From Object Detection to Archaeological Object Detection

Developing a Model for Amphora Identification of a Punic Wreck site Using Object Recognition AI.

L-Università ta' Malta

UM Maritime Archaeology Programme
Research Project

Author:
Pablo Morando

Supervisors:
Dr. Maxine Anastasi
Dr. Dylan Seychell

Dissertation submitted in part fulfillment of the requirements for the degree of Master of Arts in Global Maritime Archaeology.

Department of Classics and Archaeology.
University of Malta

March, 2025

L-Università ta' Malta

FACULTY/INSTITUTE/CENTRE/SCHOOL_____Arts_____

DECLARATIONS BY POSTGRADUATE STUDENTS

(a) Authenticity of Dissertation

I hereby declare that I am the legitimate author of this Dissertation and that it is my original work.

No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

(b) Research Code of Practice and Ethics Review Procedures

I declare that I have abided by the University's Research Ethics Review Procedures. Research Ethics & Data Protection form code _____ARTS-2024-00366_____.

As a Master's student, as per Regulation 77 of the General Regulations for University Postgraduate Awards 2021, I accept that should my dissertation be awarded a Grade A, it will be made publicly available on the University of Malta Institutional Repository.

ABSTRACT

The last years have seen an increase in the use of object detection methodologies in land archaeology during surveys, the study of archaeological assemblages, reconstruction of archaeological materials, and taphonomic studies. The advances of these methods in maritime archaeology have been more limited. This study explores how AI object detection can help identify archaeological materials underwater. It aims to explain the issues that the underwater environment presents for automated detection, to bridge the knowledge gap that exists between the practical application of this computer vision technique to maritime archaeology, and to provide a practical example of its application on the underwater assemblage of Xlendi Archaeological Park, one with which to evaluate the possibilities that the use of such methodology presents for archaeological research.

We trained, classified and named a total of seventy-two detection models based on three differentiating factors. The Progressive Complexity Index (PCI) divides them into groups based on their level of complexity and the amount of archaeological information embedded in their predictive process. The Parameter of Archaeological Identification (PAI) specifies the archaeological framework used during training to teach subjective information. Finally, the models are also different in the version and size of model they use. To fulfill the goals of this project, we used these differences to interpret the results of a series of comparatives tests made on data not seen by the algorithm during training, thus recreating a real-world situation in which to evaluate the technique.

The result is the division of the models into three progressively complex groups: nature models, state models and typological models. Nature models focus on the assessment of underwater archaeological assemblages by the nature of the materials to be found in them, classifying them based on them being ceramic, litter, modern elements, or part of the natural background. On their best iterations, these achieved an average precision of identification of 87.8%. State models, on the other hand, focus on the state of preservation of those materials. Their best iteration's average precision, while lower at 75.2%, still produced very usable models on a real-world scenario. Finally, typological models focus on ceramic materials based on their typology. Their best iteration, while not being field-ready with an average precision of 61.1%, offers a lot of potential for improvement.

This dissertation has demonstrated how subjective archaeological information can be integrated into YOLO models to develop detection models tailored to specific archaeological questions. By analyzing and comparing these models, it has outlined the technique's fundamental applications, limitations, and future potential for studying underwater archaeological assemblages.

Keywords: Maritime Archaeology, Underwater Object Detection, Ceramic artefacts, Xlendi's Underwater Archaeological Park,

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, Maxine Anastasi, for her invaluable time. Her expert guidance on all matters related to ceramics, helpful insight that helped me coalesce my chaotic ideas into the page, and unending patience at enduring my first drafts were all crucial to the completion of this research.

I am equally grateful to my co-supervisor, Dylan Seychell, for his support and for helping me navigate the intersection of maritime archaeology and computer vision.

I would also like to thank Timmy Gambin for his responsibility in igniting my ideas through our conversations,My gratitude here extends to the department's whole diving team, including Karl, Dave, Maja, Charlie, Roger, Julia from Heritage Malta, and all the others who have taught me so much throughout the past year and during the last campaign at Xlendi while conveniently forgetting about my driving skills.

John Wood has to be mentioned as one of the people without whom this research could not have been possible. Not only was he the primary source behind the data I utilized, but his dedication and insights served as a genuine source of inspiration throughout my work.

A special thanks has to go to Leonie Buchele. Having met her in Malta before she left for work, her generosity in offering her time for thorough reviews and phone calls is a testament to her character. I would also like to thank Ethan Zammit. It was through assisting him with his own project that I was introduced to the possibilities of integrating object detection into maritime archaeology. The biblical-sized stream of messages that have plagued his phone since probably have him (and his phone) wishing I hadn't.

Finally, I must thank my parents, Arturo and Almudena, for the support and unwavering trust they show me every time I return home. This is for them.

TABLE OF CONTENTS

The author of this dissertation acknowledges the use of artificial intelligence tools, specifically for grammar correction and spellchecking during the writing process, and for the standardization of the bibliographical references.

ABBREVIATIONS

AI: Artificial Intelligence.

ANN: Automated Neural Networks.

AUV: Autonomous Underwater Vehicles.

CBAM: Convolutional Block Attention Models.

CPU: Central Processing Unit.

CNN: Convolutional Neural Network.

DL: Deep Learning.

DPM: Deformable Parts Model.

DSRPAI: Dartmouth Summer Research Project on Artificial Intelligence.

GAN: Generative Adversarial Networks.

GPU: Graphics Processing Unit.

HOG: Histograms of Oriented Gradients

INA: Institute of Nautical Archaeology.

LIDAR: Light Detection and Ranging.

MBES: Multibeam Echo Sounders.

ML: Machine Learning.

MNV: Minimum Number of Vessels/Individuals.

NISP: Number of identified Speciments.

NLP: Natural Language Processing.

NMS: Non-Maximum Suppression.

PAI: Parameter of Archaeological Interpretation.

PCI: Progressive Complexity Index.

R-CNN: Region based Convolutional Neural Network.

ROV: Remotely Operated Vehicles.

RPA: Robotic Process Automation.

ROME: Radar Object Modelling Environment

SIFT: Scale-Invariant Feature Transform algorithms

SSS: Side Scan Sonar.

SVM: Support Vector Machines.

UXO: Unexploded Ordnance.

YOLO: You Only Look Once—Detection model.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Contents

Artificial Intelligence (AI) was defined by John McCarthy in 1955 as 'the science and engineering of making intelligent machines' (McCarthy, 2007: 2). It is a broad concept that encompasses a multitude of applications capable of simulating human intelligence in performing tasks such as learning, decision-making, perception and problem-solving. Since its inception, AI systems have evolved alongside hardware advancements, with researchers and experts enthusiastically exploring their potential across nearly every scientific discipline.

In archaeology, AI-based methodologies date back to the 1970s, when expert systems were first introduced (Cowgill, 1967; Doran, 1970). However, it has been during the past two decades that significant advancements in both hardware and software have made AI a truly valuable tool in the field. Within maritime archaeology, AI has proven particularly beneficial in enhancing remote sensing through robotics (Drap et al., 2015; Kamal et al., 2024) and underwater site analysis through computer vision techniques like photogrammetry (Radić et al., 2019),[1] while also reducing the time experts spend on repetitive data analysis. Despite these advantages, maritime archaeology remains a relatively young discipline, and AI-driven methodologies, while innovative and full of potential, are not yet widely embraced.

One area where AI holds great promise is in the application of automated object detection methods on underwater archaeological assemblages.[2] These, machine learning (ML) models able to localize and identify objects within an image or video, offer a powerful means of extracting information from underwater sites and optimizing archaeological interventions.[3] However, their

[1] **Photogrammetry,** in maritime archaeology, is a computer vision technique used to generate 3D measurements and models from sets of overlapping 2D underwater photographs. Using special software, the 3D data is used to record sites and study them once at the surface (Radić et al., 2019: 45).

[2] **Automated object detection** is a computer vision (type of AI) technique that enables machines to automatically identify, locate and classify objects within a set of images or a video (Yang et al., 2023: 2-3).

[3] **Machine learning** is a branch of AI that encompasses a wide array of tools. It can be generalized as the elements of AI programming that allow software applications to become more accurate at predicting outcomes without the requirement of explicit changes in the code (Argyrou and Agapiou, 2022: 5999).

application does not come without challenges: though advancing rapidly in capability and ease of use, detection methods remain conceptually complex and require a user with specialized knowledge. Hence, despite their potential, they are often underutilized outside of larger research projects with the luxury of means.

In this context, a family of detection models called YOLO (You Only Look Once), has emerged in the last years as an interesting candidate to become an archaeological tool of widespread use thanks to a combination of simplicity, autonomy and precision.[4] These models allow researchers to quickly process vast amounts of data by teaching the models the processing parameters through a small amount of that same data. The result is a trained model that can be used repeatedly on sets of data with similar parameters (Figure 1).

By using these simplified object detection models, this dissertation aims to demonstrate their value in the analysis of underwater assemblages and characterize their use for potential future applications. These include streamlining labor-intensive processes, enhancing data accuracy by reducing human error, increasing overall efficiency, opening new avenues for research, and more.[5]

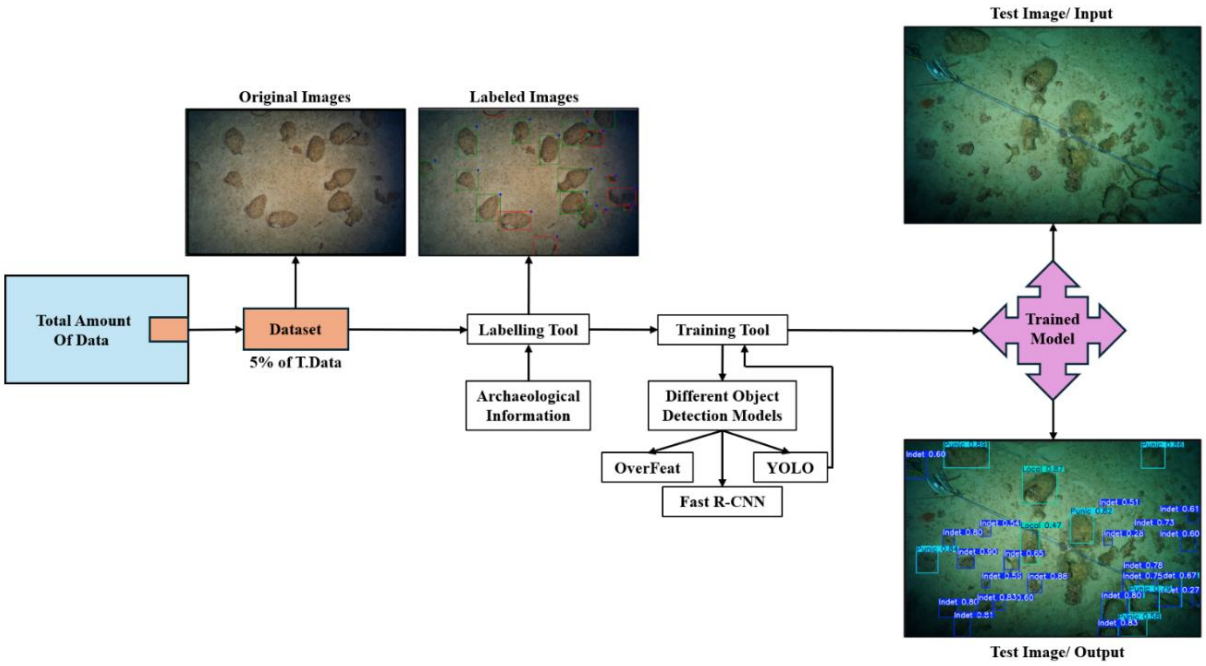


Figure 1. Diagram showing the development process of a detection model in the context of maritime archaeology. A portion of a large amount of data (dataset) is used to introduce archaeological information into the images through a labelling tool. After choosing a detection model, the labelled dataset is used to create a trained model that can be used to process the entirety of the original data. This process is explored in detail in Chapter 3.

[4] There are many types of detection models that look at the same problem (what objects do you have on an image, and where) in different ways and using different ML tools (literary review p. 10-12, Appendix I-p.153).
[5] More information on the general applications of YOLO models can be found on the developer's online repository (https://github.com/ultralytics/ultralytics).

1.1 Aims and Main Tasks

In 1970, James Doran, a computer scientist with a strong understanding of archaeology, identified what he deemed the most significant challenge that archaeologists face when encountering new technologies with potential applications. He observed that in such situations, archaeologists must choose "between losing touch with some of the most interesting and important developments in their subject, or embarking upon a course of study which nothing in their training has prepared them for" (Doran, 1970: 289). Things have changed since then, but in the case between maritime archaeology and computer vision,[6] the assessment remains relevant.

This dissertation aims to bridge the knowledge gap between maritime archaeology and widely used object detection tools. It does so by first defining and contextualizing this computer vision technique within the archaeological field and then assessing the challenges and benefits of applying simple 2D automated detection to an underwater site. The study focuses on the ceramic assemblage from Xlendi Underwater Archaeological Park (Gozo, Malta), using the site as a case study for a series of comparative experiments to assess the viability of this computer vision technique for the study of underwater archaeological assemblages. In addition, and through various means that permeate the whole project, this research aims to test the hypothesis that, unlike most archaeometric techniques, the full potential of object detection in underwater archaeology can only be realized when its users are archaeologists themselves.

We pursued these objectives through the following tasks:

-An explanation of the nature and functioning of object detection methodology that focused on its core principles and historical applications in maritime archaeology. Rather than serving as a step-by-step guide for implementing detection methods, this research intended to bridge disciplines, offering an accessible introduction that equips archaeologists with the foundational knowledge needed to adapt and benefit from this technology.

-The creation of a typological chart for the ceramic materials from Xlendi Archaeological Park was a task that served two purposes: First, it established a reference framework for the future

[6] **Computer vision** is a subset of AI whose pattern identification algorithms enable computers to interpret and understand visual information from the world. They do so by recognizing the patterns in the visual information provided to them with the help of machine learning. Both photogrammetry and object detection are archaeologically-used techniques that belong to this subset (Argyrou and Agapiou, 2022: 5999).

characterization of the site, and second, it contributed to the construction of a complex classification system with which to test the capabilities of YOLO detection models.

-The creation of a theoretical framework from which to implement and adapt the methodology to its use in archaeology. Since there is scarcely any precedent of automated detection being used to study underwater assemblages, the development of such a (testing) framework aimed to smoothen out the adaptation process of detection models in maritime archaeology, facilitate the interpretation of their output (always from an archaeological perspective), and showcase the methodology's potential.

- To assess the capabilities and limitations of YOLO detection models on underwater assemblages, we conducted several experiments entailing the design and training of multiple detection models. These were tested in a manner that best simulates a specific set of easily met real-world conditions.

-An evaluation of the worth of this tool in the context of underwater assemblages was conducted through a discussion interpreting the experimental results in relation to the study's objectives. This discussion also explored the broader significance of the research, identifying aspects of the experiment that require refinement for future studies and examining the potential for implementing detection methodologies in underwater archaeology for non-experts in computer vision.

2. LITERATURE REVIEW

Contents

2.1 Introduction to Literature Review

Relying on foundational understanding of the relationship between underwater archaeology and computer vision technology, this dissertation explores the applications and challenges of using detection methods for underwater assemblages. While aiming to bridge two fields that are fundamentally distinct, this research inevitably engages with technical concepts, making it inherently interdisciplinary. To aid in understanding these technical aspects, the following section provides background on the intersection of maritime archaeology and object detection techniques, situating this research within the broader context of their application in the field of archaeology.

2.2 Automated Object Detection. A Computer Vision Technique

"The value of any innovation lies in its own merits, and the merits of which it replaces" (Clarke, 1968: 1).

Written at a time when his research field faced the introduction of computer science, this quote from pioneering archaeologist David L. Clarke has at its core a question still pertinent to us fifty years later: What do computer vision methods such as object detection aim to replace in archaeology?

Since the discipline's inception, archaeologists have consistently sought more accurate ways to capture and understand data, aiming to unravel the complex relationships between humanity and the material culture left behind. From field drawings and survey chains to the use of laser scanners and archaeometry, technology has changed archaeological science. The latest link in this chain is

computer vision.[7] Today, with the aid of software, computers can see, analyze, and interpret vast amounts of data far more quickly than a human ever could (Drap et al., 2015; McCarthy et al., 2019: 5, Bickler, 2021: 187). This is particularly advantageous in maritime archaeology, where data acquisition faces additional challenges due to the underwater environment. In such a hostile setting, automated detection offers a clear set of benefits:

-Time efficiency: Archaeological research produces large amounts of data. Detection models are, in essence, data-processing tools. Whereas these large amounts of data produced during archaeological fieldwork would normally need to be manually reviewed by the researcher, the ability to process them with automatic techniques allows the researcher to focus on the interpretation stage of a project (Bickler, 2021: 186).

-Resource efficiency: In maritime archaeology, due to the high cost of conducting any sort of intervention in the underwater environment, it becomes critical to develop methods that allow the researcher to achieve results while reducing costs (Paraskevas et al., 2023: 1). Computer vision technologies use open-source models; and aside from requiring powerful graphics processing units (GPU), they require no additional hardware. In this way, they offer a cost-effective solution for researchers, enabling them to obtain results without the need for additional interventions, technical support, or more expensive methods (Zammit et al., 2024: 4123).

-Alternate perspective: On their own, computer vision techniques generate unique visual data that can offer a fresh perspective in the interpretative stage of an archaeological project. Thus, automated detection methods are not only tools for data management, but also potential sources of new knowledge and avenues for further investigation (Masita et al., 2020: 1; Character et al., 2021: 2).

-Heritage preservation: Computer vision techniques generate digital records of archaeological sites and avoid the damage caused by physical interventions. They are non-invasive, thus helping preserve archaeological sites from excavation and ensuring they are left for future generations of study and protection (McCarthy et al., 2019: 1-2; Character et al., 2021: 2; Yang et al., 2023: 1).

While these benefits apply to maritime archaeology, the first three are also relevant across various scientific disciplines, commercial interests, and other endeavors in the underwater

[7] Computer vision (p.3).

environment. These parties—often economically incentivized—must manage large volumes of data too. It is for this reason, along with the increasing commercial interest in underwater exploration, that the role of computer vision techniques has expanded significantly in recent years.

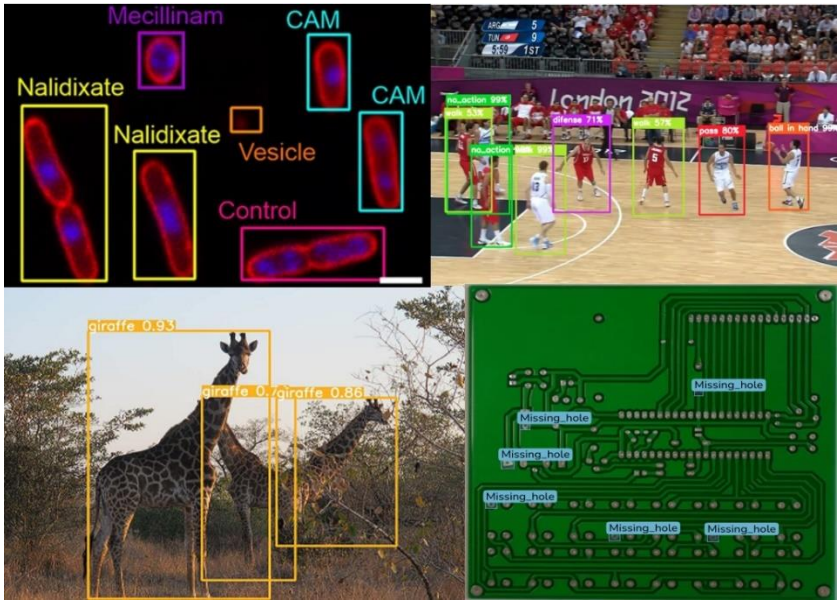


Figure 2. Examples of open-source object detection being applied to microbiology (top left, Spahn et al 2022), sports analytics (top right), wildlife monitoring (bottom left, Zhang et al., 2023), and manufacturing and quality control (bottom right).

2.2.1 Automated Detection in the Underwater Environment

Object detection algorithms are among the most powerful computer vision tools. Implemented through detection models, these programs can identify features of interest in various types of imagery by recognizing the unique visual patterns in which they are represented (Kypraios, 2012: 9; Chollet, 2021: 94). In scientific research, they enable users to automatically process large volumes of visual data. However, their use also comes with certain caveats which we will address later on, including their technical complexity, their requirement for high computing power, the potential biases introduced on their predictions during training, and their hunger for data.[8]

The difference between applying detection algorithms on land (Figure 2) versus the seabed lies in the additional challenges posed by the conditions that the underwater environment imposes both

[8] For training purposes, detection algorithms require a vast amount of pre-labeled, classified data to produce robust prediction patterns. This data must be of high quality. Moreover, without an adequate quantity of it, the detection model is of no use on a practical scenario (Tuomi, 2017: 12; Paraskevas et al., 2023: 3).

on the data acquisition process and the target objects. Yang et al. (2023) reviewed these challenges from an archaeological perspective:

-Poor quality of underwater images: Due to the methods used for capturing them—whether by hand, remotely operated vehicles (ROV),[9] or autonomous underwater vehicles (AUV)[10]—it is often difficult to maintain constant and acceptable levels of illumination in underwater imagery. The reason is the various factors that affect image quality and need to be adjusted for. Some examples of these factors include low visibility from inadequate lighting, suspended water particles, marine life, color deviations, lack of contrast, or the differential absorption of light by water depending on wavelength (Shortis, 2019: 16; Zhang et al., 2022: 3997; Zammit et al., 2024). These factors can also be expected to vary across different bodies of water and be influenced by meteorological conditions.

-Difficulty obtaining the data: Unlike aerial optical images, underwater optical imaging remains inconsistent despite the latest advancements in cameras and lenses designed for such environments (Shortis, 2019: 12-17; Hu et al., 2022). Whether data are gathered by underwater vehicles or human divers, the underwater environment always presents logistical complications. The high cost of organizing data-gathering missions, the inherent risks associated with SCUBA diving, or the unpredictable effects of weather can be mentioned as some examples in this regard.

Archaeological data: The mismatch between the characteristics of detection methods and the realities of archaeological sites is a central issue for the continued application of automated detection on underwater assemblages. Two wrecks, even if both consist solely of amphorae, will never be identical. Archaeological materials often appear in clusters due to depositional processes, and we often find ceramic items fragmented into multiple sherds. In most cases, these materials can be covered by sediments or marine life, which leads to them being easily mistaken for other natural elements such as rocky reefs (Chen et al., 2022: 1; Zammit et al., 2024: 4122). These challenges compromise the operation of detection models.

On this line, another complication stems from the fact that models need to be trained on a large set of images. While this is not an issue for large datasets such as those used in site detection or

[9] **Remotely Operated Vehicles**, tethered to a boat, can fit lights and cameras to direct video feed directly to the surface (Paraskevas et at al., 2023: 2). This video feed can be used as data for a real-time detection model.

[10] **Autonomous Underwater Vehicles** can also be fitted with cameras that take photos of the seabed and be programmed to follow specific paths to cover large areas with 2D imagery (Nayak et al., 2019: 2-4).

for Xlendi Archaeological Park, smaller and less varied sites/assemblages can face difficulties and produce less robust models.[11] A useful analogy to illustrate this issue is that of the researcher trying to obtain wood samples from a single splinter.

What we can take from these challenges is that, when it comes to the use of automated object detection, the importance of data cannot be overstated. Efforts to develop detection models that address key challenges—such as dataset size and variety, image distortion, object scaling, occlusion, and object overlap—while achieving an acceptable level of accuracy in the results have been, in fact, the driving force behind almost every project using detection in the underwater environment over the last thirty years.

The adaptation of detection algorithms to the specific conditions of underwater data gathering began in the late 1990s and continued throughout the first decade of the 21st century (Figure 3).[12] During this period, maritime archaeology was still in its formative years as a distinct discipline, and much of the early work on underwater detection came from fields with stronger incentives for technological development, such as marine biology, private-sector industries, and oceanic engineering (Barngrover et al., 2014; Moniruzzaman et al., 2017; Paraskevas & Kavallieratou, 2023). [13]

This adaptation process can be divided into two distinct stages. The first stage predates the rise of conventional ML[14] and is formed by projects concerned with developing AI tools capable of recognizing submerged objects. For instance, in 1989, Aull and Gabell introduced the Radar Object Modelling Environment (ROME), an interactive workstation that used algorithmic feature extraction tools to identify objects in radar images. While their study did not specify a particular type of target, later work by Johnson and Deaett (1994) applied a similar AI-based feature extraction approach to automatically recognize ocean-bottom toxic waste deposits. Other researchers explored different methods, such as Daniel et al. (1998) and Chew et al. (2007), both

[11] A detection model is more **robust** if trained on a heterogeneous set of images with distinctive perspectives, lighting conditions and object variety that will more easily withstand its application on a real-world scenario (Paraskevas et al., 2023: 3-4).
[12] More information on the evolution of the integration of AI into archaeology can be found in Appendix I (p.153).
[13] Some **early** examples of detection being applied to the maritime environment in the commercial, engineering and military sectors: In matters of Unexploded Ordnance (Schweizer et al. 1994; Shin et al., 1997), biota studies (Ren et al., 1996; Cutter et al., 2015; Zhang et al., 2022), seabed mapping (Schweizer and Petlevich, 1989; Zerr, 1991; Wen et al., 1995; Carmichel, 1998; Boulinguez and Quinquis, 1999; Lowe, 1999), oil and gas exploration (Gehrig and Becker, 2004) or infrastructure monitoring (Almeida and Dhanasekar, 1995; Crisp and Watson, 2001).
[14] Machine Learning (p.1).

