9). The authors of this paper state that there has been a low success rate of the earlier attempts of image recognition software in archaeology and has resulted in skepticism towards artificial intelligence. However, this is contradictory to the results of the bone mark experiment done by Byeon and her team (Byeon et.al 2019), which also gave way to further research about the subject. This can be overlooked as something the authors missed due to them not working with the same type of archaeology. The machine learning way that pattern recognition has now gone down has dramatically increased the ease of use for archaeologists. Convolutional neural networks are now reliable and they may shorten the research time dramatically (Soroush et al. 2020: 2).

The qanat

A qanat is an underground water supply canal often composed of horizontal tunnels and a series of vertical qanat shafts. The qanat collects water from underground and transports it to the desired location, which might be tens of meters to hundreds of kilometers away from the source. Researchers still debate the origin of this technology. Although it is certain that by the first centuries of the second millennium CE, many areas in both Africa and Asia used variants of qanat technology as their water supply (Soroush et al. 2020: 3).



Example of qanat shafts and tunnel (Soroush et al. 2020: 3).

In order to properly map the qanat system, all the individual shafts need to be documented. However, when mapping the qanat shafts it is easier to do so on the smaller systems. When mapping the larger ones that stretch for several kilometers, it is more common to have a line representing the canal's pathway from its source to the outlet. The complexity is lost in the picture when mapping large canals shaft points (Soroush et al. 2020: 4). It is rather easy to detect qanat shafts if they are well preserved. The shafts were originally doughnut shaped as they had dirt lying in a ring around them from the original dig. Over time and as a result of not being maintained and/or in use they have lost their doughnut shape, but if visible they remain somewhat circular in shape. These shafts are also often found in a line, something you do not often see naturally occurring (Soroush et al. 2020: 4).

The study

They had two goals in mind when conducting this study. The first goal was to test and see if deep learning could be used for automated detection of archaeological features with relatively uniform signatures, qanat shafts in this case. Their second goal was to assess the deep learning models performance on publicly available historical images. The CORONA satellite in this case gives panchromatic images with 1.8 spatial resolution (Soroush et al. 2020: 4). The choice of using the CORONA satellite was also a way of bringing down cost for this and future studies as it is available to the public for free. This has also its cons and pros as you are working with old historical photos and not the latest ones. However, this can be a good thing because new photographs from developing countries such as Iraq in this case, often have changing landscapes due to development projects (Soroush et al. 2020: 5).

The study was conducted in Iraq, in an area covered by the “Erbil plane archaeological survey” group which covers 3200 km² in the center of the Erbil Governorate of the Kurdistan region. Locally the qanat is known as karez and it has been heavily used in this region. The geographer Dale Lightfoot has previously 683 qanat systems in this area using historical maps and field observation, he did not have access to any modern computer programs so he did all of this manually (Soroush et al. 2020: 5). EPAS has been investigating and mapping qanats since 2018, however only 5-10% of those who are visible on the CORONA satellite have been documented with modern photography due to development in the area. The images from the CORONA satellite were acquired between 1960 and 1978, these images detail the area right before the development of this region started and forever changed the landscape. In preparation for this study they created a map documenting different shafts using a combination of CORONA imagery and manually marking 12 000 shafts in a GIS database (Soroush et al. 2020: 6).

Materials and methods used

The imagery used for training in this study came from images declassified in 1995,originating from the US intelligence images obtained in the CORONA program. The CORONA program was a US based intelligence program designed to gather information about the Soviet missile strength between 1960 and 1972 during the cold war. These images have been used by archaeologists since 1998 when working in the middle eastern region due to the rapid development taking place there. The spatial resolution of the CORONA images were also in the acceptable range making it easier to identify archaeological features (Soroush et al. 2020: 7). The KH-4B imagery used was downloaded from the USGS website. It has a spatial resolution of 1.8 m and was taken in the winter of 1968.

Convolutional Neural Network

The CNN used in this study used a binary classification model for qanat feature segmentation. Binary classification is fairly simple as its function is to classify objects as either true or false (Kumari & Srivastava 2017:1). Either it is a qanat or it is not.
Qanat feature segmentation means that the goal is to partition an image into distinct parts based on certain criteria, such as qanat in this case. This makes it so that the CNN more visibly detects and displays the shape of the qanats for the researchers. The architecture of the CNN used was inspired by another model called 2D U-Net as it uses some of its features. The model they used (2D FCN) consists of 22 convolutional layers in groups of two, 5 max-pooling layers between the first groups of convolutional and 5 up-sampling layers between the last groups. The model was also fitted with shortcut connections so that any

extracted information can skip to a later convolutional layer if needed (Soroush et al. 2020: 9).