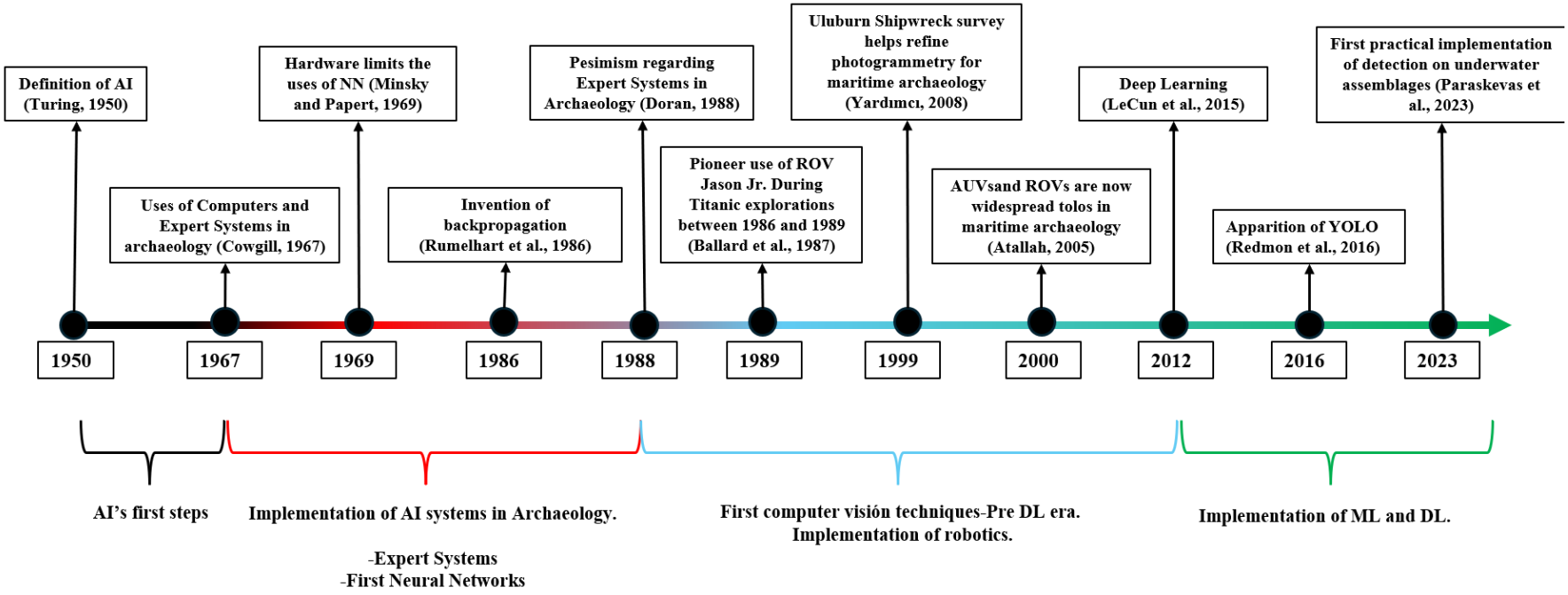


Figure 3. Timeline of the evolution of AI in Maritime Archaeology.

of whom worked with Side-Scan Sonar (SSS) imagery.[15] Daniel et al. sought to match sonar images to objects by analyzing the geometric relationships between shadows on the seabed. Chew et al., for their part, introduced an illumination-balancing technique that significantly improved contrast and object identification.

The process of adapting computer vision to the underwater environment entered its second stage with the advent of Deep Learning (DL).[16] Deep learning revolutionized object detection, drastically increasing the number of algorithms capable of recognizing intricate patterns in underwater data (LeCun et al., 2015: 436). As a result, the early 2000s brought a surge of publications that followed the same blueprint—authors developing detection models to deal with the problems they encountered during their projects with the underwater environment. Research efforts expanded, with numerous studies exploring different code architectures and optimization techniques (Figures 4, 5) as well as approaches to solving the (adaptation) problem. For example, Chen et al. (2020) utilized DL-based object detection and instance segmentation methods in marine botany research.[17] Their novel sample-weighted super network (named SWIPENet) significantly improved the detection of small objects in images affected by wavelength-dependent absorption and scattering—common sources of noise that degrade contrast and blur object boundaries (Akkaynak and Treibitz, 2018: 6723). Other researchers introduced additional refinements. Zeng et al. (2021) proposed a region-based convolutional neural network (R-CNN)[18] integrated with

[15] **Side-scan sonar** is a remote sensing technology used to create detailed images of the seafloor reading the reflections caused by emitted soundwaves (Nayak et al., 2019: 3).

[16] **Deep learning** is a specialized field of ML. It differs from normal ML in its architecture and the ways in which it processes the data. DL is very complex in this regard. Where ML excels at working with smaller data sets comprised of tables of numbers, DL was developed to deal with larger and more unstructured data sets such as photos and videos (LeCun et al., 2015, Akinosho et al., 2021; Chollet et al., 2021). Because of this, it requires more data, more processing power, and more time. However, it is also very powerful and has become very popular as a way to tackle complex tasks like image and language recognition. It is particularly useful in archaeology because of its ability to identify many different morphologies and orientations of the same features (Character et al., 2021).

[17] **Instance segmentation** is a computer vision technique sister to object detection. It is also used to identify and classify objects in 2D imagery, but it does it by analyzing the images pixels to separate the targets from the background (Chen et al., 2024: 2).

[18] **Convolutional neural networks (CNN)** are a type of DL model designed for processing structured data, such as images—similar to YOLO. They are widely used in computer vision applications, including object detection and image classification. A notable variation, **Region-based CNN** (R-CNN), introduces region proposals to improve detection accuracy. Over the years, R-CNN has been refined through Fast R-CNN and Faster R-CNN, significantly enhancing efficiency. More details on their inner workings are provided in Chapter 3, while extensive references can be found in the bibliography (Sermanet et al., 2013; Girshick et al., 2014; Girshick, 2015; Ren et al., 2017; Caspari and Crespo, 2019).

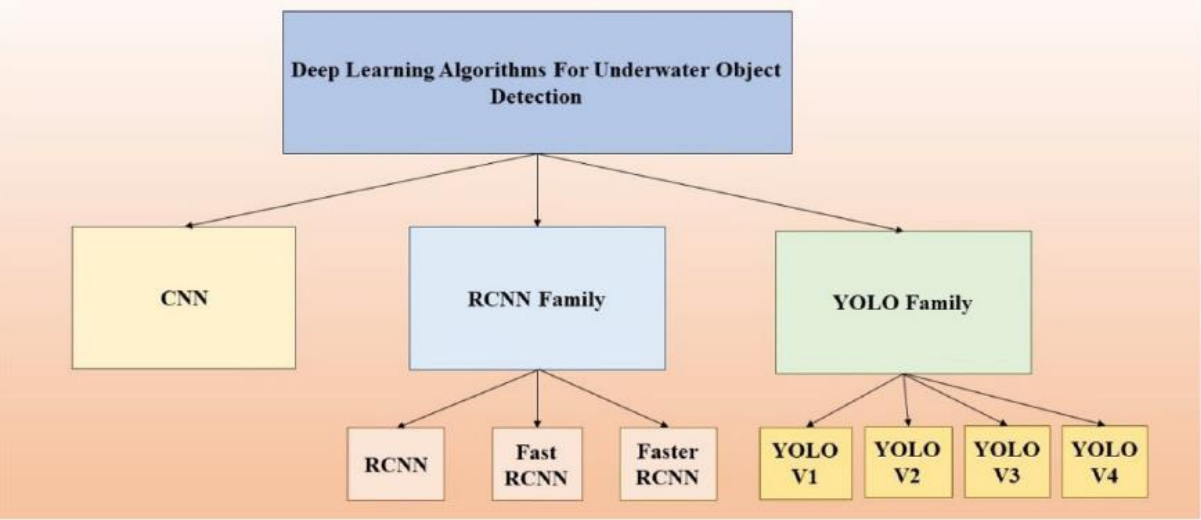


Figure 4. Categories of deep learning algorithms for underwater object detection (Fayaz et al., 2022: 20822).

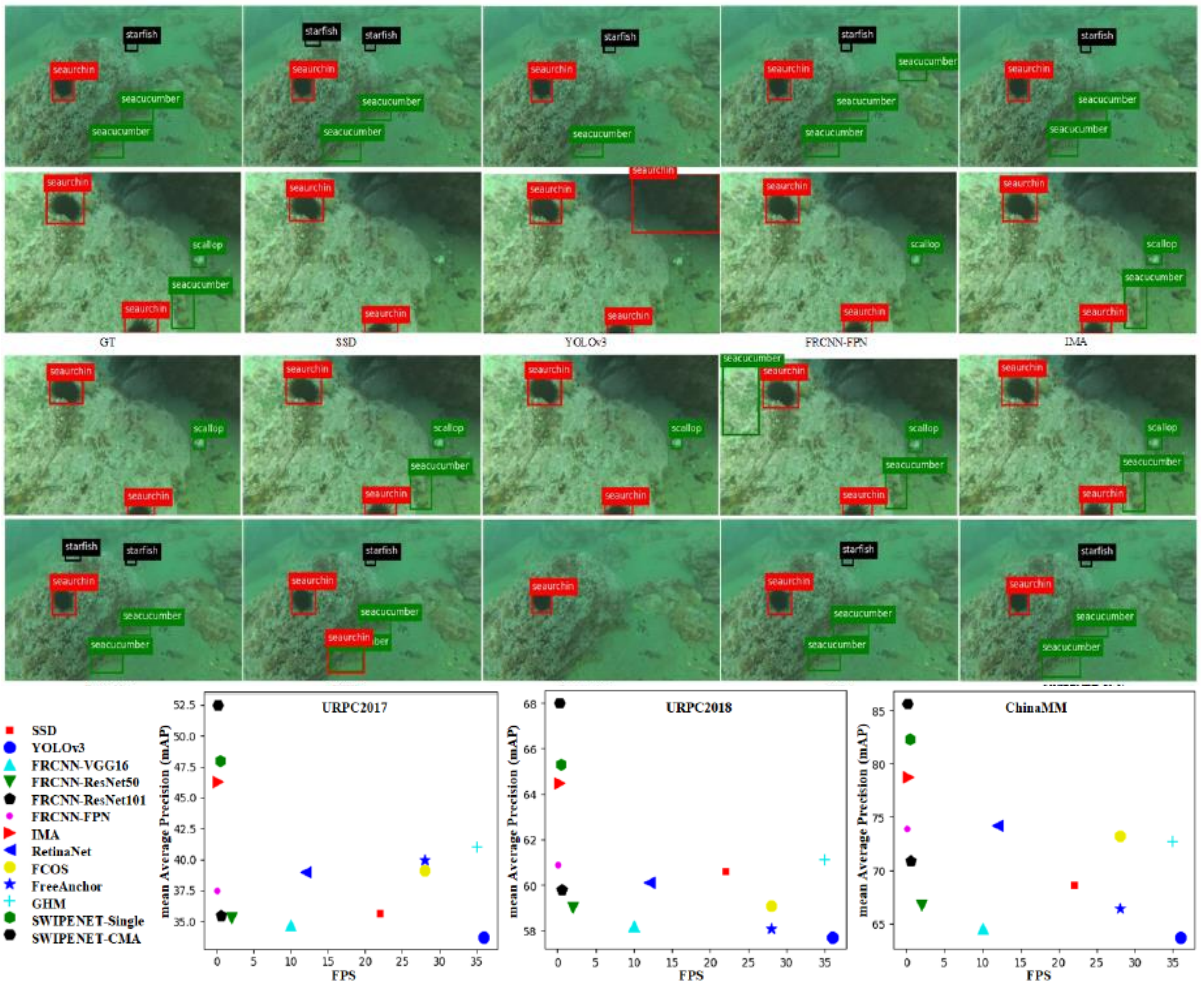


Figure 5. Visualization of different object detection models applied to noisy underwater images. Their respective speeds (Running time vs Mean Average Precision) are compared in the charts below (Chen et al., 2020: 12).

generative adversarial networks (GAN)[19] to improve object detection in cluttered, overlapping environments. Yan et al. (2022) incorporated Convolutional Block Attention Modules (CBAM) into object detection models to focus more on relevant feature information and reduce the impact of background noise.[20] Song et al. (2023) enhanced R-CNN with Boosting R-CNN, a mechanism designed to mitigate image occlusion issues. El Rejal et al. (2023) introduced the WGH-model, which was specifically designed to counteract image distortion problems in underwater environments.

At this stage, with all these publications presenting different models and processes aimed to minimize the issues presented by the underwater environment to computer vision methods, numerous comparative studies emerged in parallel evaluating and comparing their performance under varying conditions (Wang et al., 2019; Masita et al., 2020; Lei et al., 2022; Fayaz et al., 2022; Raavi et al., 2023). These reviews categorized underwater object detection models into three primary families (CNN-based, R-CNN-based, and YOLO-based architecture), proving that each one has distinct advantages and limitations depending on the environmental conditions and project requirements (Figure 4). Crucially, these studies also revealed that no single model outperforms all others in every scenario. For instance, a detection model optimized for the murky depths of a Baltic port may be less effective in the clearer waters of a Maltese bay. Hence, selecting the most appropriate objective model becomes one hundred percent a case-specific decision, particularly when maximizing detection accuracy is the primary focus of a project intending to use the methodology.

Finally, it is essential to highlight the recurring mention of YOLO in recent publications. The fact that many authors consider it an especially promising group of models due to their low computational complexity, high speed, adaptability, and suitability for real-time applications will be particularly relevant when discussing the most suitable detection options for archaeological research (Fayaz et al., 2022: 20909).

---

[19] **Generative adversarial networks** (GAN) are another type of DL model that uses two competing neural networks (see Appendix I, p.150.) to create realistic data. In the context of object detection for archaeological assemblages, these have a use for image enhancement (Goodfellow et al., 2014: 140).

[20] **Convolutional Block Attention modules** (CBAM) are a supporting mechanism designed to help CNNs focus on the most important parts of an images, thus emphasizing relevant information over irrelevant details (Woo et al.2018: 4).

2.2.2 Automated Detection in Maritime Archaeology

In 2015, Drap et al. presented an AI-based method to map shipwrecks using 3D photogrammetry and instance segmentation (Drap et al., 2015). That same year, Jaklic et al. developed a similar technique to create volumetric models of ancient Roman cargoes using 3D point cloud technology (Jaklic et al., 2015). These are just a few examples of the growing variety of computer vision techniques used to explore, record, and visualize underwater archaeological sites (Drap and Long, 2001; Menna et al., 2018; McCarthy et al., 2019; Drap et al., 2019). Though less commonly used in maritime archaeology, object detection is one of those techniques.

Most applications of object detection in maritime archaeology focus on site detection, a critical step for underwater research and heritage management (Character et al., 2021: 1758). The process begins with seabed imagery acquisition through geophysical data collection systems such as SSS,[21] Multibeam Echo Sounders (MBES),[22] or airborne LiDAR (Character et al., 2021: 1759).[23] Detection models are then used to analyze these data to identify potential archaeological sites. One of the earliest attempts was by Louis Atallah in 2005, who applied an algorithmic saliency-based detection method to analyze different materials placed on the seabed at Smelt Mill Bay, Belfast. Even before DL[24] object detection emerged, Atallah recognized the potential of computer vision for underwater archaeological surveys utilizing geophysical data (Atallah, 2005: 292). Subsequent studies have validated this notion (Plets et al., 2011; Gambin, 2014; Nayak et al., 2019: 3-4; Zhu et al., 2019; Character et al., 2021: 1758; Argyrou and Agapiou, 2022: 2-8).

Remotely sensed imagery acquired through sonar is highly valuable in maritime archaeology, as geophysical imagery—even when captured underwater—does not pose significant challenges for detection techniques. However, its applicability is largely limited to site detection. Fortunately, as previously discussed, geophysical imagery is not the only type of archaeological data suitable

[21] Side-scan Sonar (p.11).
[22] **Multibeam sonars** read the reflections on sounds emitted on a wide pattern, measuring the time it takes for the soundwaves to travel to the seabed and return to produce high-resolution bathymetric data (Brown et al., 2019: 127).
[23] **Airborne Light Detection and Ranging or LIDAR** is a remote sensing technology that uses laser pulses to map the earth's surface from an aircraft or drone. In archaeology, it has been widely use as a tool to ascertain the characteristics of the terrain through dense vegetation (much like multibeam on the sea). Some examples include the localization of Mayan sites on the Mexican Yucatán peninsula (Gallwey et al., 2019; Magnoni et al 2016; Ringle et al., 2021).
[24] Deep learning (p.11).

for automated detection. Optical images, in contrast, offer a wider range of applications.[25] In terrestrial archaeology, for instance, detection models are widely used to analyze aerial photos and satellite imagery for site identification and artefact classification (Verschoof-van der Vaart, 2020; Orengo et al., 2020; Brandsen and Lippok, 2021; Guyot et al., 2021). Various other applications exist for taphonomy (Byeon et al., 2019; Dominguez-Rodrigo et al., 2020),[26] epigraphic recognition (Heenkenda and Fernando, 2020; Tomasella et al., 2024), archaeometry (Sun et al., 2020), conservation (Willet, 2019; Camara et al., 2023; Molua, 2024), and artefact identification (Benhabiles and Tabia, 2017; Hein et al., 2018; Tyukin et al., 2018; Itkin et al., 2019; Anichini et al., 2021). In addition, the theoretical framework, implications, and best practices for these methodologies have also been well-explored (Lake, 2014; Davis; 2020; Argyrou and Agapiou, 2022; Orengo et al., 2019; Fiorucci et al., 2020; Bickler, 2021; Jamil et al., 2022; González, 2024; Clavert and Gensburger, 2023).

Maritime archaeology, by contrast, lacks a similarly established framework for optical image-based detection. This disparity likely arises from the field's relative youth compared to terrestrial archaeology[27] and the previously mentioned challenges associated with adapting land-based techniques like detection to the underwater environment. As a result, a significant gap remains in the literature regarding the potential applications of optical image-based detection in maritime archaeology.

Between 2010 and 2020, only Drap et al. (2019) and Pasquet et al. (2017) attempted to apply object detection in maritime archaeology beyond site detection. They developed a DL approach for detecting and recognizing submerged objects using orthomosaics,[28] leveraging transfer learning to compensate for their small training datasets.[29] Testing their model on an amphora assemblage at 44m depth, they achieved a 90% positive identification rate (Drap et al., 2019; Pasquet et al., 2017; Character et al., 2021: 1759). More recent studies have sought to build on

[25] By **optical images** we understand visual representations formed by light rays interacting with an optical system such as a camera.
[26] **Taphonomy** is the science that involves the study and interpretation of the marks that postdepositional and postmortem processes leave on organic remains across different environments (Bahn and Renfrew, 2011: 292).
[27] If we use the academic search engine Scopus, the results decrease from 70.000 to 2.000 when searching on the IEEE Xplore for articles with the keywords “archaeology” and “marine/maritime/underwater archaeology” respectively.
[28] **Orthomosaics** are high resolution, geometrically corrected images that overview and area accounting for distortions caused by the conditions and the terrain. They provide scale-accurate representations of the area under survey.
[29] **Transfer learning** is a DL technique consisting of starting a project from a model that has already been trained on a large and complete dataset; and then fine-tuning it on a smaller but more specific dataset that would have otherwise lacked size for the training process (Kolar et al., 2018: 58-59; Fayaz et al., 2022: 20874; Casini et al., 2023: 1).

their success. Yang et al. (2023) underlined the scarcity of research on automated detection methodologies in maritime archaeology and identified key challenges posed by submerged assemblages.[30] In addition, their study introduced a novel variation of a detection model designed to overcome these challenges, mirroring the efforts of the disciplines where detection techniques were being adapted to address the same data-collection limitations. While their approach optimized accuracy within a specific context, it lacked scalability: According to our hypothesis, the variability of underwater environments and artefact conditions means that prioritizing maximum accuracy for a single assemblage would require developing a specialized detection process for each new project, restricting broader applicability in maritime archaeology.

There are only three other publications that implement automated detection in underwater optical images of archaeological assemblages. Unlike the studies mentioned earlier, these projects largely rely on models from the YOLO family rather than developing custom models supported by additional elements. YOLO (discussed extensively in Chapter 3) was designed to be a simpler, faster, and more versatile detection algorithm capable of operating in diverse contexts (Redmon et al., 2016: 2; Bochkosckiy, 2020: 10; Terven et al., 2023: 1680). This means that in a way, its operation[31] directly addresses the concerns about the inefficiency of developing ad-hoc models for every cybernetics-archaeology project that existed in the past that are very present in the theoretical framework behind this dissertation (Cowgill, 1967: 332).

The first of these studies, conducted by Paraskevas et al. (2023), integrated one of the latest YOLO versions (YOLOv8) into an ROV to assist in the identification of ceramic sherds on the seabed. Leveraging YOLO's speed and computational efficiency, their goal was to minimize the time spent detecting material concentrations and reduce diver intervention while maintaining acceptable accuracy levels (Figure 6).

In their experiments, Paraskevas et al. used a dataset of 2,531 annotated images and determined that 150 epochs[32] provided the optimal balance for training models on underwater assemblages. Additionally, after testing the 'nano', 'small', 'medium', and 'large' versions of YOLOv8, they found that the 'small' variant offered the best trade-off between speed and precision for

[30] Challenges posed by the maritime environment on 2D imagery (p.8).
[31] By **operation**, we mean the pipeline and mechanisms within the algorithm (i.e. the way the algorithm works).
[32] An **epoch**, in DL and ML, corresponds to one complete cycle through the entire training dataset (i.e, during one epoch, the model goes through the training set once, updating its weights based on its predictions against the validation set). See Paraskevas et al (2023:3-6).

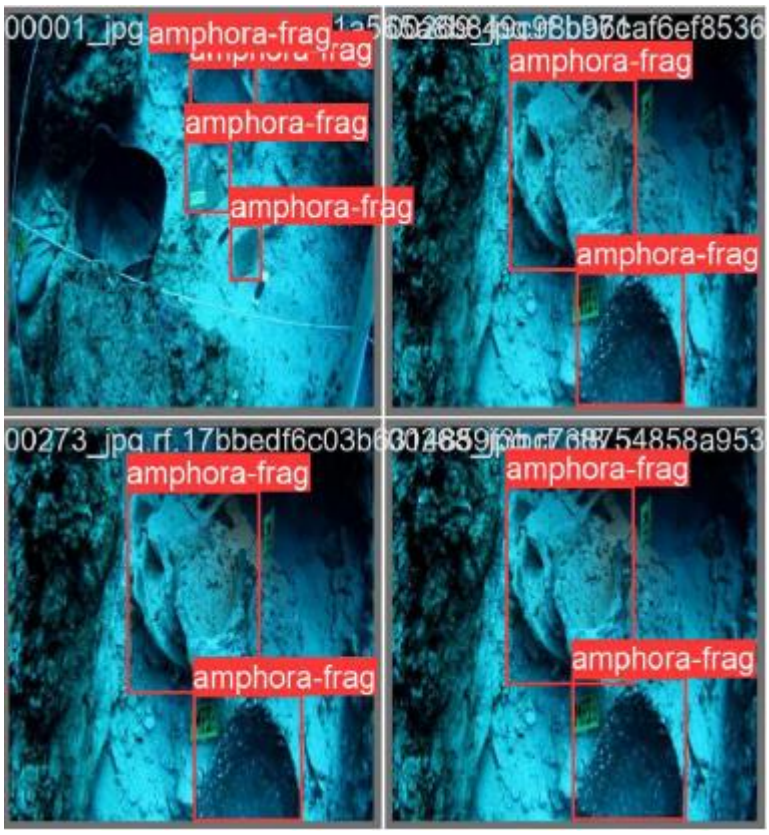


Figure 6. Sample detections from the evaluation video shown in four different screenshots of Modi Island Project, Greece (Paraskevas et al., 2023: 3).

archaeological applications. Another key contribution of their study was the discussion of parameter tuning, particularly the adjustment of confidence thresholds based on the detection task at hand. For instance, when verifying the typology of a specific amphora, a higher confidence threshold ensures that only the most likely to be successful detections are considered, improving precision for well-defined objects. Conversely, when broadly identifying potential finds, a lower confidence threshold increases sensitivity at the expense of more false positives (Paraskevas et al., 2023: 5). These insights were invaluable in shaping the experiments conducted in this dissertation, providing both a time-saving foundation and a baseline for our initial model designs. Moreover, Paraskevas et al.'s work established a performance benchmark. Their study represents the first example of a YOLO-trained model successfully identifying submerged archaeological material in a real-world scenario. It is because of this that replicating their success served as the starting point for our own experiments.