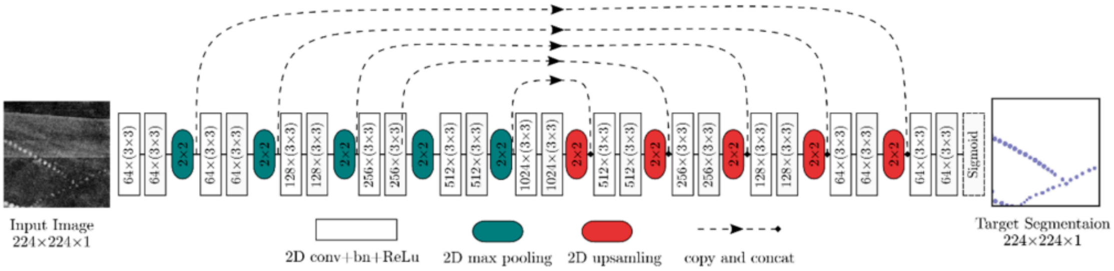


Illustration of said model with its different layers and connections (Soroush et al. 2020: 9).

Training of the model and results
For this model they used a stochastic gradient descent algorithm for the Keras framework. Stochastic gradient descent updates the parameters for each training example resulting in the training time being reduced (Alagözlü et al. 2022: 2). Annotation was done by one of the authors of the paper who specializes in remote sensing of landscape irrigation features. For data labeling, they used “3D Slicer”, an open source software platform often used in medical image processing and annotation (Soroush et al. 2020: 7). The spatial resolution (1.8m) was not enough to be able to properly separate all the different qanat shafts individually. Therefore, many shafts were connected in the annotation process and treated as one, creating larger clusters. After the annotation process they analyzed the work to find potential overlapping of labels. The data was labeled for eleven image patches over the Erbil landscape that had a dense network of clear qanat shafts visible (Soroush et al. 2020: 7). The training was performed on 11 2000x2000 CORONA image patches, an area covering around 30,2 square kilometers. Each training image was a 2D patch with a size of 224 x 224 x 1 pixel in grayscale. Data augmentation was also performed in the images by flipping some of the images upside down and left/right, creating an unknown number of more images as they do not say how many more. They used five fold cross validation in this test in order to prevent overfitting. Five fold cross validation means that they split the training set into five equally large sets of images and trained the model on four of them, then they compared the results to the fifth set. This training was repeated five times so that all the different sets of images got to be the validation set (Soroush et al. 2020: 10).

When doing the test, the model performed better on the patches of images containing more dense qanat shafts than on those with less density. For the five images with the most dense and largest number of qanat shafts the model had an mean accuracy of 0.654 (65,4%). On the other six images with less dense and a smaller number of shafts the mean accuracy was 0,34 (34%). The low accuracy of detection of qanat shafts on the six images depended on the presence of bumpy features on the ground getting recognized as shafts. These false positives that were recognized as shafts lowered the overall accuracy. The five images with denser and more shafts overall also got false positives. However, the authors think that there is a correlation between model precision and labeled features in an image. The ones with more labeled shafts had lower false detections and vice versa (Soroush et al. 2020: 11). Out of the eleven images used for the test the one that had the highest accuracy was patch 11/11. The number of detected features were a total of 1178 with an accuracy of 0,764

(76,4%). The worst performing picture was 4/11. This image resulted in 23 detected features with an accuracy of 0,107 (10,7%) (Soroush et al. 2020: 11-12).

Conclusion

This study that they made has shown that image recognition software can be used to detect qanat shafts. The results may vary depending on the density and number of shafts however, as the images used with the least amount of shafts only had a mean accuracy of 10,7% (Soroush et al. 2020: 12), which is very low. However, since qanat structures exist in many countries throughout Africa and Asia, it is possible to enlarge the training set with images from more regions in order to better optimize the CNN. It is relatively easy to validate the finding of the CNN due to the fact that qanat shafts are found in large and linear alignment with each other. In this study they were not able to validate many shafts with in person inspection due to the development of the land that has been done since the photos were taken in the 60's (Soroush et al. 2020: 13). The authors mention that other archaeological experiments with image recognition software have used models (no one is specified) trained on other things than just archaeology and once more researchers start adding data sets of solely archaeological images then the results might improve (Soroush et al. 2020: 14). The perks of using larger training sets and not just archaeological images were shown in one of the bone marks studies. There they made use of the large training set used for the "Imagenet Large Scale Visual Recognition Challenge", together with their own set of bone mark images. This resulted in the model VGG16 being significantly more precise and accurate when analyzing images (Domínguez-Rodrigo et al. 2020: 3-4).

3.2 Second study

This study done by Hector Orengo and his team was done by using satellite imagery with the goal to detect archaeological mounds in the Cholistan desert in Pakistan. The Cholistan desert has one of the world's largest concentrations of Indus civilization sites dating from 3300 - 1500 BC. Artificially made mounds are a common characteristic of both permanent and semi permanent settlements of old civilizations. These mounds are easily spotted due to their often large protruding shape. When searching for mounds in areas now populated and developed, researchers can make use of the CORONA satellite images that are available to the public for study. However, the authors of this study have made use of the Sentinel 1 and 2 satellites for their project, using new and modern imagery (Orengo et al. 2020: 1-3)

The Cholistan and Indus

The Cholistan desert stretches from the southern edge of the alluvial plains of Punjab to the north of the Sindh province in Pakistan. This desert is highly sensitive to the annual variation of the Indian summer monsoon and the intensity of which has significantly affected its population and ecological diversity throughout the Holocene period. Today in modern times the area has a lot of fossilized sand dunes as well as small shrubs and trees. This desert was once an important route between central and south Asia (Orengo et al. 2020: 2). The area was the most important for the Indus and had a large settlement concentration between 2500 and 1900 BC. The total number of archaeological sites were counted to 462 in the 1940's (Orengo et al. 2020: 2).

Materials and methods used

For this study the researchers made use of the Copernicus Sentinel satellite series, Sentinel 1 and 2. These satellites were chosen because they offer fairly high resolution and they are open access. The spatial resolution these satellites use is 10 and 20 meters per pixel, this is sufficient for the detection of mounds since most mounds in this area have an average diameter of 100 meters (Orengo et al. 2020: 3). Sentinel 1 has several scanning modes, the one used in their study was the interferometric wide swath mode. This mode is the main mode used for many satellites when scanning large land surfaces and satisfies the majority of service requirements. The resolution used with this mode was 10 meters per pixel. Sentinel 1 provides data imagery starting from 2014 and uses SAR. SAR or Synthetic aperture radar is a way of taking images by the use of radar scanning. This makes it easier to photograph land surfaces through clouds and rain. Sentinel 2 used a mix of a ground resolution of 10 meters per pixel and 20 meters per pixel. This mix was done by the use of different bands, making the images from Sentinel 2 multispectral (Orengo et al. 2020: 4). A multispectral image is a collection of several images of the same scene, each of them taken with a different sensor. Each image can have a different number of bands. Each band captures the light of a specified wavelength, giving the features in the image different colors depending on what it is (You 2021: 4). They decided on using these two satellites and their different technologies together because on their own it would not be enough. The drawback with using SAR images is the presence of noise in the pictures taken, coming from microwaves in the atmosphere. SAR can not on its own detect and isolate archaeological mounds from naturally occurring desert shapes. Also, optical multispectral imagery used by Sentinel 2 can not on its own differentiate between mounds and natural accumulation of clay that produce similar visible features on images taken. The two satellites were used together in order to produce images that could be used in this test (Orengo et al. 2020: 4).

The program used for this test was the Google earth engine cloud computing geospatial platform. The GEE platform is a free to use, web browser based program that anyone can use and it works great for remote sensing based applications. GEE was the chosen platform for this test because where it excels is at implementing large scale data analysis as it provides access to 20 petabyte of satellite imagery and geospatial datasets, which includes the Sentinel satellites. GEE uses code developed by its users in JavaScrips or Python, in this study, Java was the chosen program. The benefit of using this cloud operated platform for any test of this caliber is that Google's infrastructure has a vast network of powerful computers doing the work. This minimizes cost and time as the researchers do not have to supply their own computers for the analysis process. GEE also uses high resolution imagery that allows the evaluation of the results of the classification and the selection of new training data without needing to export them to another software (Orengo et al. 2020: 3).

The machine learning algorithm and training

“The steps for classification of mound-like signatures included gathering training data, training the classifier model, classifying the image composite, and then validating the classifier with an independent validation set” (Orengo et al. 2020: 4). When training and validating, they made use of 25 different archaeological mound sites. These sites were selected because they could be clearly found and identified in the GEE image database. Geometrical shapes were drawn in GEE to define the mounds from which the values of the pixels in the image could be extracted for the training of the algorithm (Orengo et al. 2020: 5).