Paraskevas et al.'s study also contributed by emphasizing the importance of dataset quality and outlining the best practices for dataset composition in archaeological contexts, all of which informed the design of our experiments. This includes their approach to evaluating results, the key metrics they prioritized for assessing an archaeology-focused detection model (Table 8), and the way they presented their findings. For instance, after observing their focus on Average Precision and Average Recall,[33] we decided to adopt these specific metrics while omitting others like processing speed, which did not align with the scope of this dissertation.[34]

The second study, published in early 2024 by Kamal et al., applied YOLO to data from the Phoenician shipwreck at Xlendi Bay, Malta—a site located at a depth of 115m (Drap et al., 2015, Figure 7). Rather than using YOLO for detection alone, this project incorporated it into a broader workflow. The model was applied to the 30,000 photogrammetric images captured during the site survey (Figure 8), serving as a foundational step for a subsequent 3D instance segmentation process (Figure 9).

The third relevant study was conducted by Zammit et al. (2024), who developed a method for archaeological object detection through bidirectional photogrammetric fusion. This study is particularly pertinent to us, as it originated from the same department as this dissertation and utilized similar data.[35]

With the collaboration of AI specialists, Zammit et al. extended beyond the scope of this study, showcasing aspects of the full potential of this methodology. Their five-step approach incorporated a special technique to account for variations in light absorption at different depths, enhancing the precision of the YOLO models based on depth information. They further advanced their methodology by projecting the detection results from the 2D model onto a 3D orthomosaic of the site, effectively merging detection data with photogrammetric visualizations (Figure 10).

These last two studies are relevant to our research in a different way than the work of Paraskevas et al., which serves as the primary reference for this dissertation. Both studies by Kamal et al. and Zammit et al. demonstrate the potential of YOLO models when integrated with additional DL and

[33] Appendix IV (p. 209) for a review of the metrics.
[34] This is because processing speed depends on the equipment used for the experiments. Our goal was to present results achieved with mid-level hardware available to everyone and not specialized equipment.
[35] The project uses data from some of the photogrammetric runs we used for this dissertation, albeit at a much smaller scale (864 images). It is worth nothing that for Zammit et al.'s project, the archaeological data in the images was processed by me as domain expert. In fact, the idea for state of preservation models and the concept of PAI sparked from my role in that project.

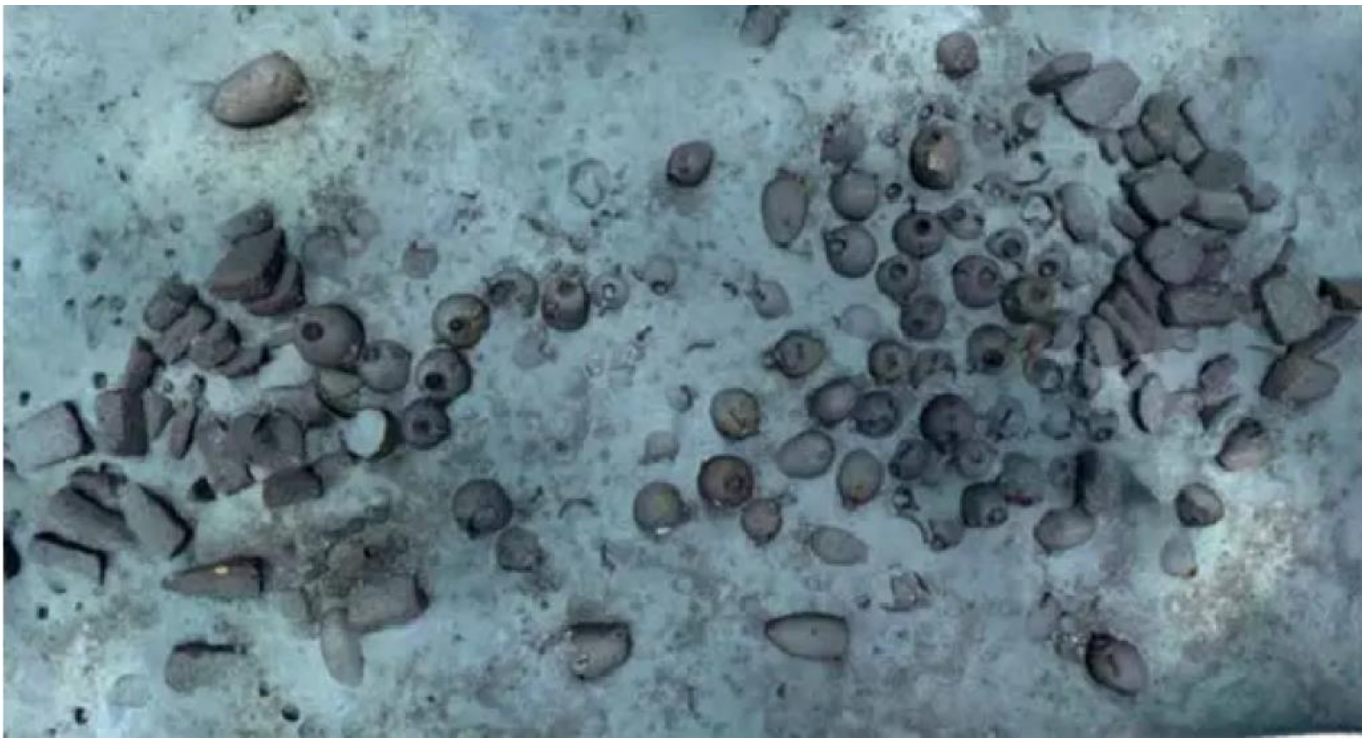


Figure 7. Orthomosaic image of the Xlendi wreck generated by photogrammetry (Kamal et al., 2024: 1333).

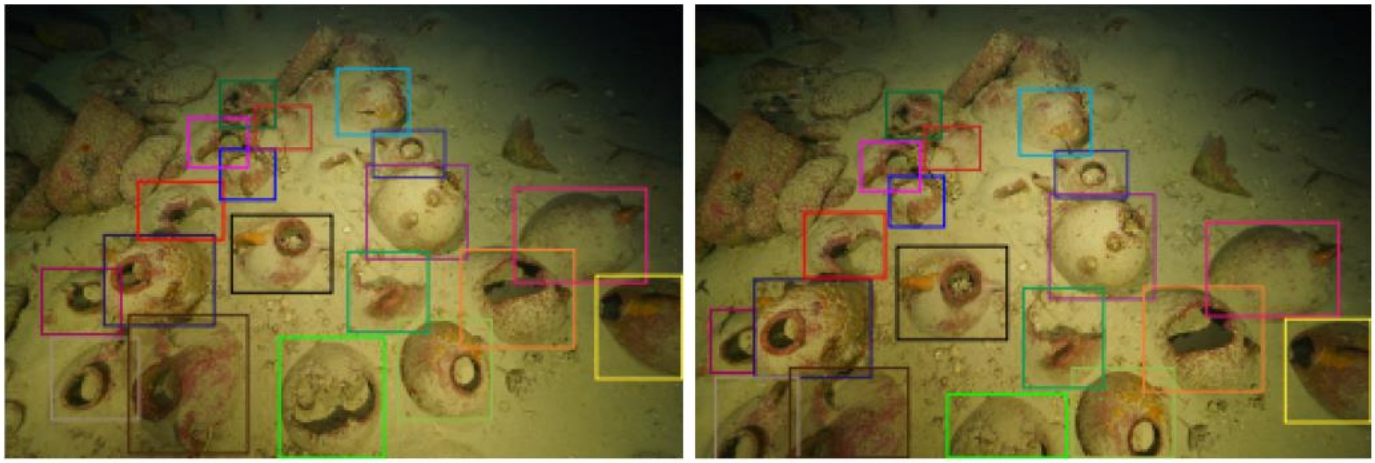
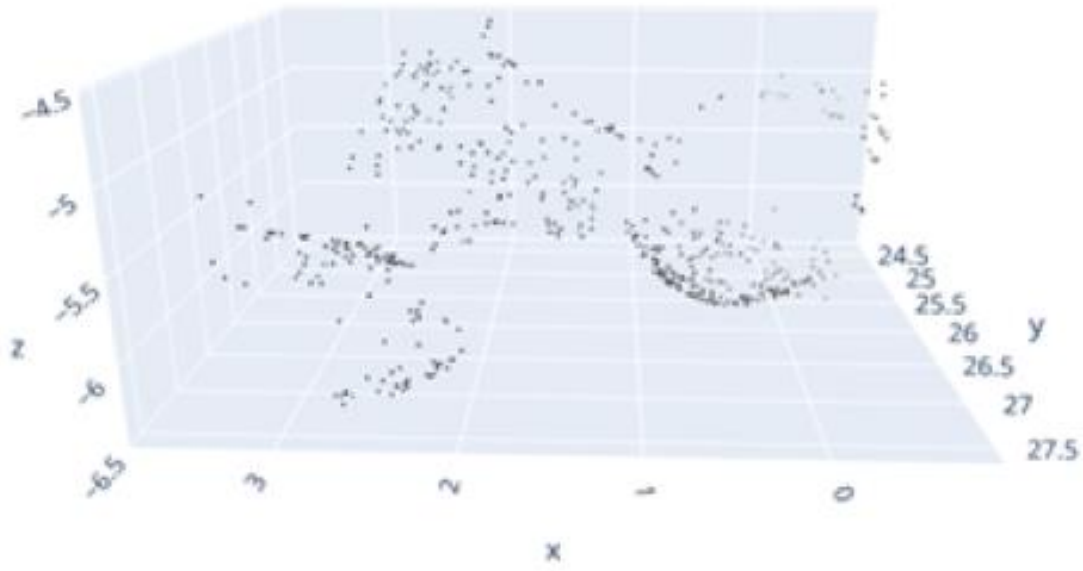


Figure 8. Example images of labelled amphorae from the Xlendi wreck (Kamal et al., 2024: 1334).

(a)



(b)

Figure 9. a) 4 amphorae detected with the YOLO algorithm (b) The same amphorae being 3D tie-point plotted for the instance segmentation method (Kamal et al., 2024: 1338).

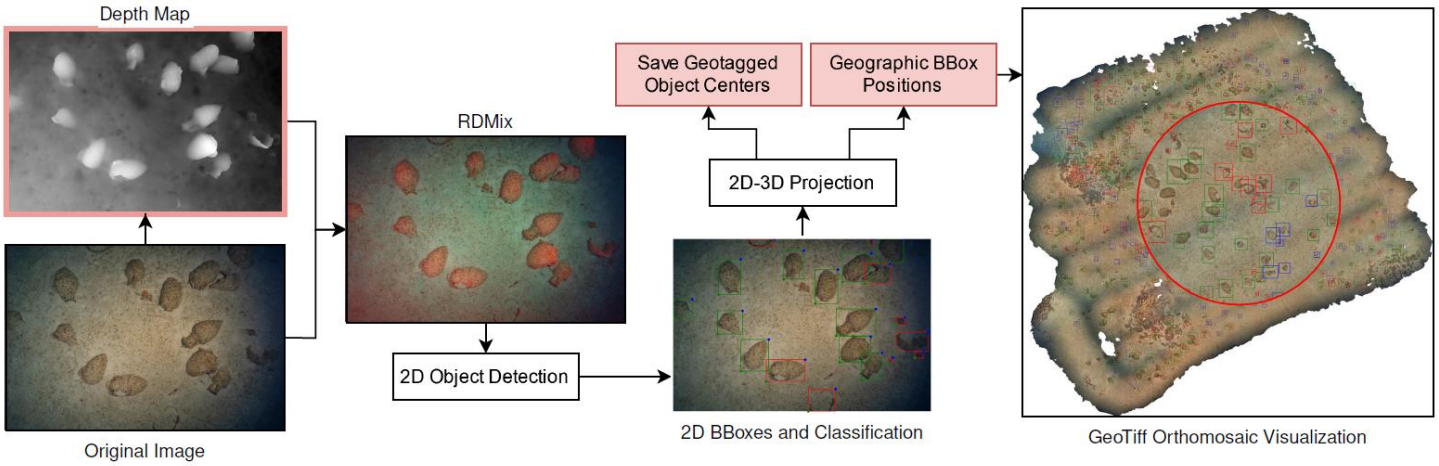


Figure 10. Proposed five-step Bidirectional Fusion Architecture including RDMix augmentation and geographic contextualization (Zammit et al., 2024: 4123).

computer vision techniques to enhance performance and expand applications. For instance, Zammit et al.'s method of projecting information from 2D images onto a 3D photogrammetric orthomosaic enables the quantification of object classes across an entire site—an approach that would otherwise be impractical if the test data consists solely of photogrammetric images. It is also fair to say that while their work highlights promising advancements for detection in maritime archaeology, it underscores the challenges of implementing such methodologies in smaller-scale projects due to the extensive technical expertise and interdisciplinary support required.

Nevertheless, these studies have also been valuable to this dissertation, offering insights into how computer vision researchers tackle the challenges of detection in archaeological assemblages. Their focus on the mathematical foundations of detection techniques contrasts with archaeology-centered approaches, broadening our understanding of the method and its output.

2.3 Conclusions to the Literature Review

The application of object detection in underwater archaeology has steadily increased in parallel with advancements in ways to circumnavigate the issues introduced by the maritime environment, computing power and detection algorithms. As these technologies become more refined and efficient, object detection is becoming increasingly accessible. YOLO exemplifies this trend as an

open-source model that does not require site-specific modifications or specialized AI knowledge for implementation.

Despite these advancements, object detection remains underutilized in maritime archaeology. When employed, it is typically in the context of site detection projects using geophysical imagery. Even fewer studies have applied detection to optical imagery to analyze material assemblages (Pasquet et al., 2017; Yang et al., 2023; Paraskevas et al., 2023; Kamal et al., 2024, Zammit et al., 2024). These studies, which examined sites such as 12th- and 16th-century Chinese wrecks, a 4th-century BC shipwreck off Modi Island, Greece, and multiple sites off Gozo, Malta, have laid the foundation for the practical use of detection models in underwater archaeology.

This dissertation builds upon their work. It does so not by developing a new method to study Xlendi Archaeological Park, but by testing detection on a controlled environment and developing an idea for a theoretical framework with which to adapt detection models for practical use on every underwater site. Additionally, it provides a perspective on fundamental questions that are crucial to consider as the discipline of maritime archaeology moves toward greater integration of computer vision and automated analysis.

3. AUTOMATED OBJECT DETECTION

Contents

3.1 Object Detection Models in Archaeology. Introduction

Archaeological materials are consistently sampled at sites and sent to laboratories that apply techniques such as radiocarbon dating,[36] thermoluminescence,[37] palynological analysis,[38] X-ray analysis, and others (Liritzis et al., 2020: 5). These methods fall under the umbrella of archaeometry, which involves the application of scientific techniques to analyze archaeological remains (Doran, 1970: 289; Ehrenreich, 1995: 3; Rousaki et al., 2018). Through this branch of science, maritime archaeology relies on methods from chemistry, biology, physics, and geology to answer questions about the archaeological materials—questions that cannot be addressed through relative dating techniques like stratigraphy or typological sequencing (Twede, 2002: 102-105; Jones, 2004: 331-332; Bahn and Renfrew, 2011: 121-124). It is among these methods that computer vision, a form of AI, finds its place.

Although AI itself may seem like a modern development, its origins stretch back further than one might expect. Its roots can be traced to the American science fiction writer Isaac Asimov and his 1940 novel *Runaround* (Haenlein and Kaplan, 2019: 6). As a concept, AI is highly complex, with a long history and a constant presence in both philosophical thought and popular imagination due to the far-reaching implications of its advancements. For archaeologists, or other researchers who

[36] **Radiocarbon dating** is a method that counts measures the presence of a specific isotope of carbon (^{14}C) on organic objects to assess their age. It is widely used in archaeology, geology and other environmental sciences (Renfrew and Bahn, 2011: 143-146).
[37] **Thermoluminescence** dating is a method that involves reheating samples of archaeological materials created by heat (ceramic, bricks, sediments) to assess their age through the measuring of the light released from the quartz and feldspar lattices that (generally) form these elements (Renfrew and Bahn, 2011: 155).
[38] **Palynology**, in archaeology, is the study of preserved pollen and spores to reconstruct environments, climates and human activities through the presence of seeds or the content of recipients (Renfrew and Bahn, 2011: 246).

may not be AI experts but wish to apply computer vision in their work, the technical concepts involved can seem overwhelming, potentially serving as a barrier to entry (Cowgill, 1967: 332). However, much like the archaeometric methods mentioned earlier, the application of computer vision such as photogrammetry,[39] instant segmentation[40] and object detection often involve collaboration with domain experts in AI. This is not inherently problematic, as interdisciplinary teams comprised of archaeologists and technology experts have demonstrated success in working together (Kamal et al., 2024; Zammit et al., 2024). Even so, this dissertation explores the idea that these results could be further enhanced if archaeologists were more knowledgeable about detection algorithms.

This idea is not new. As previous literature has suggested, archaeological studies that integrate both archaeological and archaeometric evidence inevitably require the collaboration of both archaeological and technical experts during the interpretive phase (Vitali, 1989: 389).

Based on these premises, this chapter will focus on offering a practical explanation of the implementation of YOLO detection models in archaeology projects. A more in-depth background in the development of detection systems in relation to maritime archaeology as well as an overview of the way YOLO models function is provided in Appendix I.[41] This is necessary only if one wishes to fully understand the inner mechanics—the operation—of detection algorithms.

3.2 Use of Detection Models in Archaeology. Practical Explanation

While each stage of a detection algorithm's processes is fundamentally based on mathematical principles, pre-trained models available in online repositories can be further trained and customized using programming libraries like TensorFlow or PyTorch that specialize in such tasks (Vasilev, 2019).[42]

In archaeology, these models can be trained to recognize patterns and classify cultural artifacts across large datasets. The training process is automated: the model's programming code defines a layered architecture that links mathematical operations to human-provided data. This data, which

[39] Photogrammetry (p.1).
[40] Instant Segmentation (p.11).
[41] Appendix I (p.153).
[42] Specifications and instructions for the use of YOLO models can be accessed through the developer's online repository (https://github.com/ultralytics/ultralytics).

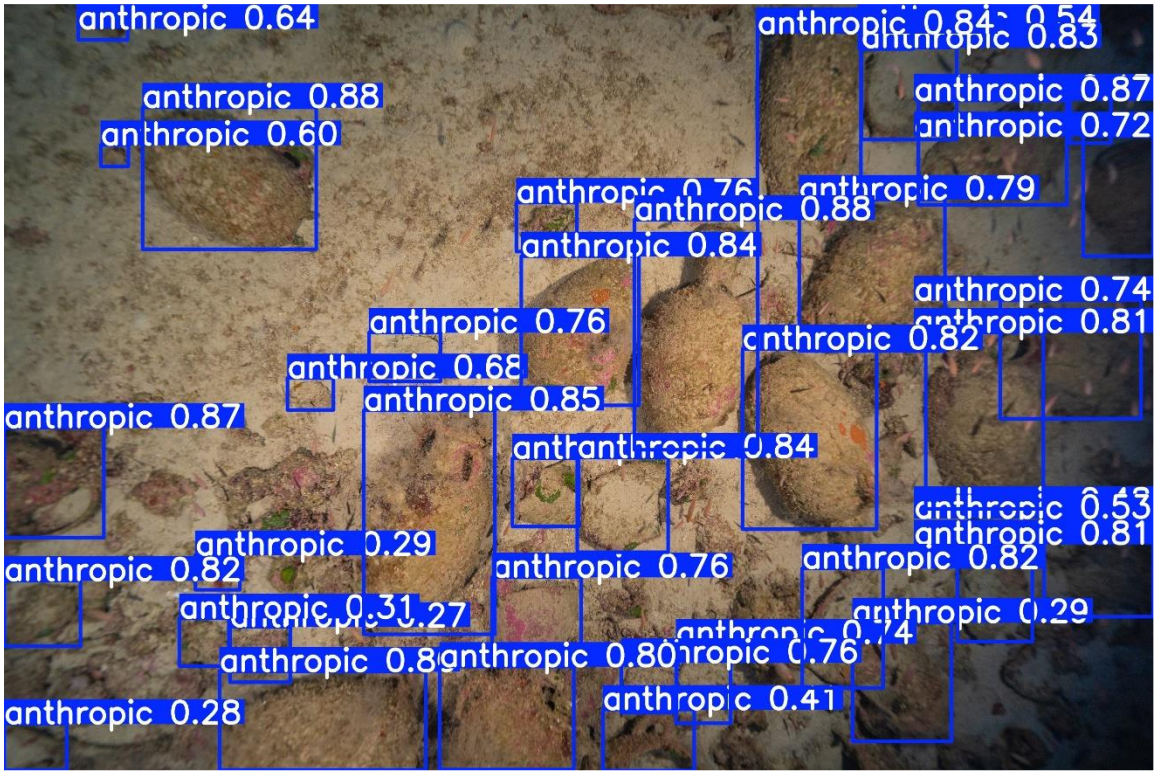
is introduced by the user (archaeologist) is known as ground truth, and its purpose is to provide correctly processed data (located, identified and classified) that the model can use to learn patterns and refine its predictive accuracy. Once this pattern recognition process is complete, the model is considered trained. A trained model can be used to make accurate predictions on new, similar data without needing human intervention for each new case (Figure 11).[43]

To illustrate this process in the context of maritime archaeology, consider a scenario where a team of researchers is tasked with developing a model to locate shipwrecks across a vast expanse of seabed using geophysical data collected through SSS.[44] The workflow is outlined in the diagram shown in Figure 12. The first step for researchers will be to gather a large dataset, such as drone imagery. Then, instead of locating and characterizing all the data manually and if they have access to automated detection, they will manually process just a small portion of this data (for instance, 5%) using free-code labeling software (Kamal et al., 2024: 1335).[45] Next, they will feed the labeled into the detection model, which, using high-performance computing resources, adjusts its predictive parameters (called weights) in a repetitive process to minimize errors and learn the patterns from the ground truth (the correctly labelled/characterized data). Once the training is complete, the model will be capable of detecting similar archaeological features in new, unseen geophysical data. In this case, by manually characterizing the shipwrecks in just 5% of the seabed area to train a detection model, the researchers will then make use of it to automatically and accurately find and characterize shipwrecks in the remaining 95%, thus saving hours of a repetitive task, limiting human error with the inherent consistency of automation, and being resource efficient.

This process is widely used in archaeology across various contexts, data types, and purposes. The ArchAIDE project is an application designed to use smartphone photography to recognize decoration patterns on pottery sherds and identify typologies based on shape (Gualandi et al., 2016; Anichini et al., 2021). There are also many site detection projects that use detection to automatize the process of identifying sites in very large areas using geophysical data (Atallah, 2005; Zhu et

[43] A **trained model** is a model that has gone through the training process and learned the patterns of the ground truth by adjusting a set of parameters (called weights) that can be then used to make predictions on unseen, unclassified data (Paraskevas et al., 2023:3).
[44] Side-scan Sonar (p.11).
[45] Some examples of this labelling software are online free tools like Roboflow, Labellmage or the one used for this project: Makesense.ai. It can be found at https://www.makesense.ai.

a)



b)

Figure 11. Example test image from the project: **a)** Before predictions. **b)** After predictions.

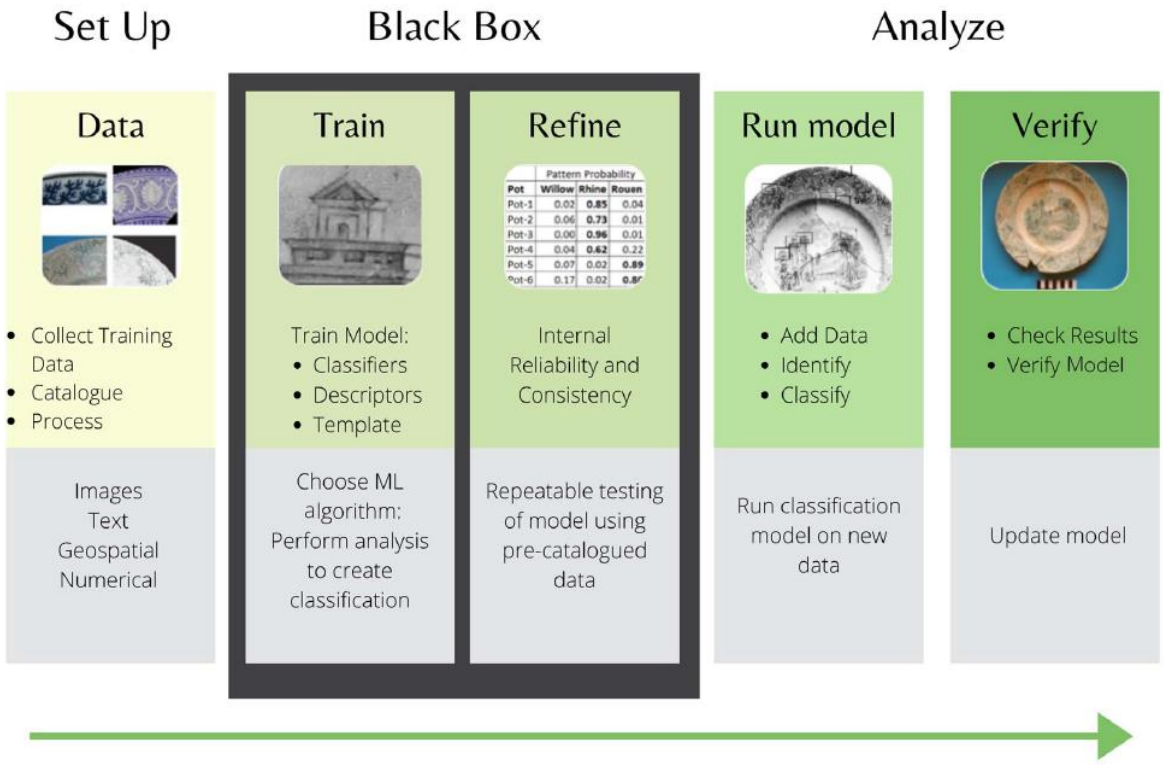


Figure 12. Diagram of object detection algorithms being applied to archaeological data (Bickler, S. H., 2021).

al., 2019; Nayak et al., 2021). Similarly, in taphonomic studies, detection models have been trained to classify post-depositional marks on animal bones—such as cut marks made by stone tools, metal implements, or animal teeth—based on nanometric structural differences captured with a binocular microscope (Byeon et al., 2019: 38; Domínguez-Rodrigo et al., 2020: 7-8).

These examples illustrate the versatility of detection methods across different data sources: optical images, geophysical data, and microscopic imagery. Despite the variation in input data, the underlying process remains the same. A subset of the data is manually analyzed and labeled as ground truth, which is then used to train the model.

Once fully trained, detection models can process both images and videos, identifying objects within bounding boxes, classifying them based on the parameters defined during training, and providing a confidence score for each prediction. Additionally, as shown in Figure 11, each detected object is assigned precise coordinates in the form of a bounding box with numerical values that locate it within the image. These are stored in a file accompanying every output image.

3.3 Ultralytics YOLO Models

Recently, YOLO (You Only Look Once) models have emerged as the most efficient choice for applications requiring real-time detection.[46] They enable instant predictions with minimal run-time and have seen widespread use in various fields, including agriculture, security, healthcare, transportation, and more (Figure 13).[47]

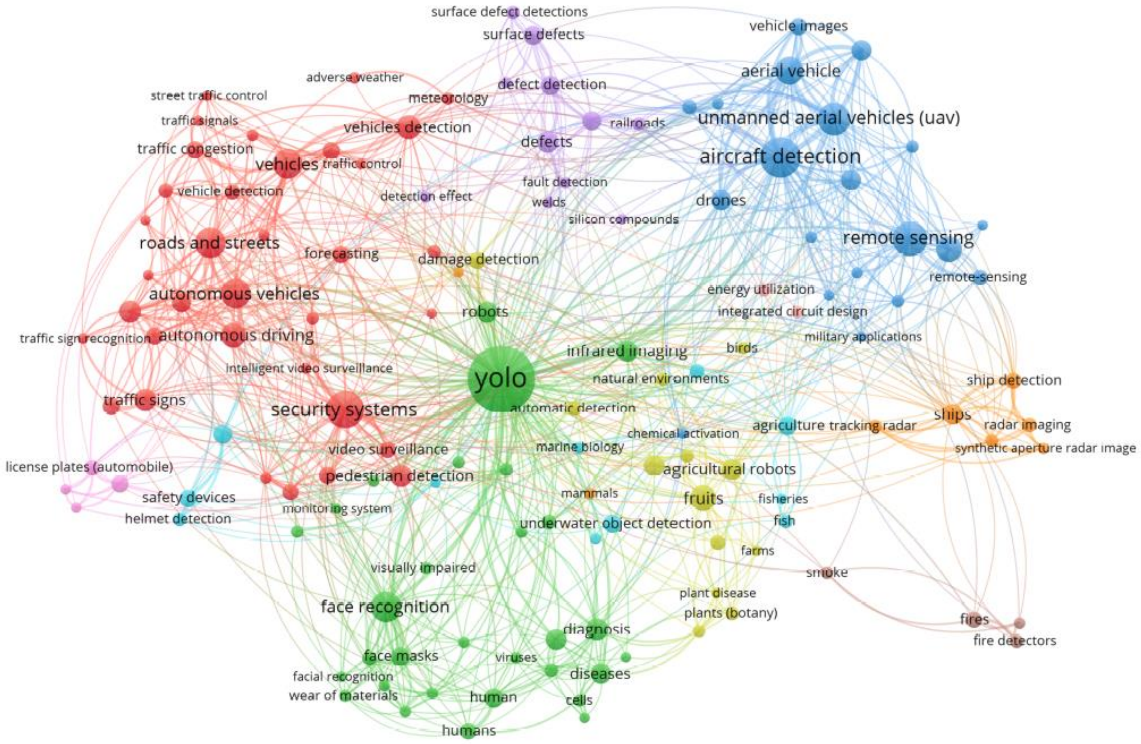


Figure 13. Bibiliometric network visualization of the main YOLO applications (Terven et al., 2023: 1682).

[46] In object detection, **real-time** detectors are those versions able to **instantly** carry out the tasks of their class (identifying, locating and classifying objects). For instance, face recognition systems on security cameras (Redmon, 2016: 785-8).

[47] In agriculture, YOLO is used to detect and classify crops and diseases (Tian et al., 2020; Wu et al., 2020) and assists in automated farming techniques (Lippi et al., 2021). In security, YOLO has applications in biometric padlocks, video-feed analysis for surveillance, and facial recognition systems (Wang and Zheng, 2018; Kumar et al., 2021). In the medical field, YOLO has been used for cancer detection (Al-Masni et al., 2018; Nie et al., 2019), dermatology (Ünver and Ayan, 2019), and pill identification (Tan et al., 2021). Governmental agencies apply YOLO in traffic and transportation systems for license plate detection (Chen, 2019), traffic sign recognition, and traffic management (Dewi et al., 2022). In the private sector, YOLO has been employed in wildlife monitoring (Roy et al., 2023), robotics (Dos Reis et al., 2019; Kulik and Shtanko, 2020), product quality control (Li et al., 2018; Du et al., 2021), sports analysis (Zheng and Zhang, 2022), and remote sensing (Pham et al., 2020; Cheng et al., 2021).

3.3.1 Why is YOLO Ideal for Archaeological Research. Strengths and weaknesses.

In the context of archaeology, YOLO (You Only Look Once) models are notable for their reduced computational complexity and exceptional real-time processing capabilities, making them particularly effective for rapid object detection tasks. This is especially advantageous in maritime archaeology, such as during real-time site monitoring with AUVs/ROVs[48], where immediate feedback is essential. As an end-to-end architecture, YOLO consolidates all stages of object detection—feature extraction, classification, and bounding box prediction—into a single, streamlined workflow. This eliminates the need for multiple software tools and simplifies deployment. Such characteristics enhance YOLO's versatility and robustness, particularly when applied to novel domains or when handling unexpected inputs. Additionally, its open-source availability contributes significantly to accessibility, enabling researchers and practitioners without extensive expertise in artificial intelligence to experiment with and deploy advanced object detection systems. For these reasons, YOLO has been increasingly adopted in recent maritime archaeological projects seeking to automate visual analysis (Character et al., 2021: 1759; Paraskevas et al., 2023: 2; Kamal et al., 2024: 1335).

That said, the application of object detection models to underwater assemblages comes with certain inherent challenges. These include dependency on annotated training data, a frame-by-frame detection approach that lacks contextual scene awareness, and the variability of underwater environments that often necessitate retraining or adaptation of the model. While not unique to YOLO, these issues are generally unavoidable and must be accounted for during implementation.

When considering YOLO models specifically, several limitations emerge that—although often outweighed by their advantages—can influence their performance in underwater archaeological contexts:

-Low visibility and lighting sensibility. YOLO models, particularly those optimized for speed, are sensitive to poor lighting and low-contrast imagery (Character et al., 2021). This can be problematic when data is captured manually by divers using basic underwater lights, which may not illuminate the entire scene uniformly. Such conditions are representative of the dataset used in

[48] AUV, ROV (p. 8).

this dissertation, which aims to assess YOLO's applicability using widely accessible, non-specialized technology.

-Limited overall precision. While YOLO is not inherently inaccurate, its architecture prioritizes inference speed and simplicity over detailed localization. As noted in several studies (Girshick, 2015; Redmon et al., 2016; Character, 2021; Fayaz et al., 2022), this results in well-documented weaknesses when detecting small, overlapping, or partially obscured objects. YOLO's single-pass, grid-based detection structure—lacking the region proposal stage found in R-CNN-based models—can lead to missed or imprecisely located objects in cluttered scenes with many competing features.[49] Such conditions are typical of archaeological sites, especially when working with overlapping/buried amphorae, dense pottery scatters, or small sherds.

These limitations are more pronounced in earlier versions of YOLO (v1–v3). More recent iterations (such as YOLOv5 and v7), including those used in this study, have improved accuracy and robustness through architectural updates such as multi-scale feature detection and attention mechanisms. Nevertheless, many of the underlying trade-offs between speed and precision persist and remain important considerations in evaluating the technique's suitability for archaeological datasets.

[49] see Appendix I (p.153).

4. XLENDI UNDERWATER ARCHAEOLOGICAL PARK. THE TOWER WRECK

Contents

The Xlendi Underwater Archaeological Park, located off the coast of Xlendi Bay in Gozo, Malta, was established in 2022 as the world's first deepwater archaeological park. It encompasses a 67,000 square meter assemblage of (mostly) Punic amphorae at depths ranging from 105 to 115 meters. The site's morphology and the characteristics of the materials found in it make it the perfect scenario for the training and testing of detection models.

The following section provides an overview of the site's archaeological significance and an explanation of the reasons that make it ideal for the training of detection models. An in-depth background of the geographical and historical contexts of the site as well as a summary of all the past interventions are provided in Appendix II.[50] Additionally, the ceramic catalogue of the site, which was created specifically for our experiments, is available in Appendix III.[51]

4.1 Xlendi Archaeological Park

The seabed around Xlendi Bay is rich in submerged archaeological artefacts, with amphorae being the most abundant. Since the 1950s, numerous interventions have uncovered a significant array of materials, primarily from around the cliffs and reefs bordering the inlet. However, despite this extensive evidence, the only officially registered shipwreck in the area remains the 7th-century BC Phoenician shipwreck (Gambin et al., 2018, 2021). As for the rest, it has often been regarded as a decontextualized spread of artefacts with uncertain origins (Atauz, 2004: 376-377; Azzopardi, 2006: 153, 2013: 293).

[50] Appendix II (p.166).
[51] Appendix III (p.175).

It was not until the archaeological surveys conducted between 2000 and 2007 with the assistance of an ROV[52] that patterns within this spread of material began to emerge (Atauz, 2004: 23-24). Reaching depths of up to 115 meters, the terrain around Xlendi Archaeological Park consists primarily of silt, sand, and rocky outcrops, with countless archaeological objects—primarily ceramics—embedded in the seabed, and many more still buried.

While at least one shipwreck dated to the 3rd century BC has been identified, the sheer quantity of material suggests a more complex scenario. Several hypotheses have been proposed to explain the extensive distribution of artefacts, including the gradual displacement of objects due to maritime factors, the accumulation of multiple shipwrecks across many centuries of commercial activity, or ritual loss overboard (Azzopardi, 2013: 293). Moreover, from a research standpoint, Xlendi Archaeological Park emerges as a crucial site in the past debate about the role of the Maltese Islands in ancient trade and seafaring (Figure 14).[53] Addressing these varying opinions regarding the uniqueness of the site would undoubtedly be valuable information, and detection methodology is one of the ways in which this could be achieved in the future.

Some valuable insights can already be gained from the site through a dedicated study of the assemblage of ceramics (Azzopardi, 2013: 290-4). Each ceramic typology has its own geographical origin and contextual significance, making the study of these artifacts akin to assembling a puzzle that reveals trade relationships between different regions and cultures (Twede, 2002: 100-107). Beyond commerce, such interactions played a fundamental role in the exchange of ideologies, technologies, and cultural practices, shaping the broader historical landscape (Gibbins and Adams, 2001: 281, Twede, 2002: 100-108).[54] Reconstructing the timeline of these connections is therefore essential, and in archaeology, ceramic analysis remains one of the most

[52] ROV (p.8).

[53] This debate revolved around the fact that there are not many ancient shipwrecks around the Maltese Islands, which some interpreted as a sign of limited maritime activity (Atauz, 2004:269) while others have presented other evidence to suggest the contrary (Gambin 2005; Bruno, 2004, Azzopardi, 2013). This debate has been settled as of late with publications showing strong evidence that the Maltese archipelago was indeed of great importance in the wider Mediterranean (Gal et al., 2022:3).

[54] It is true that ceramic studies improve our understanding of the morphological variations of the materials depending on their origin, the contemporaneity and significance of different forms, the scale of their production, their contents, and their epigraphy. This notion is applicable not only to ceramic, but also to other materials and to the entire archaeological sites themselves. Ships, for example, would also have been conceived and designed according to certain mental templates and ideologies.

In general, underwater archaeological sites act as time capsules that answer social questions and illustrate processes of cultural change across time (Gibbins and Adams, 2001:281).

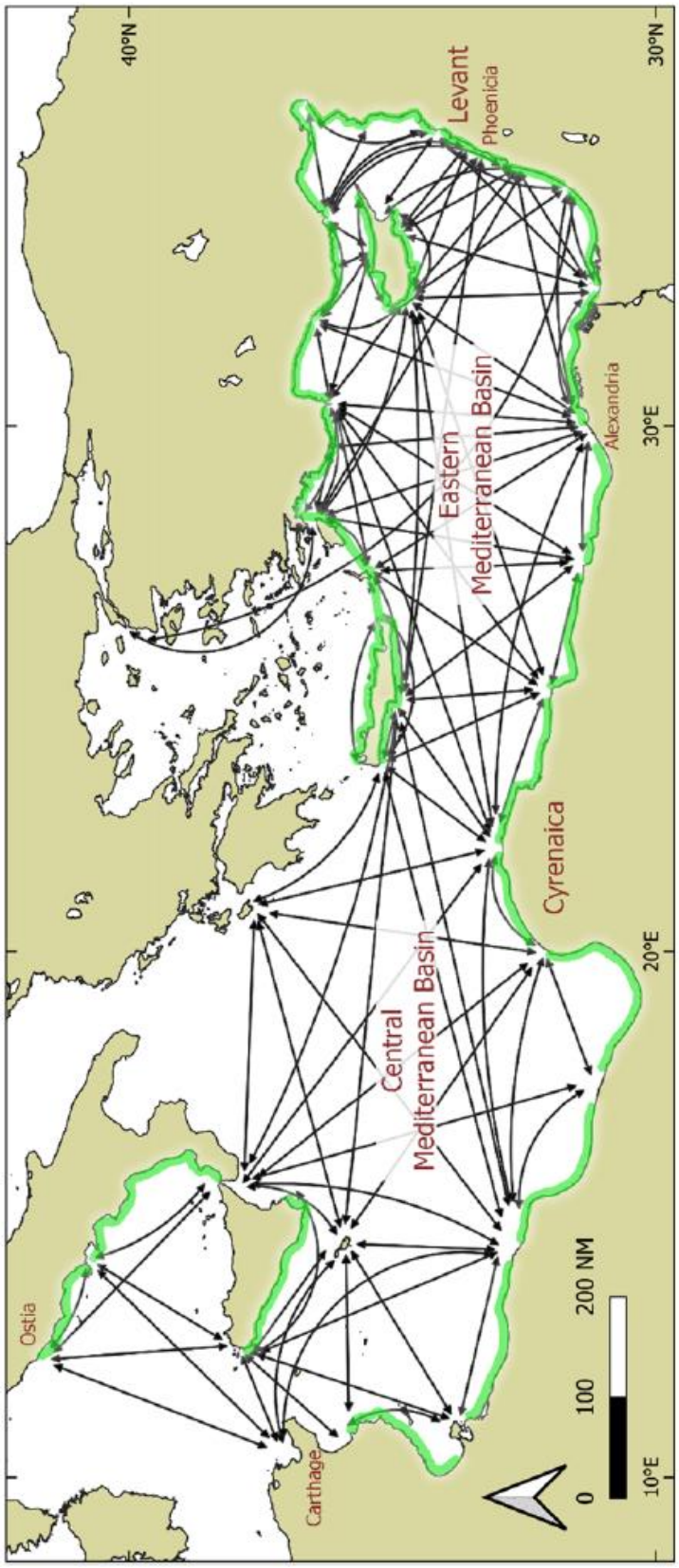


Figure 14. Scope of Gal's mobility measurements for ancient seafarers. Black lines indicate bidirectional links, and green shading represent the coasts where coastal sailing is a possibility (Gal et al., 2022:3).

effective ways in which this might be achieved (Peacock and Williams, 1986: 54-63; Rice, 1987: 329; Orton et al., 1993: 23-34; Azzopardi, 2013: 289-90).

Xlendi Archaeological Park contains a diverse assemblage of archaeological materials from across the Western Mediterranean (Appendix III) which could provide valuable insights into the significance of Xlendi Bay, both locally and within the wider Mediterranean context. While a full analysis falls beyond the scope of this dissertation, the research on typologies conducted here lays the groundwork for future studies in which our detection models designed could further support these efforts by offering a clearer understanding of the site's nature and the role of its pottery.

Following the designation of Xlendi Archaeological Park as an Archaeological Zone at Sea under Maltese heritage legislation (Gambin & Bonanno, 2006), the Tower Wreck Project (2021–present) was launched to study and document the site's morphology. In 2022, Xlendi was declared the world's first deep-water archaeological park, and by 2023, it became accessible to divers through SCUBA clubs and Heritage Malta (Figure 15).

4.2 Xlendi's Value for Automated Detection Methodology

When studying underwater archaeological assemblages, it is essential to consider the specific configurations of shipwreck sites. Extensive research on shipwreck formation processes has demonstrated that, in cases of undisturbed deposits, cargo typically clusters around a wooden hull's original shape (Muckelroy, 1978). As shown in Figure 16, this clustering effect makes ancient shipwrecks relatively easy to identify using remote sensing techniques, exhibiting defined patterns of site morphology that detection models can easily learn to recognize.

Detection models rely on large datasets to improve their accuracy and adaptability. A dataset with greater variability results in more robust final models, enhancing their ability to be exported across different archaeological sites. However, if a detection model is trained exclusively on a single, well-preserved wreck—where amphorae remain stacked in their original position—it may struggle to perform well in more complex or dispersed assemblages. A model trained under such conditions may fail to recognize amphorae in varying states of preservation, across different typologies, or scattered across larger areas due to post-depositional processes.

Based on these and as shown in Figures 15 and 17, Xlendi Archaeological Park provides an ideal

Figure 15. Photogrammetry and image (down) of Xlendi Archaeological Park (from Heritage Malta Website).

environment for training and implementing object detection models due to several factors:

-Large-scale artefact distribution: The extensive spread of materials across the site ensures diverse training data.

-Consistent archaeological presence: The park contains artefacts in various orientations and states of preservation, enhancing model robustness.

-Consistent site morphology: Xlendi Archaeological Park spans a mostly flat surface of sandy terrain. Unlike shallower and rockier sites, which often feature more life and other disturbances that complicate detection, Xlendi's more stable environment facilitates clearer detection.

-High typological diversity: The presence of at least 16 distinct ceramic forms, ranging from small rim sherds to fully intact vessels, allows models to learn a broad spectrum of variations.

Given these advantages, Xlendi Archaeological Park plays a crucial role in this study, serving as a testing ground for different approaches to the detection model's training process. The insights gained from this work have significantly contributed to the development of more efficient and adaptable automated detection models, and such a thing would have been impossible to achieve using a more uniform or limited dataset or working on a shallower and confusing site.

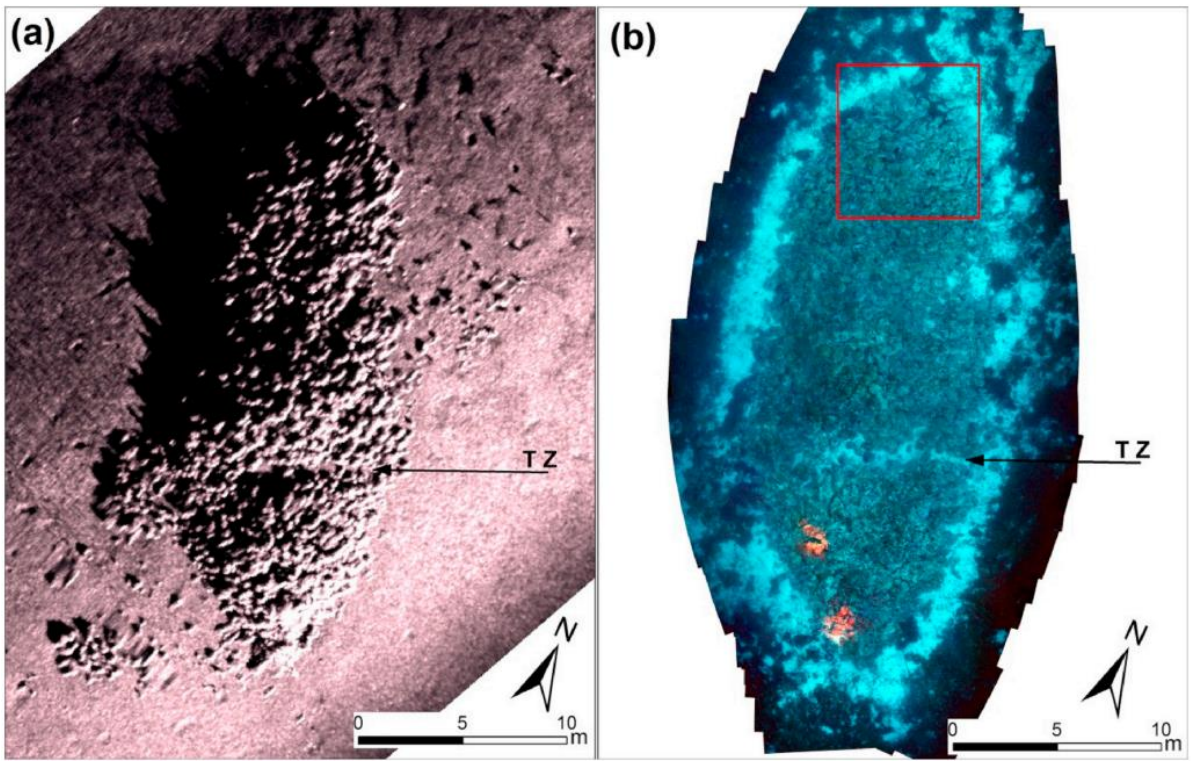


Figure 16. View of traditional site morphology of an ancient shipwreck **a)** High resolution imgery showing the cargo outline of the "Fiskardo" shipwreck. **b)** ortophoto of the oval shape of the amphorae cargo (from Ferentinos et al., 2020: 6).

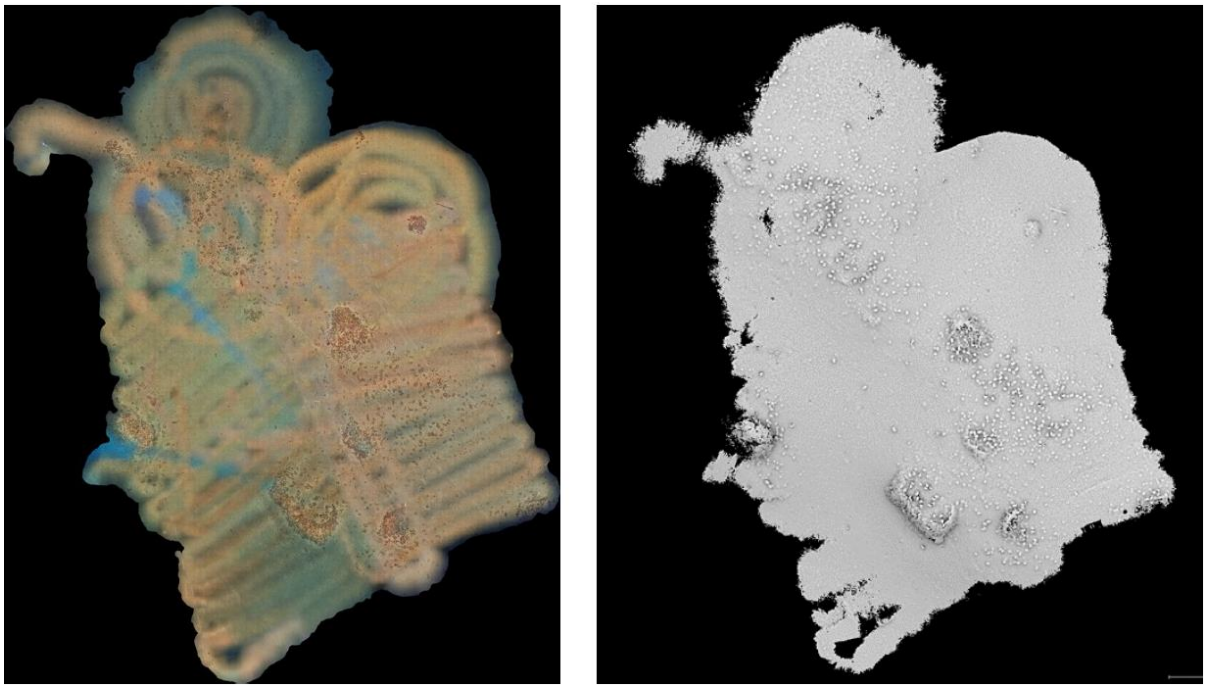


Figure 17. Orthomosaic showing the distribution of material present that is the norm at Xlendi Archaeological Park. Though this data is only from the 2021 season, it is possible to notice the combination of clusters and widespread density of material (Department of Classics and Archaeology, University of Malta).