The model also made use of a random forest classifier, which is a collection of individual decision trees. Each tree predicts a class of whatever they are trying to find, and the tree with the highest probability is selected for the result; this is not done on every result. They decided on 128 trees for an optimal result without increasing computational power. Even though they use Google's cloud computers, there is still a waiting time on their end for any result. The experimentational architecture went through three versions. The first version successfully identified 20, already known mounds, as well as a number of potential ones. However, they needed to make two new versions of the test because the percentage of pixels being identified as potential mounds were too high. They decided that everything that resulted in under a 0,55 probability of a potential mound should be filtered out and not registered, thus minimizing the potential for error (Orengo et al. 2020: 5).

Results and conclusion

After deciding to filter out the results under 0,55 (55%) probability they got a result of 337 clusters of potential mound structures, 25 of which were the ones used in training. The random forest method used only managed to produce a few mound like structures as result, they do not state exactly how many. Out of the 337 results, only 71 mounds could be identified as previously known sites (Orengo et al. 2020: 6). The mounds were all detected in desert areas. The authors think that the new distribution of found mound structures indicates that the Thar desert has expanded considerably since the Indus period. There are known medieval settlements that form an arc right at the north west end of the Thar desert. The authors also think that this may indicate that the expansion of the desert has been a long term process taking place over several millennia, resulting in the abandonment of settlements (Orengo et al. 2020: 8). The conclusion from this study is that computer based image recognition software can be used to find archaeological mounds. Through their research they were able to locate 266 not previously known potential mound structures. They, however, did not authenticate their findings with in person visits to these sites. They mention the need for further research on the unknown sites. Even though the model found new sites and they verified the existence of mound like structures on GEE, there is still a chance that some of these are archaeological mounds but rather natural occurring structures (Orengo et al. 2020: 7,9).

3.3 Third study

The purpose of the third study was to investigate the possibility of using pattern recognition software in detection of archaeological sites of the semi-arid Khorramabad plain located in west Iran. The place chosen has been suitable for human settlements for over 40 000 years. The area has, however, been the victim of erosion and sedimentation during the pleistocene and holocene periods, resulting in the disappearance of archaeological sites. The authors of this paper writes that this study will be the first of its kind, using artificial intelligence in a semi-arid context (Sharafi et al. 2016: 1-2). They make it very clear that they aim to use one-class classifiers as their model. One-class classifiers are the models that are trained solely on the object they are searching for, just as Orengo and his team only trained their model on archaeological mounds (Sharafi et al. 2016; Orengo et al. 2020). This study also makes use of GIS spatial analysis. GIS spatial analysis is used in the GIS program and its function is to answer “where questions”. Where something located or where geographical change occurred. The team used these two tools in order to try to locate archaeological sites in the Khorramabad plain (Sharafi et al. 2016: 2).

The Khorramabad plain

Located in west Iran, these plains are one of the oldest where residential remains have been found, dating from the paleolithic through the islamic era. Previous archaeological excavations have been done since the 1890's and it has shown the importance of the area. Environmental changes transformed the area during the end of the pleistocene, like the formation of Kar-Gah lake and the formation of different paths in the Khorramabad river. The team thinks that changes like these caused archaeological sites to be destroyed and/or buried under soil.

Methods used and the model

Their test consisted of a two stage study. The first stage was to collect the environmental factors of 43 different archaeological sites in the Khorramabad plain using ArcGIS. The second step was to use the results collected from the first step to create a predictive one-class classifying model. The data from the 43 sites included elevation, slope, precipitation, distance to river, distance to accessible roads (ancient roads), and water resources. These different factors were then turned into raster layers using ArcGIS (Sharafi et al. 2016: 4). "A raster layer consists of one or more raster bands — referred to as single band and multi band rasters. One band represents a matrix of values. A color image (e.g. aerial photo) is a raster consisting of red, blue and green bands" (QGIS). In other words, the data were converted to pictures representing the different distances and locations. All examples except precipitation were generated using topography maps. Precipitation was generated using data from synoptic and climatology stations. The values for each data point were exported to an excel document in order to use min-max normalization to reduce the effect of the measurement unit on the learning process of the model (Sharafi et al. 2016: 4). Min-max normalization is often used to scale your data values so that they fall within a specified range, typically between 0 and 1. Since the data they had was only on the 43 archaeological sites, they artificially created another 43 sites, but without archaeological presence. This was done to make the training of the model more successful. The dataset which contained 86 archaeological and non-archaeological samples was created and used to train, validate and test the predictive model using a nested ten-fold cross validation method. The method used is similar to the one used by Soroush and his team (Soroush et al. 2020: 10). However this method divides the training set into ten groups, nine for training and one for validation. This is repeated ten times so that every group can be used as both training and evaluation. The "nested" part of the procedure means that they also split the training set in three parts and trained them against each other (Sharafi et al. 2016: 4).

Results and conclusion

After training their model, all algorithms were implemented in the Matlab program. Matlab is a common tool used to create datasets and build models like the one used in this study. The experiment itself was repeated ten times for each variable to ensure that the results were reliant. The mean accuracy for their test was 0,91 (91%), a fairly high accuracy. This high accuracy gives support to their claim that one-class classifiers can be as good as, and sometimes even better, than other models (Sharafi et al. 2016: 8-9). In addition to this they also tried another experiment. They used real world non archaeological sites as well as artificially generated non archaeological sites. This was to fully evaluate the models performance when classifying sites. All together they had 87 different non-archaeological sites that they ran through their model in three different groups of 29. In the first group the model classified all the 29 samples as non-archaeological with 100% accuracy. Group two and three both classified 28 out of their respective 29 samples as non archaeological, giving them both an accuracy of 96,5% (Sharafi et al. 2016: 9-10). Both experiments done show great promise when using artificial intelligence to detect and differentiate between what is and is not archaeological. Their approach of using one-class classifiers gave them good accuracy when classifying both kinds of sites. However, the dataset for their training and experiment was quite small, making it less objective than ideal.

Conclusion from all three studies

Given the information from all three studies (Orengo et al. 2020; Sharafi et al. 2016) (Soroush et al. 2020), it is safe to say that artificial intelligence can be very successful when trying to locate archaeological sites. The test done by Orengo and his team can be replicated by any archaeologist wanting to try it out for themselves. Both the code and the satellite data can be accessed through the links in their paper. They have also provided instructions for how it can be modified to fit other areas (Orengo et al. 2020: 5). All of the three studies were made in order to further the development and understanding of artificial intelligence in archaeology. They all give fairly detailed descriptions on how they trained the models and carried out the test. One of the things that could be improved upon in future tests of similar structure is to use larger training sets. The one used by Sharafi and his team for example used a rather small sample size and that might correlate to their rather high accuracy (Sharafi et al. 2016: 9-10). Deep learning models such as the one used in these three experiments can assist archaeologists in detecting objects faster than standard conventional methods whilst both saving money and time. Ultimately, today there is no open and public access standard data that can be used by archaeologists worldwide. Therefore, further developing this technology and sharing the data is of great importance to the field. The need for a collective repository has never been greater (Jamil et al. 2022: 13).

4. Artificial intelligence and its challenges in archaeology

Archaeologists are experiencing a dataflood, fueled by a surge in computing power and devices that enable the creation, collection, storage and transfer of an increasingly complex amount of data (Casini et al. 2021: 1). The introduction of artificial intelligence and the different ways it can be implemented have changed every field that can use it. All of this is fueled by the different companies that make it possible through new and more powerful iterations of computing parts. Companies like Intel, AMD and Nvidia are some of the ones that stand at the forefront of technology and what is possible. Whilst CPUs and GPUs both grow more and more powerful every year, they still struggle when training on 3D data and sometimes the funding for projects involving this technology might not be enough. The studies researched in this paper (Orengo et al. 2020; Sharafi et al. 2016; Soroush et al. 2020; Byeon et.al 2019; Domínguez-Rodrigo et al. 2020) all shows great promise for the future of artificial intelligence in archaeology. However, it also shows some of its challenges we archaeologists face going forward. Small training sets, subjective bias, lack of shared data are just a few of them. This part of the thesis explores some of the main challenges and problems involving artificial intelligence in archaeology and some possible solutions.

Bias

Bias is defined in the Cambridge dictionary as “the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment” (Cambridge Dictionary 2023). In the context of artificial intelligence this would mean that the people who trained said model showed bias in their input and training, resulting in a model who also shows bias in its algorithmic behavior. The quick rise of artificial intelligence in archaeology gives way to many new opportunities but also risks algorithmic bias if not careful. Algorithmic bias is the presence of systematic and repeatable errors in a computer system that create unfair outcomes (Florida State University 2023). Creating training sets for artificial intelligence can be a lengthy task. It often involves searching data sets and/or making sets of your own. When dealing with deep learning models, it may include making sure that small sample sizes are over represented to give the model enough training data and learn the supposed pattern (Roselli et al. 2019: 3), this can be done through artificially creating more samples from just a few real ones.