5. METHODOLOGY

Contents

5.1 Method Overview

The goals of this dissertation are threefold: to describe and contextualize object detection methods for archaeological use (addressed in Chapter 3 and Appendix I);[55] to evaluate these methods as tools for extracting information from underwater archaeological assemblages by integrating archaeological data into the training process; and to test the hypothesis that an archaeologist's personal involvement in the development and implementation of the methodology improves its performance and utility from an archaeological standpoint.

We pursued the last two goals by testing and comparing two of the most widely used and accessible detection models (YOLOv8 and v11) on the ceramic assemblage of Xlendi Archaeological Park. Of the site's total area (67,000 m²), we selected 2,268 photos from

[55] Appendix I (p.153).

photogrammetric runs covering 625 m² as the primary dataset for this dissertation (Figure 11). The number of images matches or surpasses those used in similar studies (Paraskevas et al., 2023: 3; Zammit et al., 2024: 4124). As for the experiments themselves, we conducted on a single-node setup with an RTX 3070 8GB GPU, making the process replicable on any mid-range desktop.

Throughout the project we trained a total of 72 detection models. While the models vary in terms of the algorithm versions and sizes used, we considered it necessary to further classify and differentiate them from an archaeological perspective as well. To accomplish this by providing extra options, we constructed a complete catalogue for the site's ceramic typologies (Appendix III).[56] Following this, the models were evaluated and had their performances compared. The knowledge gained throughout this process allowed us to more confidently contextualize the methodology within maritime archaeology while assessing its prospects.

The purpose of this chapter is to provide the essential tools to understand the process of designing and training a new detection model. It also addresses the need to explain the approach used to construct a theoretical framework with which to classify and evaluate the different trained models. Finally, to serve as a detailed reference for the interpretation of the results provided in the next chapters, we provided an overview of the metrics used during the comparative process in Appendix IV.[57]

5.2 Implementation of Automated Detection

The subject of this section is to outline the steps we followed during this project to create a single detection model from scratch. The entire process can be followed in the flowchart provided in Figure 18:

-Preliminary stage: The first step in the process is defining and acquiring the dataset necessary to create robust and versatile models. For an experiment aiming to train multiple models and compare them in the way we aim to, this is done only once.

[56] Appendix III (p. 175).
[57] Appendix IV (p. 209).

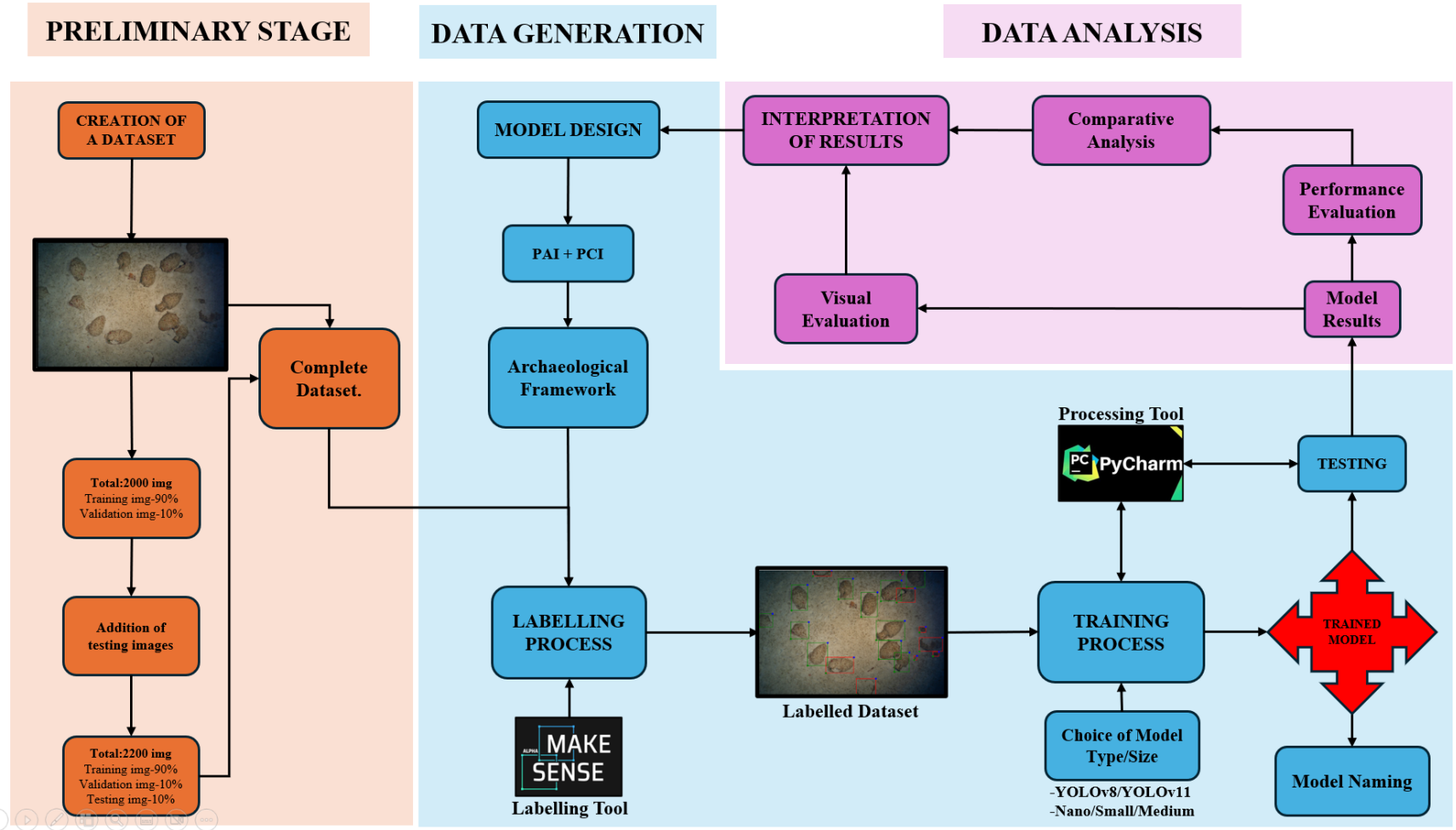


Figure 18. Flowchart of the methodology followed during the experiment.

In this case, images were selected from multiple photogrammetric runs used to document Xlendi Archaeological Park. To ensure a diverse dataset with varying heights, angles, and perspectives, images from three different divers were chosen for the training set. This resulted in a total of 2,200 images that were divided into different folders as follows:

• A **training** set of 1,760 images (80% of the total).

•A **validation** set of 220 images (10% of the total).

•A **test** set of 220 images (10% of the total).

The division and naming of the three data subsets adhere to the formatting requirements for implementing YOLO models. The specific split ratios between the training, validation, and test sets are typically adjusted according to dataset characteristics (e.g., size), project needs (e.g., the need for robust evaluation or to address overfitting), and model complexity (e.g., the size and depth of the model variant used). In our case—aiming to process highly specific and complex information from a relatively small dataset—we initially adopted an 80%/10%/10% split, informed by the only prior application of detection models to underwater assemblages (Paraskevas et al., 2023). This choice was considered preliminary and subject to revision depending on training outcomes. However, early model results indicated no need for adjustments.

-Data generation stage: This stage involves designing, training, and testing individual models. The model design process begins with the definition of the archaeological purpose sought for each model, ensuring alignment with specific research questions. Once the objectives are established, objects within the dataset must be identified, located, and classified using free online labeling software (Figure 19). For this study, Makesense.ai was chosen as the labeling tool due to two reasons: First, because we wanted to use software that aligned with the project's positive outlook on making the methodology exportable and readily available to everyone; and second, because of familiarity.[58] This is the most important step, as it is only during labelling that we can integrate archaeological information into the training process.

After labeling, the dataset consists of images with each relevant object enclosed in bounding boxes as well as text files that record the coordinates and class of each annotation. The next step is to feed the data into the YOLO algorithm using a Python script, which is executed in software

[58] During the comparative AI study carried out by Zammit et al., in 2024 (also in Xlendi Archaeological Park), we contributed as domain experts regarding the processing of their archaeological data.

like PyCharm (Figure 20).[59] During training, the model learns to recognize objects by analyzing the annotations in both the training and validation sets. In this dissertation, each model was trained

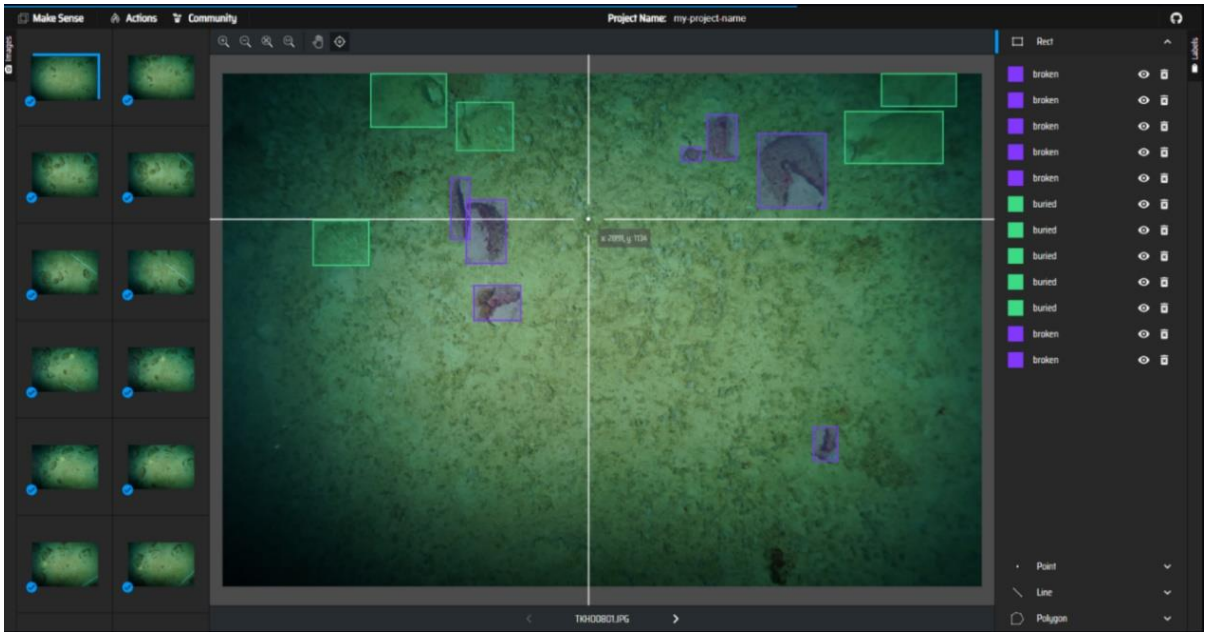


Figure 19. The labelling process involved identifying and locating every object of the dataset using Makesense.ia.

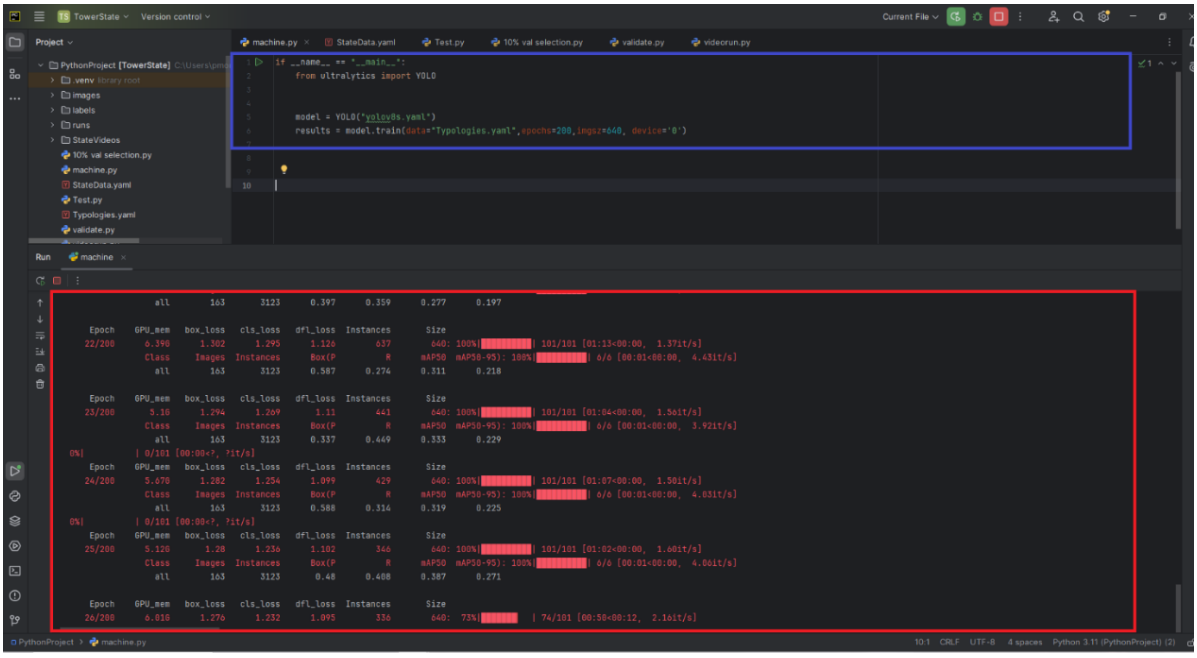


Figure 20. Capture of ongoing training process. Notice the few lines of code necessary to start the process (blue) and the output showing the cyclic nature of the model's adjustments to its prediction parameters (red).

[59] **Pycharm** is an integrated development environment designed by JetBrains for Python development. A few examples of the interface are shown in Figures 20-23.

for 200 epochs[60] or until improvements in prediction accuracy plateaued.

The final step is testing the trained model (Figure 21). At this stage, the model is capable of making predictions based on prior training. To evaluate its performance, the test set—a previously unseen collection of images—is used to generate performance metrics for comparison and assessment.[61] An example of an image passed through a training model can be seen in Figure 22.

-Data analysis stage: This stage involves testing the trained model on the test set and studying the results. The information gathered can be used to help design future models[62].

As shown in Figure 18, the implementation of automated detection follows a three-stage process: preliminary, data generation, and data analysis. It is important to note that the preliminary stage—outlined on page 41—is conducted only once at the beginning of the project. Its primary goal is to create the dataset that will be used throughout the experiment. This step remains consistent across all detection projects.[63]

In contrast, the data generation and data analysis stages are more complex. Operating in a cyclical manner, they continuously refine each other and are unique to our approach. For this reason, while the preliminary stage requires no further discussion, the following sections will explore the data generation and data analysis stages in greater depth.

[60] Epochs (p.16).

[61] The test set of images is unseen because, although it has been labelled, unlike the training and validation sets, it has not been seen by the algorithm at any point during its training—it is new data to the predictive model.

[62] See model design (p.47).

[63] Every detection pr Their mAP50 improves from 87.1% oject in every field, whether professional or informal, needs to go through a preliminary stage in which a dataset to train models is assembled. This process is always the same, as the separation of the data into three sets as stated in pages 40 and 41 responds to the format needs of the YOLO model and cannot be changed. Beyond the considerations regarding the construction of a varied dataset, which are also stated above in our case, there is no need to analyze it further. Changing the dataset during the experiment would defeat the purpose of testing models on the same data, which is another reason for it to remain static in this case.

During this dissertation, however, and as a result of the experiment's own results, we will consider adjustments to the dataset that would benefit future experiments.

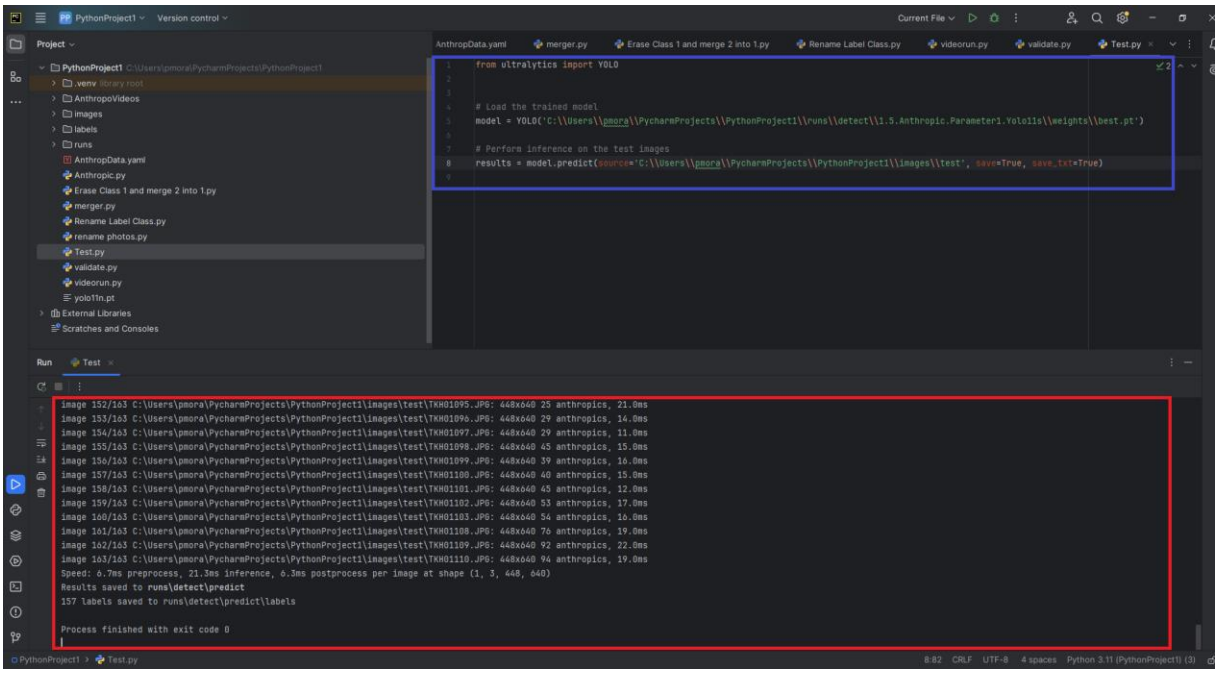


Figure 21. Capture of the implementation of a trained model on unseen data. Notice the few lines of code necessary to start the process (blue) and the output showing the saved location of the output images (red).

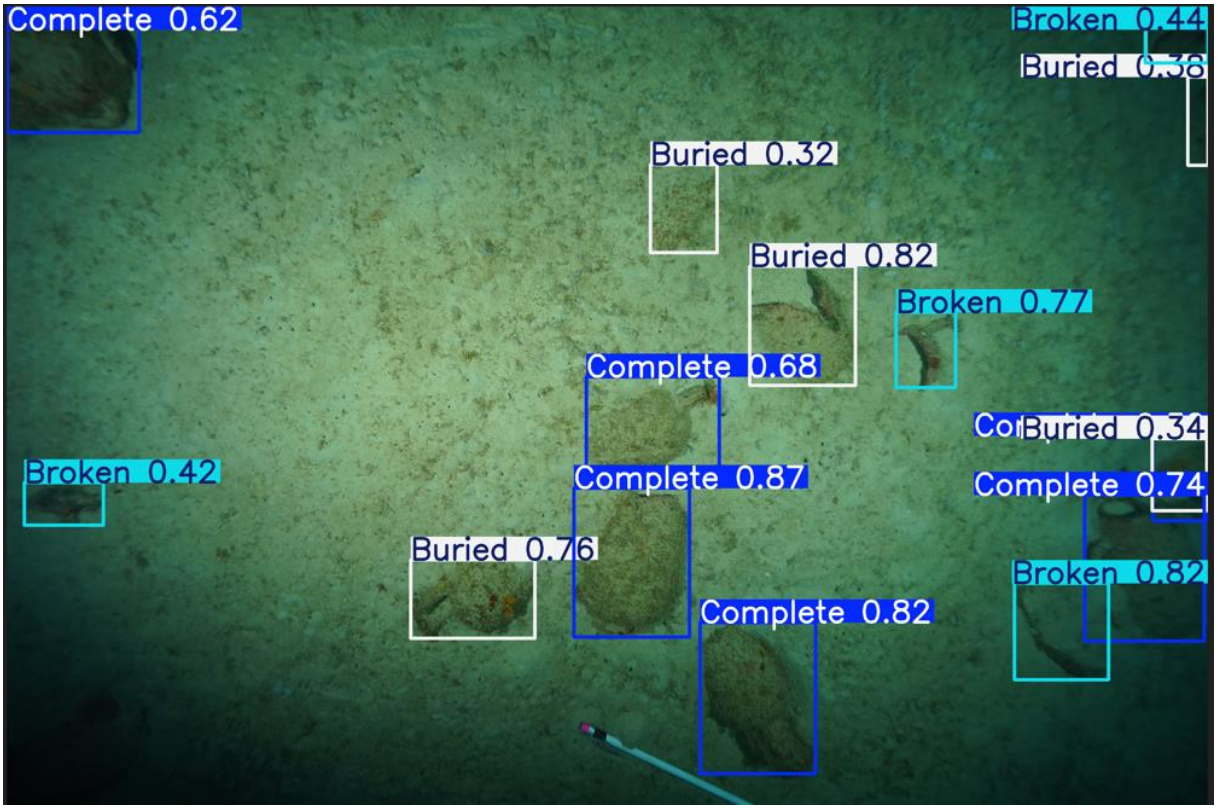


Figure 22. Example of output data after being fed to a model trained to distinguish items by state of preservation. Notice the location boxes, class identification and ratios of precision for each item.

5.3 Data Generation

Having reviewed the general steps followed during this project to produce the detection models that will be the focus of this study, we must now specify the method used to differentiate and design each model in a way that allows a comparative study to yield answers to the aims we have set for it. Classifying the trained models was necessary not only from a results standpoint but also to construct a flexible framework that allows for future additions. At the same time, this framework will provide the tools to develop an equally flexible 'model nomenclature' that will help reduce confusion for all involved.

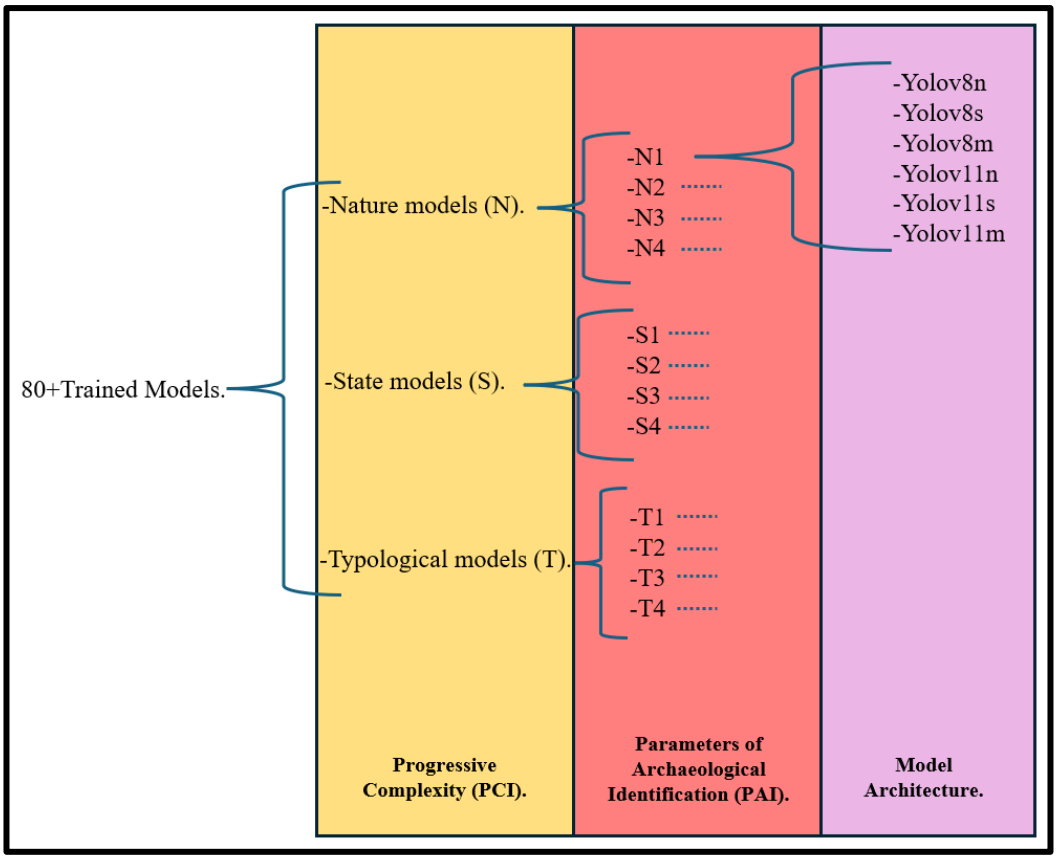


Figure 23. Layers of classification for the models trained during the project.