The creation of data sets and training of a model is a lengthy task and can often lead to some of its problems, like bias. The training sets used are manipulated by the archaeologist in charge of the project and as so, influenced by that person's bias. Archaeological data sets are always fragmentary, they change over time and therefore one needs to be careful when training a model. Training should preferably be done on samples similar to the ones later used in the test. Even well trained models do not get 100% accuracy every time (Roselli et al. 2019: 3-4). Classification is basic to every comparative analysis and it is the classificatory process that may introduce potential bias. When creating the data set for a deep learning model it is up to the archaeologists involved to classify and select the relevant information, the accuracy of this relies on human perception. When humans are involved there is always variation in classification and answers, even if the criteria are strictly defined (Beck, Jones 1989: 244). Bias in humans and algorithmic bias in models does not occur randomly, it is produced by factors that skews the data distribution. This may come as the result of a difference in perception among archaeologists or just a change in perception over time. The difference in perception between archaeologists may lead to a difference in classification, resulting in the training data being biased, thus doing the same to the algorithm (Beck, Jones 1989: 245).

The difference in perception may also come due to a difference in experience. In the study by Byeon et al., the human experts that they used in their experiment as comparison only got 63% positive classifications of bone marks (Byeon et al. 2019: 41). Using the above example, an analyst may be quite liberal at the beginning of the analysis and identify slightly concave or convex edges as straight, but with time may become more conservative so that in the end, only absolutely straight edges are so categorized (Beck, Jones 1989: 245-246). Depending on its nature, errors like these introduced during classification of bone marks or other archaeological features might lead to grossly incorrect interpretations and ultimately bias in the algorithm if it is allowed to be used as training material. Given the number of potential ways that bias can be introduced into an algorithm, it can be a difficult task to mitigate this. There is not a single approach that can solve all potential bias but there are a few solutions that may reduce the risks. One solution proposed by Roselli et al., is that any process to evaluate an AI for potential bias must be carried out by those who are not involved with the project (Roselli et al. 2019: 4). If this is done on both the training data and the model it will be a good solution to verify if all the data is accurate as a non-participant, preferably experts in the same field, is able to overlook it. Another thing that can be done is to monitor the “production data”. The data that the model gives you needs to be consistent with the data used in training for any successful classification. If this is not the case then the model might need to be retrained on another data set.

Data sets

Almost all current deep learning systems rely on what is called a supervised model of learning. The supervised model is based on training data that has been labeled by humans so that the model can be adjusted when the labels for data are wrongly predicted (Tuomi 2018: 15). This data that is being used in models, most of the time, comes from and represents the region and/or things that the archaeologists are studying. Examples as the bones used by Byeon et al. (2019) or the mounds used by Orengo et al. (2020). The model used by Byeon and her team was only trained on a set of bone marks that they themselves had created and then tested on another set, also made by them.   (Byeon et al. 2019). The same goes for Orengo and his team as the mounds that were used for training also came from the exact same area that they used in their testing (Orengo et al. 2020: 6). These small data sets may give an inflated success rate due to them only being trained and tested on the same marks and locations. Domínguez-Rodrigo et al. notes that “some geometric morphometric analyses have also overemphasized their success, using very small sample sizes and by showing accuracy rates that are similar to less sophisticated methods” (Domínguez-Rodrigo et al. 2020: 1). This is one of the main problems when dealing with artificial intelligence as it needs to be trained on large data sets with a variety of examples to be able to give reliable results, otherwise the success rate will be grossly overemphasized and the model itself might not be useful on anything other than the training data/region. The need for large data sets has led the European Commission to call these data based AI models “datavores” (Tuomi 2018: 3). A datavore is a system that requires vast amounts of data for training to be able to give reliable results. One of the biggest technical bottlenecks at present is the availability of relevant data. This is a very present problem in archaeology. As written in Jamil et al., there is no open and public access standard data that can be used by archaeologists worldwide (Jamil et al. 2022: 13). The need to share data has never been higher than now due to all the different AI projects taking place around the world. Orengo et al. made their project available for anyone who wants to use it for their research (Orengo et al. 2020: 5). However, their project has been trained on archaeological mounds in Pakistan, giving little transferability to use in Sweden for example due to the landscape being completely different, thus it would probably need to be retrained. Introducing a collective repository of data that models can be trained on could solve the “lack of enough data” that is currently the trend.

AI for an archaeologist

One of the largest problems when involving artificial intelligence in archaeology is that most archaeologists do not have the know-how to create a model by themselves, or even understand the inner workings of said model. Given the popularity of deep learning models in the present, many archaeologists are probably wanting to adopt solutions to their work, even if it is not necessary. The result of this complex system might lead to the black box approach. The black box approach involves the archaeologist in this case, relying on a previously created classification model and a need to accept the applicability to new data without getting too concerned over the mathematics and its possible limitations (Bickler 2021: 188). Bickler also brought up Byeon and her team (Byeon et al. 2019) that suggested that their model was more reliable than the human experts that were used as comparison (Bickler 2021: 189). Given real world examples and not artificially created ones, the model might be a lot less reliable due to the fact that archaeological data has a lot of inconsistencies and variability. Most models work best and will give the most reliable results when trained on large data sets of a variety of things, such as the version of VGG16 that was used in one of the studies on bone marks (Domínguez-Rodrigo et al. 2020). The version of VGG16 that got the best score in the entire test had the training data from the "Imagenet Large Scale Visual Recognition Challenge", containing over one million images from one thousand different categories, as well as the bone marks (Domínguez-Rodrigo et al. 2020: 3-4). The goal of archaeologists is often to make new discoveries or learn more about what has already been found. For this, artificial intelligence works great and will continue to evolve and become more useful. The use of already existing models makes it a lot easier for archaeologists and saves a lot of time. However, the lack of understanding might hinder them from using the model to its full capabilities. The need for easy to use programs that have a user friendly interface might solve this issue, but getting this product to the market is another problem entirely.

5. Conclusion

In this paper I have thoroughly examined the current state of artificial intelligence in the field of archaeology with a focus on two main applications, such as image recognition for bone mark identification and satellite imagery analysis for the identification of new archaeological sites (Orengo et al. 2020; Domínguez-Rodrigo et al. 2020). The ability of artificial intelligence to process and analyze archaeological data, like the examples used in this paper, has revolutionized the field and demonstrates the immense potential and impact that artificial intelligence can have on archaeological research. With the ability to analyze and correctly identify markings on bones with a 92% accuracy and even discover new, never before seen qanat shafts, this paper has shown and proven the usefulness of artificial intelligence for the field. This paper has also analyzed the challenges and limitations that archaeologists face when using artificial intelligence. These include the need for large and reliable data sets, the existence of algorithmic bias and the complexity of the different architectures in artificial intelligence models. Given the current problems that have yet to be overcome, it is safe to state that I demonstrated in this text that artificial intelligence is not a replacement for human expertise, but rather a tool that can augment and enhance archaeological practice, as long as it is used with caution. It is also suggested that artificial intelligence can offer new opportunities and perspectives for archaeologists, allowing them to analyze and interpret data in new ways.

It is argued in this paper that a collaborative approach to the integration of artificial intelligence is needed in the field of archaeology given all its challenges. The need for different scientists to work together is necessary to ensure the mitigation of algorithmic bias and strengthen the validity of future endeavors. Even though artificial intelligence has existed for several decades in archaeology there is still much to learn as it is now the big leaps are being taken. As the computing power available to archaeologists continues to increase and through the use of cloud computers we are able to develop and use more powerful systems than ever before. However, machines can only learn from the information that we archaeologists provide them with. Therefore, applying proper training is of the utmost importance to ensure that a functioning model with accurate results is created.
Semi-automated data interpretation must also be in focus as long as there exists bias and poorly trained models, ensuring that all information generated is correct. In conclusion, this

paper has provided a comprehensive overview of the role of artificial intelligence in archaeology. It has highlighted the benefits and challenges, and underscored the need for a collaborative and critical approach to the use of artificial intelligence in this field. The future of archaeology will undoubtedly be influenced by these revolutionary methods, and this paper contributes to the ongoing dialogue on how best to navigate this exciting future.

6. References

- Tuomi, I.2018 *The Impact of Artificial intelligence on Learning, Teaching, and Education. Policies for the future,* Eds. Cabrera, M., Vuorikari, R & Punie, Y., EUR 29442 EN, Publications Office of the European Union, Luxembourg

- Cowgill, G. 1967. *Computer applications in archaeology.* Comput Hum 2, 17–23

- Horn, C. 2022. A Boat Is a Boat Is a Boat…*Unless It Is a Horse – Rethinking the Role of Typology*. Open Archaeology 8: 1218-1230. De Gruyter

- Domínguez-Rodrigo M, Cifuentes-Alcobendas G, Jiménez-García B, Abellán N, Pizarro-Monzo M, Organista E, Baquedano E. 2020 *Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications.* Scientific Reports 10

- Byeon W, Domínguez-Rodrigo M, Arampatzis G, Baquedano E, Yravedra J, Maté-González M, Koumoutsakos P. 2019 *Automated identification and deep classification of cut marks on bones and its paleoanthropological implications* Journal of Computational Science 32 36–43

- Coste A. 2018 *Project 1:Histograms CS6640 Image Processing Report*. University of Utah.