This framework was formed as the structure of layered differentiation factors shown in Figure 23. Two of these layers distinguish between models based on archaeological parameters, which, in turn, are to be reflected in the results as differences that can be used to compare models. In order of application, these layers are the Progressive Complexity Index (PCI) and the Parameter of

Archaeological Identification (PAI). The third and final layer of differentiation is generated by the specifics of the DL[64] model used to create the trained models. In our case, since we are limiting the study to the latest versions of the YOLO model, these specifics are the different combinations of two versions and three model sizes.

Next, we will discuss the particulars of each factor. This is necessary to understand why, how much, and what type of inherently subjective archaeological information we can hope to embed on a tool from a discipline so dependent on mathematical absolutes rather than treating object detection as just another archaeometric study.

5.3.1 Progressive Complexity Index (PCI)

This factor was devised to classify the trained models into different groups based on the complexity of the task assigned to the detection algorithm, specifically in terms of the number of classes they must process as well as the amount of archaeological information embedded in them during the training stage. Based on these parameters, three different groups were created for the purpose of this dissertation: nature models (N), state models (S), and typological models (T). Models from each group are assigned the corresponding letter of their group as part of their name. Going forward, models from the nature group are called N models, state models are S models, and typological models are T models.

5.3.1.1 Nature Models (N)

These models form the baseline for how object detection can be used on underwater assemblages. The features identified by these models respond uniquely to the dichotomy between the natural background and anthropic objects. Among them, we find models trained solely to identify all traces of pottery. Others incorporate additional elements into the classification, such as discarded plastic. The models in this group have from one to two different classes, and although they do incorporate archaeological information into their predictions, their low number of well-defined classes works in their favor in terms of complexity.

[64] Deep learning (p.11).

5.3.1.2 State Models (S)

The models in this group have two to four classes and focus solely on locating archaeological materials. Once located, these materials are classified according to their state of preservation on the seabed. As not everyone has the same ideas when it comes to the preservation of ceramic material (Fabri et al. 2016; de Lapérouse , 2020: 169), these group's models introduce a clear layer of subjective archaeological information into the training process.

5.3.1.3 Typological Models (T)

Implementing the site's complete typological chart available in Appendix III,[65] the models in this group have from 5 to 16 different classes. Many of the types of pottery are very similar in shape and are found in situ in varying states of preservation. In addition, the maritime environment and the passage of time blur the differences these materials might have displayed in terms of texture, shape, and color. These two factors mean that, despite the theoretical ability of object detection models to handle more than 16 classes, the models in this group still test the model's ability to consistently identify materials correctly. On the other hand, as they also depend on the materials state of preservation, using typologies as classes introduces a vast number of variations in the layers of archaeological information that are being embedded into the models' training stage, which bolsters their potential applications.

5.3.2 Parameter of Archaeological Interpretation (PAI). Model design.

The idea behind the second factor used to further subdivide the models with the same PCI is that the archaeological information embedded into the models through the training stage is often entirely subjective, as assessing matters like a vessel's state of preservation or typology from low-quality pictures[66] of an underwater site is subject to opinion or differences in the user's school of thought (both the trainer, and the interpreter of the model's output). For example, when reviewing

[65] Appendix III (p.175).
[66] While the quality of our original images is high, the specific YOLO models that we are using resize them for computational speed purposes. More information on the influence of architecture size on model performance is available in p.62.

one of the images from the training dataset used in this dissertation (Figures 24-26), two independent experts may differ in how they consider some of the materials during the labelling stage. Since detection training is a process based on repetition and large datasets, occasional minor discrepancies are unlikely to significantly impact the model's development. However, what happens when the principles behind the experts' pair of classifications diverge entirely? What if experts A and B have fundamentally different definitions of what constitutes a Dressel 1 amphora, or what should be considered a 'complete' amphora?

Returning to the previous example, expert A may define a 'complete' vessel as one that is fully above the seabed's plane, while expert B might classify any piece of material with at least 50% of its surface above the plane as 'complete'. This seemingly minor difference can, in fact, have a significant impact not only on the model's metrics and the usefulness of its output, but also on its overall validity. In such a scenario,[67] the model's results would only be reliably interpretable by the person who trained it—and only if they remember.[68]

This variability of possible interpretations is common in archaeological fields such as the study of ceramic typologies, where each researcher brings their own perspectives and interpretations. While such diversity normally fosters fresh ideas and scientific progress, it can also lead to confusion when different scholars assign varying names to the same ceramic types or express similar concepts in different ways. This, in turn, makes it difficult for others to navigate the extensive body of existing knowledge (Azzopardi, 2006: 30). The same challenge applies here. If we aim to create reusable, exportable models, ensuring consistency in classification criteria is essential.

The solution devised for this dissertation was the Parameter of Archaeological Identification (PAI), which has the format of a text file attached to each model trained during the project. This file specifies and thoroughly describes the archaeological parameters applied during the labeling stage of said model. In addition, the PAI also helps to further classify and track all the trained models by expanding model nomenclature. Each trained model is now assigned both a PCI and a PAI classification. The PCI categorizes the model based on its complexity (N for nature, S for state, T for typology). The PAI further differentiates models within these categories by adding a

[67] This is just one example of a minor difference in opinion that can lead to variations between trained models. Often, these differences are more significant and accumulate by the dozens, resulting in an exponentially larger effect.
[68] From personal experience, this becomes very hard once a few models been trained unless each model's labelling stage is thoroughly documented using the PAI.

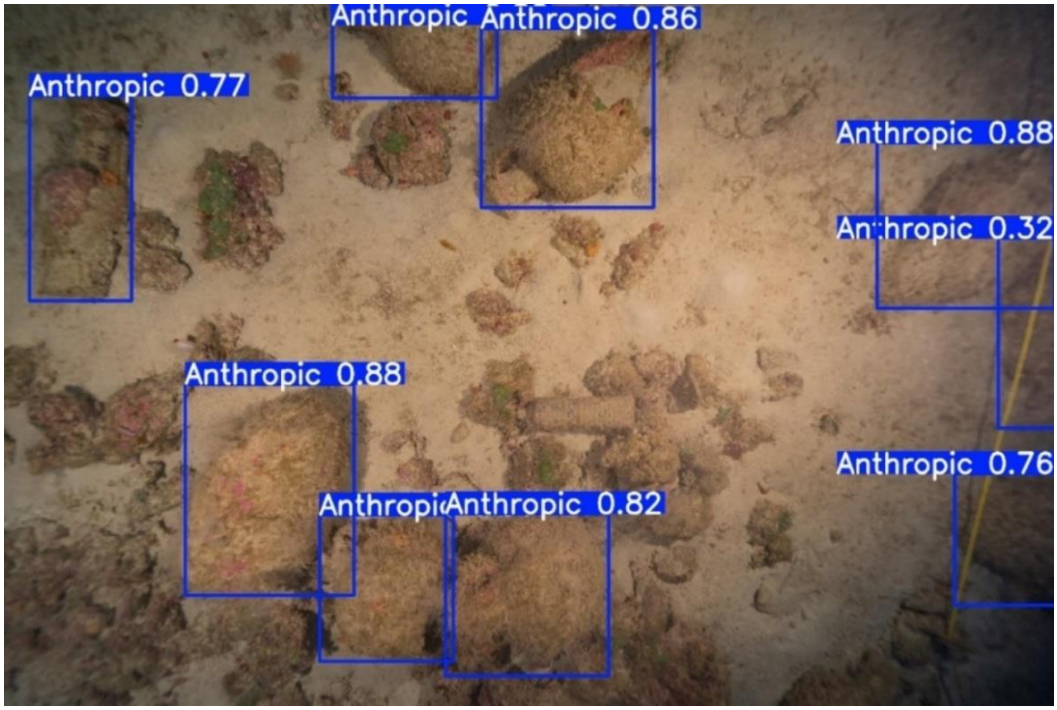


Figure 24. Example of the output on N3 model.

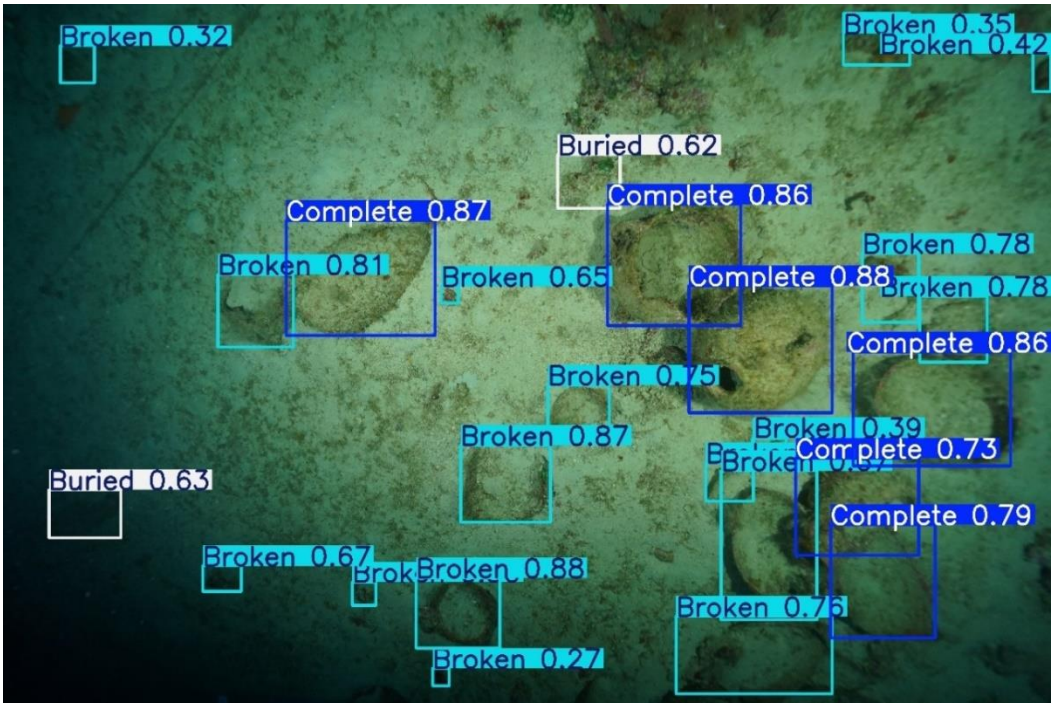


Figure 25. Example of the output on S1 model.

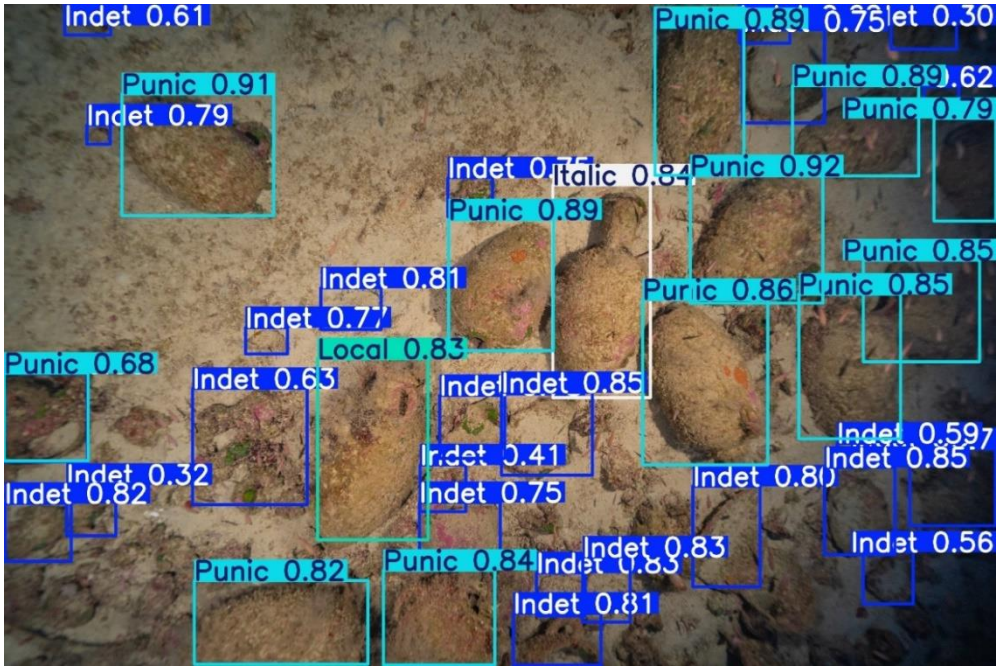


Figure 26. Example of the output on T4 model.

sequential number corresponding to the specific archaeological labeling parameters applied. For example, 'S2' represents a model trained using the second set of PAI for models with PCI of ‘S’ (or state models). This system provides a flexible and effective method for ensuring that each model's output can be accurately interpreted and referenced throughout the project at both general and class levels.

5.3.2.1 Interpretation at General Level

At the general level, the PAI considers the user’s criteria for general classification problems during the labeling process, addressing challenges such as objects partially visible at the frame edges or poorly illuminated by divers. While a general interpretative framework is not required for a project aiming to create a single model, we want to highlight its usefulness when integrating subjective archaeological information during training.

The first thing to consider when talking about the information provided with each PAI is the following: why are general level interpretations even necessary? In object detection projects conducted in terrestrial settings, where colors and textures are less affected by environmental conditions and image quality is generally higher, a common rule of thumb is to consider each

object in an image as if it were fully visible. For example, in a model that is being trained to identify different animal species, if an animal is partially cut off by the image frame, the human operator can still infer its full shape and verify their classification using adjacent images during the labelling stage to set the ground truth. For its part, during training, the algorithm will use features such as fur color, texture, body shape, and size—which the human cannot use to the same level—to arrive at the same conclusion the user did without having to "cheat" by looking into other images.

Underwater archaeological assemblages like Xlendi Archaeological Park present a different challenge. Our dataset primarily consists of amphorae, which are highly uniform in size, shape, and features. Poor image quality[69] and the maritime environment further homogenize their appearance, making distinctions even more difficult. In this context, human users can still rely on multiple images from the same photogrammetry to confirm an object's classification—just as it happened on the terrestrial example. Also like in the previous example, this is something the algorithm cannot do. The difference is that the objects under consideration now—the amphorae—all share very similar characteristics. There are not very pronounced differences in colour, shape, size or texture, and the algorithm has an even harder time telling them apart when they appear cut or obscured by their position on the image. This creates a significant issue: on a situation where a user might have classified an amphora as 'Malta 1' because they saw a different image where the defining characteristics of that type are clearly visible, the model will learn to apply this class even in cases where, based on the available visual data alone, the object should be labeled as 'Indetermined' according to the model's own predictive features. This happens because the model processes images by image and cannot use contextual reasoning to verify using other images, and since its predictive capabilities are diminished by the environment, they rely entirely on the ground truth. Hence, in this example in which the users are helping themselves with other images during labelling, they run the risk of the model associating indistinct, featureless, or poorly illuminated fragments with the 'Malta 1' typology, potentially compromising the model's accuracy.

This raises an important question: should we adhere to the standard approach used for training object detectors, or should we intentionally adjust the classification process by labeling all featureless, frame-cut materials as 'Indetermined,' even when our expertise suggests otherwise?

[69] Examples of the image quality of underwater assemblages when taken manually and unrefined without color-correcting software (p.49).

Before answering, it's crucial to acknowledge that such an adjustment could be interpreted by non-archaeologists as introducing bias into the model's training.[70]

In this thesis, the general rule of thumb for object identification can be followed without issue by utilizing a much larger dataset to normalize the predictive spikes caused by these ambiguous cases. Since the goal of this dissertation is not to produce a field-ready model optimized for maximum precision, we opted to adhere to this rule for simplicity purposes. The impact of this decision on precision during testing will likely be negligible for nature models, mild for state models, and potentially significant for typological models, where certain typologies have very few examples in a dataset consisting of naturally overlapping photogrammetric data.[71]

5.3.2.2 Interpretation at Class Level

At class level, the PAI outlines the interpretative nuances for each class present in the model, describing the archaeological parameters that led to the identification of objects of the same class. The more detailed and specific the descriptions, the better they will aid during implementation and interpretation.

In summary, the PAI serves a triple purpose. First, it assists in differentiating and classifying models, ensuring consistent organization. Second, it provides a solution to the previously discussed interpretative challenges underwater archaeology models face. Third, it adds a layer of complexity

[70] In this context, it is important to clarify that **bias**, as commonly understood, refers to a tendency to favor certain outcomes or perspectives systematically rather than randomly. However, in archaeology, bias does not necessarily imply distortion or unfairness; rather, it refers to the interpretive lens through which evidence is analyzed. The use of hypothetico-deductive reasoning in archaeology allows for a structured approach to answering research questions, where evidence is evaluated from multiple perspectives within an established theoretical framework. The key point is that researchers must define their perspective or theoretical standpoint **at the outset**, ensuring that their interpretation is transparent and justifiable—even if others might interpret the same evidence differently. In this sense, while biases (whether cultural, theoretical, or interpretive) might influence conclusions, they are not inherently problematic as long as they are acknowledged and systematically applied within the reasoning process.

Therefore, in this case, and in contrast to computer vision, bias is not about distorting data. It is about the interpretive framework used by the researcher to understand and explain the complexities of the past. This is the role of the PAI: to express and describe this archaeological lens.

For more information on the ethical implications of AI in archaeology, refer to the dedicated bibliography for general ideas: a) regarding external understanding of AI (Mitchell, 2009; Lake, 2014; Casilli, 2019; Casini et al., 2021; Argyrou, 2022; Gonzalez, 2024) and b) for the issue of AI-induced bias in archaeology (Barriedo et al., 2020; Gattiglia,, 2020; Fisher et al., 2021; Bender et al., 2021; Bickler, 2021; Cobb, 2023; and Clavert and Gensburger, 2023).

[71] This is an issue that will have to be considered on any project making use of trained detection models (the aim of this dissertation is to assess the methodology on Xlendi Archaeological Park and does not pretend to be creating trained models ready for field implementation).

by enabling the design of specialized models that can address specific archaeological questions. The different PAI created in this dissertation were designed to demonstrate all these functions; however, it is important to note that the potential for developing new ones is vast. In total, as shown in Figure 23, only four different PAI were deemed sufficient to serve as examples for each PCI.

5.3.2.3 PAI used in Nature Models: N1, N2, N3 and N4

These models typically consider just one or two classes, primarily focusing on distinguishing archaeological materials from the natural background. At the general level, their PAI have minimal impact, as the presence of only one archaeological class means that the issue of materials being cut by the image frame is less meaningful. However, it is at the class level where their PAI begins to make a meaningful contribution to model design. These variations are described below.

-**N1** models (Table 1a) were developed as a working baseline, inspired by the models of Paraskevas et al. (2023). They were trained to identify all visible instances of archaeological materials, regardless of size, and are particularly valuable when the goal is to assess the raw density of material on a site or to gain an overall perspective of the material in relation to the background environment. They can also be considered as models providing insight into the number of identifiable specimens on a site. In truth, they are providing a visual representation of the site's Number of Identifiable Specimens (NISP)[72] on an image-to-image basis.

-**N2** and **N3** models (Tables 1b, 1c,) differentiate themselves from the **N1** models by applying stricter criteria for the classification of materials within their single class. These models progressively filter out smaller ceramic sherds with unclear origins (out of context in archaeological terms). In theory, this makes them better suited for answering questions about the number of individual pots at a site. Specifically, **N3** models are designed to provide a perspective on the Minimum Number of Vessels (MNV),[73] offering a more refined count of individual vessels present on an area or site (Figure 24). The use of these two PAI was also expected to enhance the

[72] In ceramics studies, the **Number of Identifiable Specimens** refers to the count of all individual, ceramic fragments or sherds. It helps quantify density and abundance of such finds across different situations (Marshall and Pilgram, 1993:261).

[73] In ceramic studies, the **Minimum Number of Vessels** refers to the lowest possible number of individual vessels out of an assemblage. It helps to reconstruct the number of vessels on an assemblage full of fragmented material (Marshall and Pilgram, 1993:261).

model's performance by reducing the influence of the small and visually diffuse ceramic sherds that resemble background elements and often interfere with predictions.

- **N4** models (Table 1d) differ from the previous ones in that they incorporate two classes: 'ceramic' and 'litter.' The first is identified in the same way as in the **N1** models, focusing on archaeological materials. The introduction of the 'litter' class aims to evaluate how adding classes with low presence influences the overall performance of a model that is already functional. In essence, these models should provide insights into the possibilities of distinguishing between genuine archaeological material and non-archaeological debris in the same model**.**

PAI N1	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (1)	
Anthropic	Models trained with this PAI consider every ceramic element under the same 'anthropic' class.

a)

PAI N2	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (1)	
Anthropic	Models trained with this PAI consider as 'anthropic' all ceramics **with a defined shape.** This means that pieces belonging to elements like rim and base are considered, but loose sherds with no discernable origin are ignored.

b)

PAI N3	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (1)	
Anthropic	Models trained with this PAI consider as 'anthropic' only those materials with a defined shape and/or more than 50% of their total surface. Everything else is not considered.

c)

PAI N4	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (2)	
Ceramic	This model considers instance of archaeological material under the same 'ceramic' class.
Litter	Plastic refuse and glass bottles.

d)

Table 1. Summarized examples of PAI text files for N1 models (a), N2 models (b), N3 models (c) and N4 models (d). Notice the general/class level divisions.

5.3.2.2 PAI used in State Models: S1, S2, S3 and S4

The state models tested during this dissertation are more complex than the nature models, with two to four classes. While the general approach remains consistent across nature models, state models vary at the class level as well, incorporating a more nuanced interpretative framework to address the complexities of introducing the concept of state of preservation into the models' training process.

-**S1** models (Table 2a, Figure 25) include three classes: 'complete', 'buried' and 'broken'. These were developed as an initial attempt to classify visible archaeological materials based on subjective archaeological information, specifically regarding their in-situ state of preservation.

-**S2** and **S3** models expand on the **S1** models by altering the interpretative parameters used to classify the materials. While they retain the same classes, adjustments were made to achieve better visual separation between them and minimize background interference with the goal of positively influencing model performance. Specifics on these changes are given in Tables 2b and 2c.

-**S4** models (Table 2d) aim to further enhance model performance by simplifying the state of preservation classification from three classes to two. In this model 'buried' and 'broken' materials are merged into a single class 'buried-inc.', which streamlines the categorization process. By eliminating distractions from the other states, these models are intended to be particularly useful when the user's attention is focused specifically on 'complete' materials.

PAI S1	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (3)	
Complete	Materials that mantain their general shape.
Buried	Materials that mantain their general shape but more than **50%** of their total surface remains underground.
Broken	All other ceramic materials.

a)

PAI S2	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (3)	
Complete	Materials that mantain their general shape and at least two features (handles,base, rim, large portion of the body).
Buried	Materials with more than **50%** of their total surface still underground.
Broken	All materials that noticeably don't mantain their general shape including those that, while buried, show clear evidence of breakeage. Very small fragments without context are not considered.

b)

PAI S3	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (3)	
Complete	Materials that mantain their general shape, two features (handles,base, rim, large portion of the body), and at least **60%** of their surface visible above ground.
Buried	Materials with more than **40%** of their surface still underground.
Broken	Materials that, while not mantaining their general shape, are representative of the presence of a single vessel. Everything else, including smaller sherds and loose, shaped fragments, are not considered.

c)

PAI S4	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions.	
CLASS LEVEL (2)	
Complete	Materials that mantain their general shape, two features (handles,base, rim, large portion of the body), and at least **60%** of their surface visible above ground.
Buried-Inc	Every other ceramic material.

d)

Table 2. Summarized examples of PAI text files for S1 models (a), S2 models (b), S3 models (c) and S4 models (d). Notice the general/class level divisions.

5.3.2.3 PAI used in Typological Models: T1, T2, T3 and T4

From this project's outset, training models capable of processing typological information were loosely considered to be an almost impossible task for a YOLO model without the help of additional techniques to enhance its capabilities. Given the vast complexity of typology studies, implementing multiple typologies into detection classes requires significantly more precise interpretative parameters than those used in nature or state models, not to mention the 18 different ones identified for the site. In this regard, many of these "typological" classes will differ from each other only in subtle nuances, like unique recognizable features in certain parts of the vessels that will be heavily homogenized by preservation and environmental factors.

However challenging, the possibility of such a model had not been systematically tested before. A key part of this project was to attempt it, not necessarily to achieve a fully functional model, but to assess how far we are from that goal.

The examples of possible PAI for these models range from 5 to 16 different classes:

-**T1** and **T2** models (Tables 3a, 3b) both implement the site's complete typological chart as classification categories with one key distinction: **T1** includes an additional class 'Indetermined' to account for archaeological materials without a clear typology.[74] In contrast, **T2** models consider only objects with a defined type. Both PAI were designed as a baseline for further tests, aiming to assess how the inclusion or exclusion of an 'Indetermined' category impacts the performance of a typological detection model.

-**T3** models (Table 3c) were designed to try to improve the performance shown by **T1** and **T2** models by merging classes that have both similar significance and physical attributes. For example, different versions of functionally equivalent Ramon Punic types were grouped together. As a result, these models have 13 classes.

-**T4** models (Table 3d, Figure 26) were trained following the same principle of merging similar classes used on the previous typological PAI but using a different approach. Instead of grouping types based purely on typological similarity, **T4** models consider geographically induced dysmorphism between typologies.[75] As a result, they feature only five classes: 'Indetermined,'

[74] There are only 16 classes because two typologies from the catalogue (Forms 8B and 11b) are less represented variations of more general ones with whom they share most contextual characteristics.

[75] Typologies from certain geographic-cultural contexts tend to share physical aspects that can be used as a way for detection algorithms to tell them apart.

‘Punic,’ ‘Italic,’ ‘NAfrican,’ and ‘MalteseLocal.’ This makes **T4** models a preliminary attempt at integrating the concept of geographical origin into an object detection model (thus the name of the classes).

PAI T1	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions. **Material is only identified as a specific type when absolutely certain**. This applies to every type save the Punic Form1 vessels. These are overwhelmingly prevalent on the site, and so many buried materials showing the general features of Form1 Punic vessels have been considered as such.	
CLASS LEVEL (16)	
Indetermined	Every piece of archaeologica material not identifiable with a type.
Form1	Form 1 directly from the site's catalogue.
Form2	Form 2 directly from the site's catalogue.
Form2b	Form 2b directly from the site's catalogue.
Form3	Form 3 directly from the site's catalogue.
Form4	Form 4 directly from the site's catalogue.
Form5	Form 5 directly from the site's catalogue.
Form6	Form 6 directly from the site's catalogue.
Form7	Form 7 directly from the site's catalogue.
Form8	Forms 8 and 8b directly from the site's catalogue.
Form9	Form 9 directly from the site's catalogue.
Form10	Form 10 directly from the site's catalogue.
Form11	Forms 11 and 11b directly from the site's catalogue.
Form12	Form 12 directly from the site's catalogue.
Form13	Form 13 directly from the site's catalogue.
Form14	Form 14 directly from the site's catalogue.

a)

PAI T2	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions. **Material is only identified as a specific type when absolutely certain**. This applies to every type save the Punic Form1 vessels. These are overwhelmingly prevalent on the site, and so many buried materials showing the general features of Form1 Punic vessels have been considered as such.	
CLASS LEVEL (15)	
Form1	Form 1 directly from the site's catalogue.
Form2	Form 2 directly from the site's catalogue.
Form2b	Form 2b directly from the site's catalogue.
Form3	Form 3 directly from the site's catalogue.
Form4	Form 4 directly from the site's catalogue.
Form5	Form 5 directly from the site's catalogue.
Form6	Form 6 directly from the site's catalogue.
Form7	Form 7 directly from the site's catalogue.
Form8	Forms 8 and 8b directly from the site's catalogue.
Form9	Form 9 directly from the site's catalogue.
Form10	Form 10 directly from the site's catalogue.
Form11	Forms 11 and 11b directly from the site's catalogue.
Form12	Form 12 directly from the site's catalogue.
Form13	Form 13 directly from the site's catalogue.
Form14	Form 14 directly from the site's catalogue.

b)

Table 3. Examples of PAI text files for T1 models (a) and T2 models (b). Notice the general/class level divisions.

PAI T3	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions. **Material is only identified as a specific type when absolutely certain**. This applies to every type save the Punic vessels. These are overwhelmingly prevalent on the site, and so many buried materials showing the general features of Punic vessels have been considered as such.	
CLASS LEVEL (13)	
Indet	Every piece of archaeologica material not identifiable with a type.
Punic	Forms 1,2 and 2b directly from the site's catalogue.
Form3	Form 3 directly from the site's catalogue.
Form4	Form 4 directly from the site's catalogue.
Form5	Form 5 directly from the site's catalogue.
Form6	Form 6 directly from the site's catalogue.
Form7	Form 7 directly from the site's catalogue.
Form8	Forms 8 and 8b directly from the site's catalogue.
Form9	Form 9 directly from the site's catalogue.
Form10	Form 10 directly from the site's catalogue.
SmallA	Forms 11, 11b and 12 directly from the site's catalogue.
Form13	Form 13 directly from the site's catalogue.
Form14	Form 14 directly from the site's catalogue.

a)

PAI T4	
GENERAL LEVEL	
Every object is considered absolutely. This means that several images may have been used to determine object class. It includes those objects cut by the frame and those under precarious lighting conditions. **Material is only identified as a specific type when absolutely certain**. This applies to every type save the Punic vessels. These are overwhelmingly prevalent on the site, and so many buried materials showing the general features of Punic vessels have been considered as such.	
CLASS LEVEL (5)	
Indetermined	Every pieece of archaeologica material not identifiable with a type.
Punic	Forms 1, 2, 2b, 3 and 10 directly from the site's catalogue.
Italic	Forms 4, 7 and 14 directly from the site's catalogue.
MalteseLocal	Forms 5, 8, 8b, 9, 11, 11b, 12, 13 directly from the site's catalogue.
NAfrican	Form 6 directly from the site's catalogue.

b)

Table 4. Examples of PAI text files for T3 models (a) and T4 models (b).

5.3.3 Detection model type and architectural size

Model architecture forms the last layer of model differentiation (Figure 23), as distinct versions and sizes of the YOLO model have a significant impact on prediction performance. Different versions often introduce improvements or complex structural changes to the base model, which in turn affect training, data processing, and prediction speed, ultimately influencing practical outcomes. Size variations also affect the algorithm's feature extraction capabilities[76] in ways that also alter the predictive outcome and that are more explainable.[77] For practical purposes, these can be summarize by stating that larger models generally have higher capacity, superior feature extraction systems, and improved Non-Maximum Suppression (NMS) performance than smaller models,[78] which effectively means that larger models are more computationally intensive and less suited for real-time applications due to their slower processing speeds. Summarizing, this means that larger models offer advantages such as greater precision and efficiency when handling clusters of closely located objects at the cost of speed.

Two different model versions (YOLOv8 and YOLOv11) and three different size architectures (nano, small and medium) were used to account for variations in detection performance that are expected when working with relatively small datasets such as the one used for this dissertation (Paraskevas, 2023: 5; Zammit et al., 2024: 4125). Among the different available versions of YOLO, YOLOv8 was selected for this project because it has been underlined in recent underwater assemblage detection projects as the version that yields the best results (Paraskevas et al., 2023: 4; Kamal et al., 2024: 1335; Zammit et al., 2024: 4125). YOLOv11, on the other hand, is the latest version of the algorithm, which was released in September 2024. Prior to this project, it had not been tested on underwater assemblages. It was selected for this dissertation due to the expectations

[76] **Feature extraction** is a concept related to the operational functioning of detection algorithms. An explanation and references to more information can be found in Appendix I (p.153).

[77] **Size variations** across different models primarily affect model performance, accuracy, speed, and more importantly, input resolution. In general, smaller models resize input images to 640x640 resolutions, reducing their computational requirements to increase their speed and be suitable for real-time detection. Larger models, on the other hand, can use higher resolution images to leverage their greater capacity for feature extraction. This makes them capable of processing smaller details, making them better suited for more complex objects (such as typologies). Their downsides lay in the fact that they demand more GPU memory to work, which in turn limits their real-time applications while requiring higher-end hardware (Bochkovskiy et al., 2020).

[78] **Non-Maximum Suppression** is a post-processing detection technique present in many different models as a way to eliminate duplicate and overlapping bounding boxes by choosing the most relevant one in terms of model confidence (Neubeck and Van Gool, 2006:1-2). An explanation of how it works and more information about it can be found in Appendix I (p.153).

of it being an improvement over earlier versions (according to Ultralytics, the developer) and to serve as a counterpart for YOLOv8.[79]

Six combinations of model version and size were tested for each PAI. These were subsequently incorporated into model nomenclature (e.g., model T3 YOLO11s = typological model trained with the third PAI defined for typological PCI and run on a 'YOLO11s' frame).

5.4 Data Analysis

Evaluating the capabilities of detection models as a tool for a maritime archaeology project is not a straightforward task. There is a learning curve involved that becomes steeper the more we consider additional supporting elements and techniques that could be integrated with DL[80] methods to improve model performance (Kamal et al., 2024; Zammit et al., 2024). However, given the aims of this project, for the purposes of this methodology and experiments, the factors represented by these supporting elements or techniques were set aside. The focus was narrowed to establishing the basic uses and limitations of detection when applied without project-specific systems, ensuring the approach remains replicable by anyone with basic computer knowledge.

To analyze the data generated by the trained models, the evaluation process was structured as a two-stage approach (Figure 18). The first stage involved a visual test in which the trained models under scrutiny were used to generate predictions on a set of images that had not been seen by the trained model before. This stage replicates real-world conditions and was essential for confirming the proper functioning of models. The second stage of evaluation consisted of a comparative analysis of the performance metrics produced by each group of trained models during testing. In this section, we will explain both stages in detail.

5.4.1 Visual evaluation. Model validity

In this dissertation, the archaeological validity of a trained detection model is considered as the model's ability to produce useful archaeological information. While it is true that the general validity of trained detection models used for other purposes can be established from their raw

[79] More information can be found on the official YOLOv11 directory: https://docs.ultralytics.com/models/yolo11/
[80] Deep learning (p.11).

metrics, this criterion must be adjusted in the context of archaeology, where archaeological validity takes precedence: some models, despite having poor mathematical performance, can still provide valuable archaeological insights or reveal issues that can guide improvements in model design. Conversely, some models that seem to produce precise results in terms of metrics may not yield archaeologically valid data. Both scenarios occurred during this dissertation. It was because of this that we decided that, before evaluating raw metrics, models first had to undergo visual testing.

For the purposes of visual testing, the only way in which variations in model architecture affect a model's output is in raw metrics. They have no effect on whether a model is working as intended or not.[81] As a result, of all 72 models, only the 12 best-performing (one from each PAI) needed to be evaluated this way. To provide a well-rounded assessment, this visual test was conducted in two parts. The first part involved generating predictions on a set of 200 unseen images that were specifically set aside during training to simulate real-world model implementation.[82] The second part tested the model on a video created by stitching together unseen images from yet another photogrammetric run that also had not been used in the dataset.[83] Both the resulting images and video were then evaluated to preliminarily confirm that the trained model performed as expected based on the following criteria:

-Proportion of missed (not located) materials of significance.

-Proportion of misidentification among similar objects.

-Proportion of misidentification between distinctly different objects.

-The model's ability to consistently track correct identifications across multiple perspectives of the same object (i.e., the same object across multiple slightly overlapping images).

5.4.2 Performance evaluation. Metric comparison

The evaluation and comparison of the performance of different trained models is a crucial focus of this research. While the visual evaluation provides a rough estimation of a model's capacities, the mathematical data from the tests results allow for a more precise performance evaluation based on all the factors differentiating our models (PCI, PAI and model version/size). During testing,

[81] We can consider that a model is working as intended when it is trying to classify the archaeological materials following our exact design. Whether it succeeds or not is another matter entirely.
[82] Test set (p. 41-2).
[83] Examples of the videos are available in Appendix V (p. 214).

trained models generate predictions on the images from the test set and compare them to that set's ground truth (introduced by the user during the labelling stage) to measure their own performance.[84] We tested out all our models this way, with the resulting measurements being our results or performance metrics.

However, the performance metrics from the first tests closely mirrored those inherently produced by the models during training. It was thus that it became clear that since all images had originated from the same photogrammetric survey—carried out by a single diver following a consistent trajectory and altitude—the dataset lacked sufficient variation. This raised concerns about the model's ability to generalize beyond these conditions, prompting a search for a solution.

Two different approaches emerged from the literature review as a precedent for this situation. The first, presented by Zammit et al. (2024), was a computer vision-driven study focusing on comparative analysis, specifically assessing the impact of a precision enhancing technique on YOLO's performance. Working on images from Xlendi Archaeological Park as well, their comparative evaluation had been carried out using the data achieved from testing model performance in a set of metrics extracted from the same area and photogrammetric run as the testing just as we had. To clarify this choice, the lead author was contacted. He explained that the decision was based on the study's primary objective—mathematical comparison. This is to say that while the test conditions did not fully reflect real-world performance, they provided a controlled environment in which the effects of depth-channel augmentation could be more effectively measured through the proportionally superior differences to be found in higher numbers/metrics.

The second approach, outlined in an archaeologically focused study by Paraskevas et al. (2023), took a different route. Instead of using test images drawn from the same area as the training set, this study evaluated YOLO models on an AUV[85] system using images from a completely different location of the same site. The goal was to simulate real-world conditions, forcing the model to encounter variations in object arrangements, environmental factors, and visual features (Paraskevas et al., 2023: 4). As the authors had expected, the test metrics were significantly lower

[84] The testing set, as the training set, is accompanied by our labels specifying the location and classification of its objects. The difference is that the testing set was set apart at the beginning to serve as reliable data that simulates a practical scenario. The trained model tests its predictions on these fresh images (p.41-42).
[85] Autonomous Underwater Vehicles (p.8).

than those recorded during training. However, they provided a more realistic representation of the model's actual performance in the field.

The difference between the two approaches appeared to be in their focus: the first, rooted in computer vision science, prioritized high-performance metrics for comparative analysis while the second, an archaeological study, emphasized real-world applicability because the author was trying to make practical use of the methodology on an assemblage.

For this research, despite having a strong comparative component at its core that would have benefited from Zammit et Al's approach, we deemed the second option—testing the models on entirely unseen data—entirely more adequate to the pursuit of a set of aims framing what is, in truth, an archaeological project. Moreover, to explore the model's limitations and applicability, testing needed to take place in a practical scenario; otherwise, the results would have lacked meaningful real-world relevance. Following this decision, a new test set of 200 images was created using data from a third photogrammetry survey of area of the site totally removed from the original dataset.

Once all metrics for all trained models were obtained with this new test set, a comparative analysis was conducted at both a broad and specific levels. At a broad level, performance variations were assessed in relation to the models' PCI. At a more detailed level, a PAI-based analysis was carried out to better understand how archaeological interpretations and training parameters influenced model performance. Finally, the algorithm's version and size were considered as well.

These were the crucial questions that guided our comparative analysis:

1. How does model performance evolve at the PCI, PAI, and variant/architecture levels? What does this reveal about the models' capabilities?
2. Is it possible to introduce highly subjective archaeological information into a model's training stage?
3. Can the experiment's results be improved? If so, how?

5.5 Typological Chart

As part of the evaluation of detection methods in maritime archaeology assemblages, training models for automatic typological classification presented a suitable challenge. To implement this

idea, a typological chart of the site had to be developed. This section explains the methodology used in constructing the typology, with the complete chart available in Appendix III.[86]

5.5.1 Identification of Ceramics

The primary identifying features of ceramic vessels are their shape and fabric (Frendo, 1988: 119; Williams, 1986: 4). Fabric analysis techniques can be broadly categorized into physical analysis, which examines inclusions in the clay, and chemical analysis, which detects less conspicuous factors visible only through microscopy. Both methods require direct access to the artefacts and specialized expertise in geology. Since Xlendi Archaeological Park lies at a depth exceeding 100 meters, these were two resources unavailable to us for this study. As a result, the identification of ceramics in this assemblage had to rely entirely on the study on the shape of the materials.

The study of ceramic shapes, as every vessel was assembled from distinct components, reveals a remarkable diversity in body forms, rims, bases, necks and handles many other features. To organize this complexity, archaeologists classify ceramics into typologies, grouping them based on shared morphological characteristics. While this approach has been (justifiably) criticized for being subjective and sometimes confusing (Ramon Torres, 1995: 158; Azzopardi, 2006: 31), visual comparison with established prototypes remains the most practical and widely used method for identifying ceramic objects in newly discovered assemblages (Ramon Torres, 1995; Sagona, 2002; Atauz, 2004; Anastasi, 2019). In our case, this translated to analyzing a dataset comprised of thousands of images taken by divers.

While a study relying on photography might seem limited to distinguishing fine details, the sheer volume of high-quality imagery,[87] coupled with the site's depth—where vessels remain fairly undisturbed by biota—allows for clear, museum-like visualization of the spread assemblage, providing ample data for **accurate typological comparisons**.

---

[86] Appendix III (p.175).

[87] We have stated before how low image quality difficulties the algorithms of detection models during the feature extraction (finding and classifying objects). This remains true. The reduced quality of the photos comes from the wieldier versions of YOLO models themselves resizing the original images to 640x640 pixels. The images of our dataset, though belonging to photogrammetric runs, are of very high quality. These are the ones we used for ceramic identification.

On the topic of the materials themselves, most of the ceramics from Xlendi Archaeological Park are amphorae. Designed primarily for the marine transport of goods, amphorae stand out among ceramics in the maritime context as they enable us to infer connectivity between different regions, which carries significant implications: A well-established typology can reveal cultural connections, place of origin, date of manufacture, likely original contents, and facilitate comparisons with similar examples to understand distribution patterns and even the evolution of manufacture techniques (Peacock and Williams 1986; Rice, 1987; Orton et al., 1993; Tweed, 2002).

However, such an interpretive study of the Tower Wreck's ceramics falls outside the scope of this dissertation. This research is focused solely on the physical attributes of the items and the development of a preliminary typology to inform the detection methodology. Thus, a simple analysis of shape and features, using visual comparisons of all the site's data, was considered sufficient for identification purposes.

5.5.2 Constructing a Catalogue

During the identification of the ceramics of Xlendi Archaeological Park, we relied almost exclusively on the catalogues of Ramon Torres, Claudia Sagona and Elaine Azzopardi (Ramon Torres, 1995; Sagona, 2002; Azzopardi 2006). Given that most of the site's material is of Punic and early Roman origin, the works of Ramon and Sagona were particularly relevant and complementary. The first focuses on Punic amphorae produced and distributed in North Africa and the Western Mediterranean. Sagona's is a focused study on the local evidence of Punic material, showing some intersectionality with Ramon's work and covering all the necessary bases for a comprehensive identification process. In addition, as a study of the underwater ceramics of the whole of Xlendi Bay, Azzopardi's work was an invaluable reference on its own. Further information was obtained from other catalogues[88] and online repositories such as the Archaeology Data Service.[89]

[88] These catalogues were used to confirm and reference the identification parameters of some specific finds: Beltrán Lloris, 1970; Will, 1982; Freed, 1996; Aubet and Barthelemy, 2000; Bruno, 2000; Bruno and Capelli, 2000; Atauz, 2004; Bechtold, 2010, 2012, 2018; Ben Jerbania, 2017; and Anastasi, 2019.

[89] **Archaeological Data Service:** University of Southampton (2014) Roman Amphorae: a digital resource [data-set]. York: Archaeology Data Service [distributor] https://doi.org/10.5284/1028192.

It should be noted that even when using all these different sources, we encountered issues identifying some materials from Xlendi Archaeological Park. Some pieces had shapes that were too general, and their state of preservation hindered definitive identification. On other occasions, the quality of the images, interference from biota, or dust clouds raised by the divers proved to be additional obstacles.

In the catalogue presented for this project, each identified type is accompanied by a description of its physical characteristics, its correlation to materials from other classifications, and a preliminary assessment of date, place of production and possible distribution. It is also necessary to point out that the primary focus of the ceramic study was the materials physical properties, as these are the most relevant aspects for this dissertation given the way detection models identify objects. Aside from this, both a drawing and a photograph of in situ examples are provided for each type.

6. RESULTS

Contents

In this research project, 72 trained models were evaluated. Their viability was first assessed through visual inspection, followed by a practical test to measure performance on unseen data. This chapter presents the evaluation results, highlighting key model characteristics in preparation for further interpretation.

The comparative results are structured in two parts. First, a series of tables provides a comprehensive overview of each model's performance metrics, serving as a reference throughout the chapter.[90] Then, these are analyzed and compared across three levels of differentiation: PCI, PAI, and version/size. An expansion of the metrics used during the process can be found in Appendix IV.[91]

6.1 Visual Evaluation Results

The purposes of the visual evaluation are to verify that the model is working as intended by its design and to get a grasp of its archaeological validity based on its proportion of false positives, false negatives, and the way particular items are identified across several images made at different angles. In the latter case, we are looking for consistency of identification. For example, a model

[90] Three specific metrics were chosen as the most representative of the models from an archaeological standpoint. They are average recall (AR) and two different measures of average precision by class (mAP50 and mAP50-95).
[91] Appendix IV (p.209).

that is consistently tracking the same 'Punic' amphorae as being 'Italic', 'North African' or 'Local' depending on the angle suggests a model that is not fit for field implementation.

Out of the 72 trained models, we selected 12 for a thorough visual evaluation—the best performer from each PAI.[92] The results of those evaluations are presented here, accompanied with representative images and videos.[93] To facilitate comparison on how different PCI, PAI and YOLO versions influence model predictions, the same images were used consistently to show the differences across related models as much as possible.

6.1.1 Nature models (N1, N2, N3 and N4)

Designed as a baseline for all others, nature models demonstrate consistently high confidence rates. From a visual standpoint, they make very few mistakes while correctly identifying nearly all ceramic fragments as intended. A notable trend is the increase in confidence rates from **N1** to **N3** models, suggesting improved precision, as evidenced by the scarcity of false positives and false negatives.

When tested on video, these models demonstrate the ability to track predictions across multiple frames, reinforcing their apparent reliability. They also seem highly capable of discerning small, diffuse ceramic sherds from similar elements of the background. Based on this initial evaluation, they can be considered archaeologically valid models.

-**N1** model (N1 YOLOv8s)—Figure 27.

- •**N1** models successfully identify and locate all ceramic elements within the site under a single class 'anthropic'.
- •The lack of false negatives in the images suggests very high precision and recall values.

-**N2** model (N2 YOLOv11n)—Figure 28.

- • **N2** models build upon **N1**, maintaining the single-class approach while incorporating parameter adjustments to ignore the smallest sherds deemed out of context.
- • Non-relevant fragments are effectively filtered away. This is apparent in both the images and videos.

-**N3** model (N3 YOLOv11n)—Figure 29.

[92] The parameters for the evaluation are available in p. 63.
[93] Some example videos can be found on Appendix V (p.214).

•**N3** models expand on **N2** by filtering out all non-representative broken pieces instead of only the smallest sherds.

•The models function as intended in terms of the pieces being filtered out. Even though we cannot confirm it without metrics, **N3**'s predictions look much cleaner without the confusion generated by large numbers of broken fragments. This refinement leads to an apparent increase in model confidence at identifying significant archaeological fragments.

-**N4** model (N4 YOLOv11m)—Figure 30.

•**N4** models introduce an additional class 'litter' alongside ceramic detection.

• In addition to identifying ceramic elements, these models successfully distinguish other 'anthropic' materials present on site from the ceramic ones.

Figure 27. Examples of N1 model predictions on unseen test set.