- Keras. 2023. *Release history*. https://pypi.org/project/keras/#history

- Krizhevsky A, Sutskever I, Hinton G. 2012*. ImageNet Classification with Deep Convolutional Neural Networks* p.1-9

- Simonyan K, Zisserman A 2015. *Very deep convolutional networks for large-scale image recognition*. Visual Geometry Group, Department of Engineering Science, University of Oxford p. 1-14

- Orengo H, Conesaa F, Garcia-Molsosaa A , Lobob A , Greenc A , Madellad M and Petrie C. *Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data* Proceedings of the National Academy of Sciences July 2020

- Forte M, Danelon N. 2019 *Cyber-Archaeology* Oxford Bibliographies

- Domínguez-Rodrigo M, S. de Juana, A.B. Galan, M. Rodríguez 2009. *A new protocol to differentiate trampling marks from butchery cut marks* p.1-12

- Kingma D, Lei Ba J. *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.* ICLR 2015 1-15

- Soroush M , Mehrtash A, Khazraee E, Ur J. 2020 *Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq.* MDPI

- Sharafi S, Fouladvand S, Simpson I, Alvarez J. 2016 *Application of Pattern Recognition in Detection of Buried Archaeological Sites based on Analyzing Environmental Variables, Khorramabad Plain, West Iran.* Journal of Archaeological Science: Reports, 8, p. 1-13

- Canan C, Cottrell G. 2012. *Color-to-Grayscale: Does the Method Matter in Image Recognition?* PLoS ONE p. 1-7

- Kumari R, Srivastava S 2017. *Machine Learning: A Review on Binary Classification* International Journal of Computer Applications Volume 160. p.11-15

- Alagözlü M, Zechen W, Xuetian S. 2022. *Gradient Descent and Stochastic Gradient Descent Variants, applications, and more* Università della Svizzera Italiana p.1-17

- You L. 2021. *MULTISPECTRAL IMAGE PROCESSING*. Brno University of Technology. Master's Thesis Specification. p.1-54

- QGIS Documentation 2023. *Using Raster Layers*

- Jamil A, Yakub F, Azizan A, Roslan S, Zaki S, Ahmad S. 2022. *A Review on Deep Learning Application for Detection of Archaeological Structures* Journal of Advanced Research in Applied Sciences and Engineering Technology 26, Issue 1 p.7-14

- Lui L, Wang Y, Chi W. 2017 Image Recognition Technology Based on Machine Learning IEEE Access p. 1-9

- Wayne W. 2022 Diffusion of Innovation Theory Boston University School of Public Health.

- McPherron, S, Alemseged, Z, Marean C, Wynn J, Reed D, Geraads D, Bobe R, Bearat H 2010. *Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia*. Nature 466, p. 857–860

- Cambridge University Press 2023. *Bias*. Cambridge dictionary.

- Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. *Managing Bias in AI*. In Companion Proceedings of the 2019 World Wide Web Conference, San Francisco,

- Beck, C, Jones, G. 1989. *Bias and Archaeological Classification.* American Antiquity, 54, p 244–262.+

- Bickler, S 2021. *Machine Learning Arrives in Archaeology.* Advances in Archaeological Practice. 9. 186-191.

- Turing A. 1950 *Computing machinery and intelligence* Mind 49 p.433 - 460

- Vitali V. 1989. *Archaeometric Provenance Studies: an Expert System Approach* Journal of Archaeological Science 1989 16, p.383-391

- Rogers E. 1962. *Diffusion of innovations.* The Free Press, A Division of Macmillan Publishing Co

- Casini L, Roccetti M, Delnevo G, Marchetti N, Orrù V 2021 *The barrier of meaning in archaeological data science* p. 1-5

- Noble S. 2016. *Challenging the algorithms of oppression.* Florida state university library

References - Pictures

- Cover page - Made by Philip Gonzalez in Gencraft picture generator 01/12/23
- Qanat shafts p.20. Soroush M , Mehrtash A, Khazraee E, Ur J. 2020 *Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq.* p.3
- Model depiction p.22 Soroush M , Mehrtash A, Khazraee E, Ur J. 2020 *Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq.* p.9