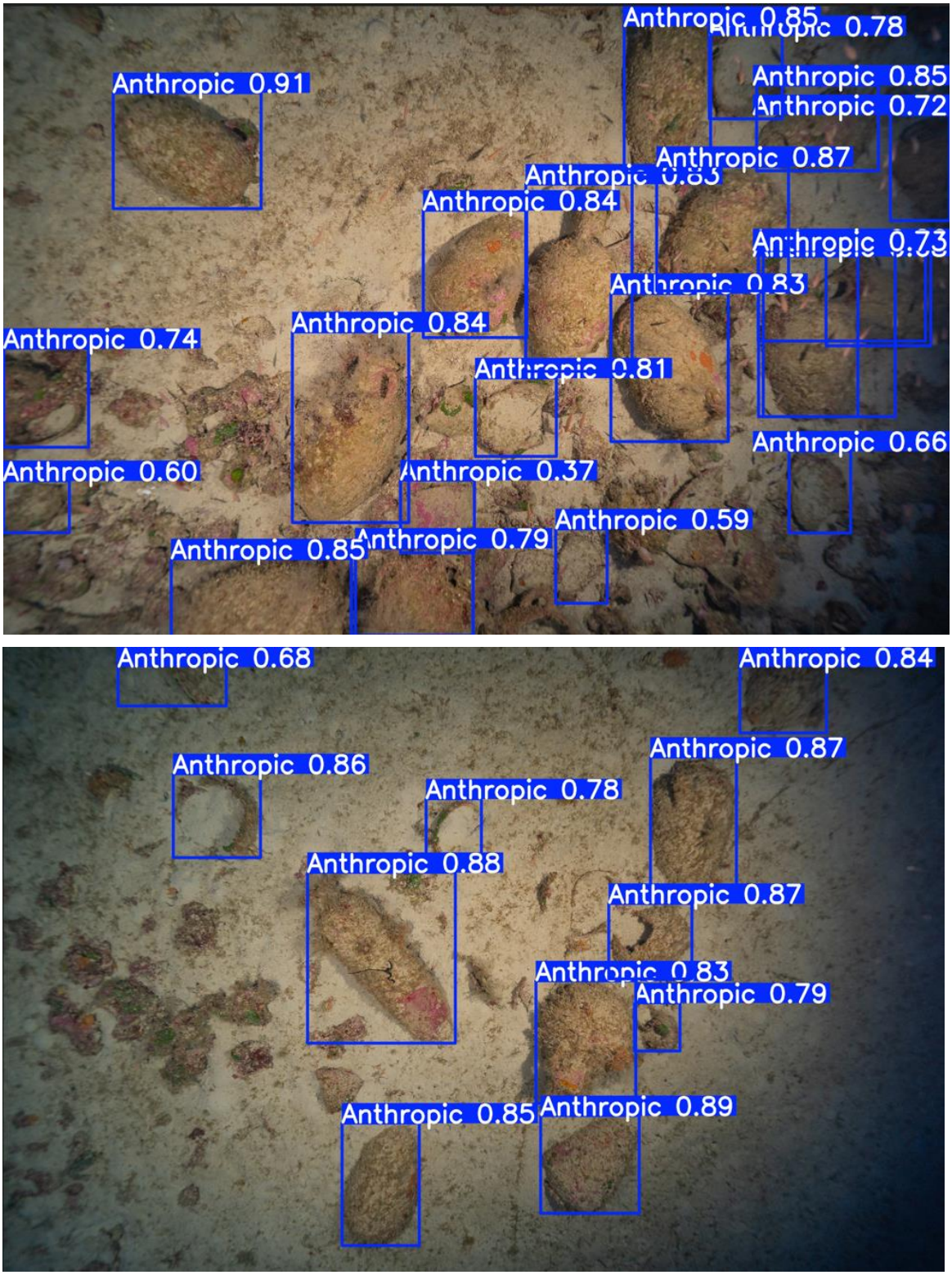


Figure 28. Examples of N2 model predictions on unseen data. Notice the selectivity the model shows when it comes to broken ceramic fragments.

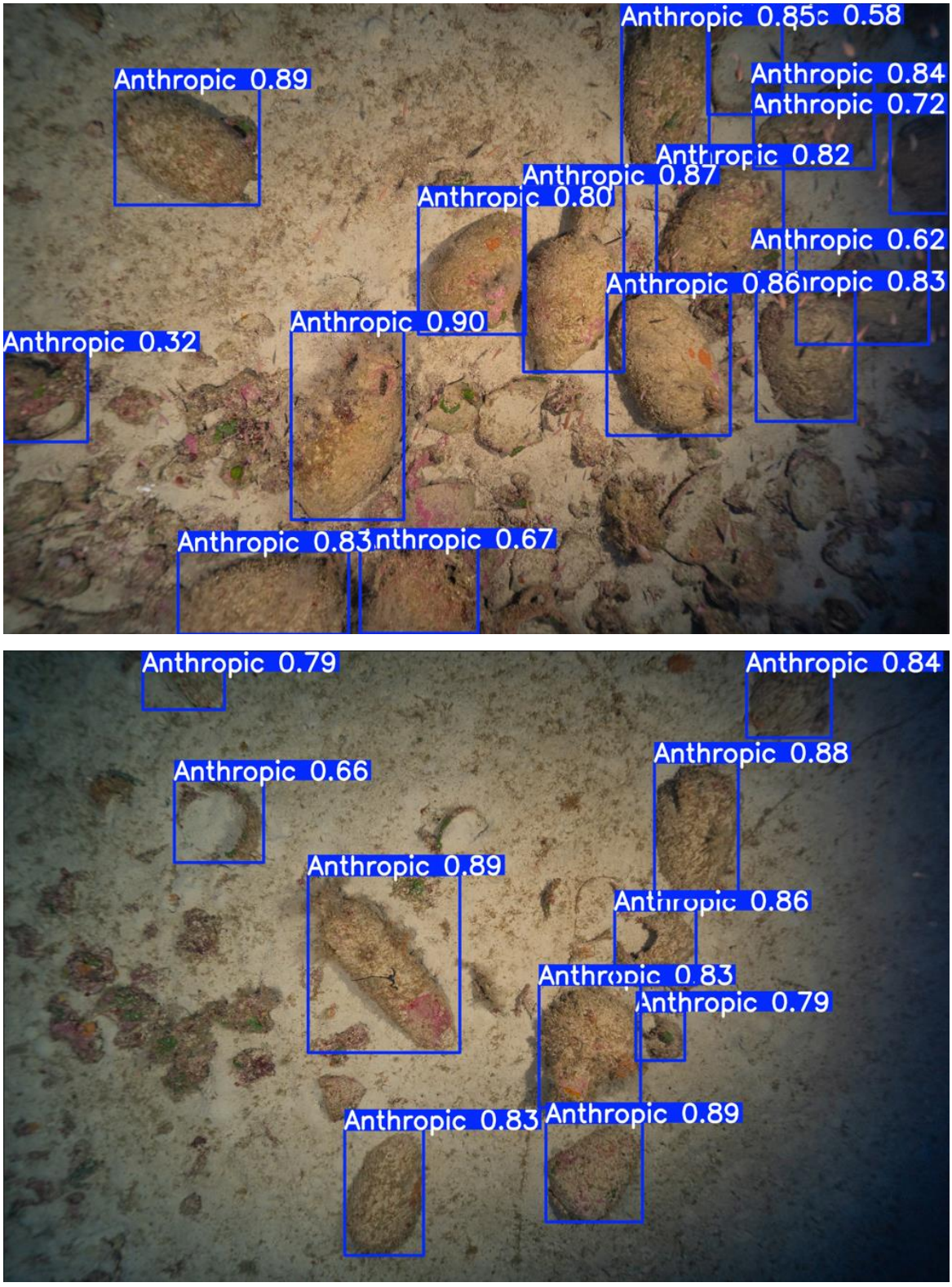


Figure 29. Examples of N3 model predictions on unseen data. Notice the increased selectivity the model shows when it comes to broken ceramic fragments in relation to N1 and N2 models.

Figure 30. Examples of N4 model predictions on unseen data.

6.1.2 State models (S1, S2, S3 and S4)

Designed as a first attempt to incorporate substantial archaeological information into a training stage, state models show generally high confidence prediction rates when evaluated visually. However, all four PAI exhibit some misclassification errors along with false positives and false negatives as well. These errors appear to increase gradually from **S1** to **S2** to **S3**, coinciding with a slight decline in identification confidence. In examples provided in Figures 32 and 33, such mistakes are highlighted in red boxes.

Despite the added complexity and the increase in number of classes, state models remain consistent in their predictions across multiple images. Based on this visual evaluation, all models function as intended and can be considered valid.

-**S1** model (S1 YOLOv11s)—Figure 31.

•**S1** models incorporate state of preservation through three classes. These are 'complete', 'buried' and 'broken'.

• Most apparent errors involve misclassification between 'broken' and 'buried' materials.

-**S2** model (S2 YOLOv8n)—Figure 32.

•**S2** models are a modification of **S1**, retaining the same classes.

•While these models also function as intended, the test images show an increased number of false negatives and false positives compared to **S1** models.

-**S3** model (S3 YOLOv11s)—Figure 33.

•As a modification of **S2**, the true performance of the **S3** models is also not fully visible in this test.

•**S3** models exhibit both more errors than the previous two models and lower confidence rates.

-**S4** model (S4 YOLOv11s)—Figure 34.

•**S4** models reduce the number of classes by merging the 'broken' and 'buried classes into a 'buried-Inc' class.

• These models show less misclassification errors than those in **S2** and **S3**.

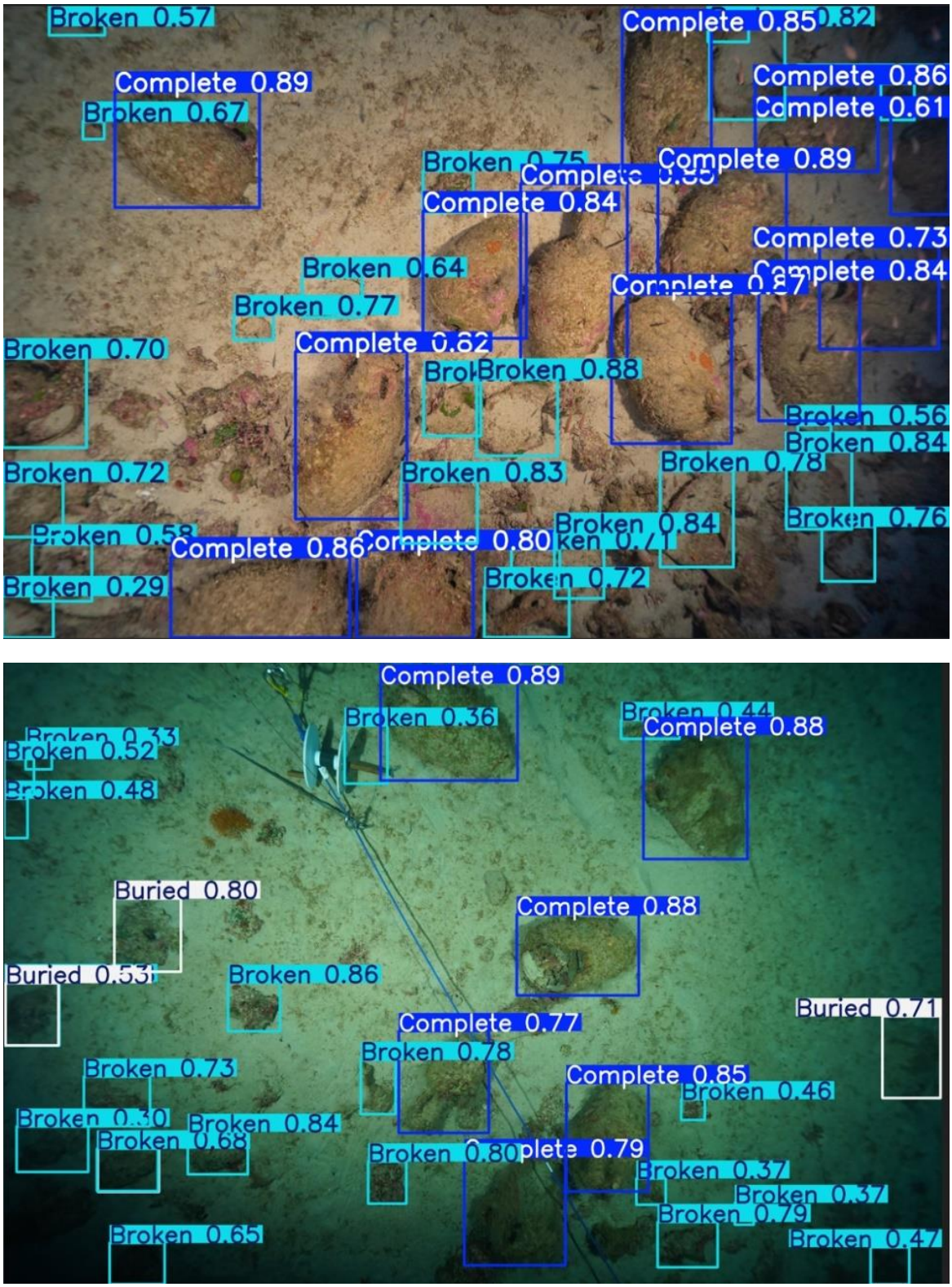


Figure 31. Examples of S1 model predictions on unseen data.

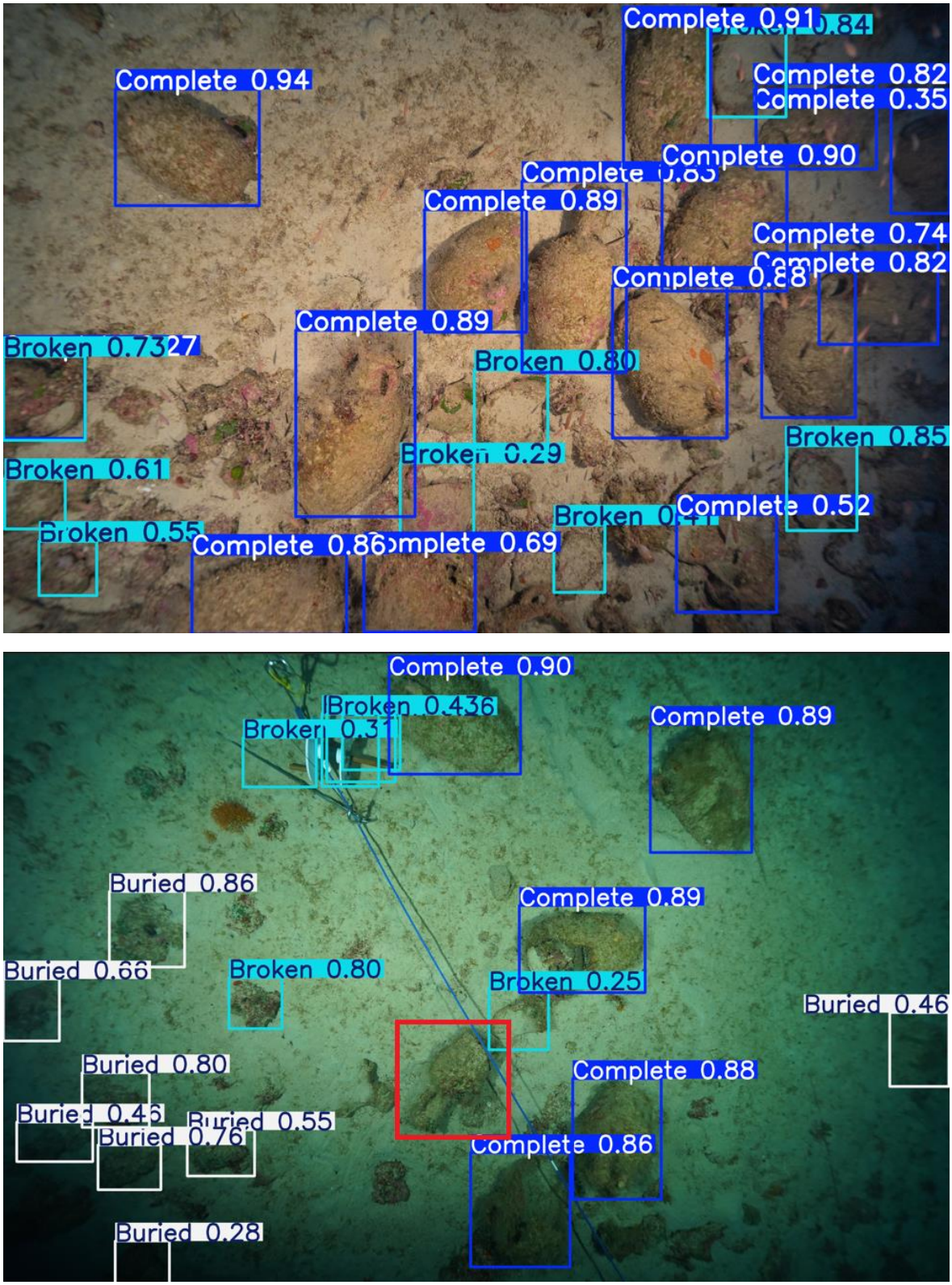


Figure 32. Examples of S2 model predictions on unseen data. Notice the selectivity the model shows when it comes to broken ceramic fragments and the example of a misidentified object (red).

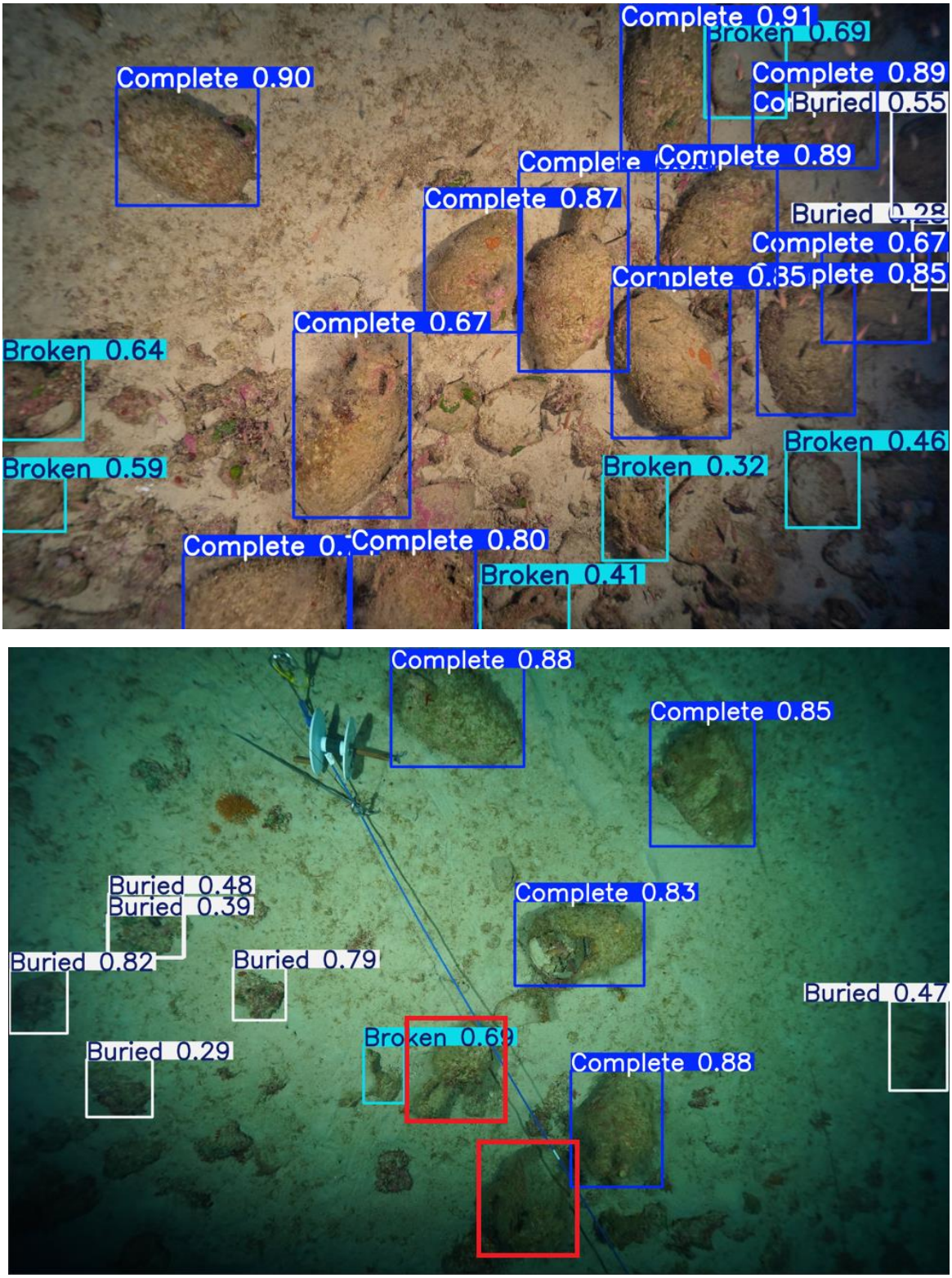


Figure 33. Examples of S3 model predictions on unseen data. Notice the increased selectivity the model shows when it comes to broken ceramic fragments in relation to S1 and S2 models and the way it reduces image clout at the expense of some misidentified objects (red).

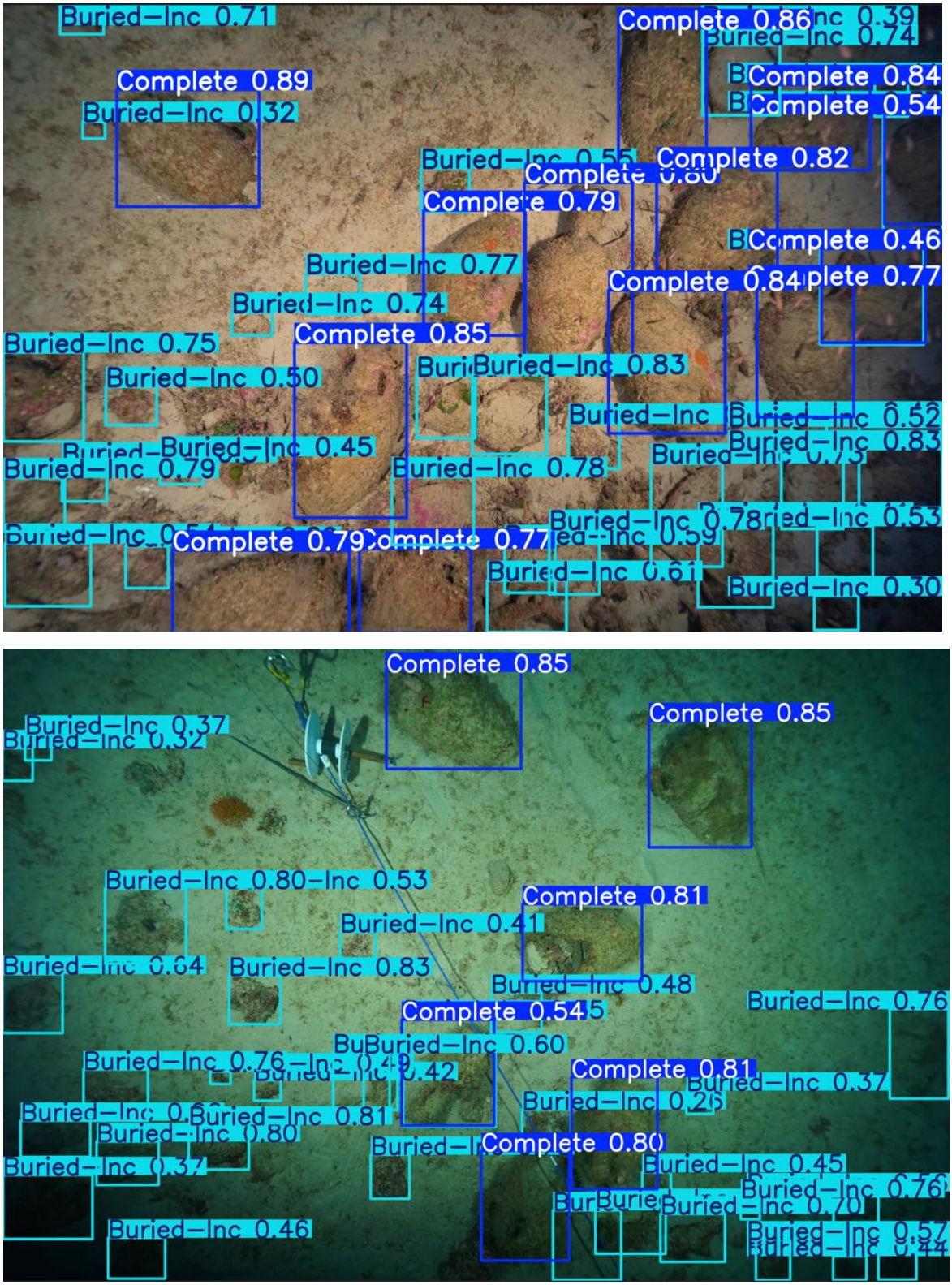


Figure 34. Examples of S4 model predictions on unseen data. Best performing models of state models.

6.1.3 Typological models (T1, T2, T3 and T4)

Typological models were designed to test the limits of the YOLO model by progressively increasing complexity through the number of classes and type of archaeological information they use. Overall, all four typological PAI exhibit numerous errors and struggle to consistently track the same object across consecutive and slightly overlapping images. Based on this visual evaluation, models **T1**, **T2**, and **T3** are not yet suitable for practical in-field implementation. However, **T4** models show noticeable improvements.

-**T1** model (T1 YOLOv8m)—Figure 35.

•**T1** models implement 16 different classes directly from the typological chart.

•They include an additional 'Indetermined' class for anthropic elements that cannot be confidently classified. Thus, every instance of archaeological material is located.

-**T2** model (T2 YOLOv8s)—Figure 36.

• T2 models use the same typologies as **T1** but omit the 'Indetermined' class, leaving ambiguous anthropic elements unidentified.

•This omission leads to even more errors, including false negatives among objects that **T1** models successfully identified (marked in red).

-**T3** model (T3 YOLOv11m)—Figure 37.

•Reintroduces the 'Indetermined' class and strategically groups similar typologies into functionally equivalent categories, reducing the number of classes from 16 to 13.

•Visually, this adjustment appears to improve performance, as the model makes fewer errors—though still a considerable number.

-**T4** model (T4 YOLOv11m)—Figure 38.

•Like **T3**, **T4** models merge elements from the typological chart to enhance classification accuracy. However, they do so based on geographic dysmorphism, apparently leading to significantly better results.

To sum up the evaluation, a visual analysis of the models' predictions underlines the increasing specificity and refinement of nature models, demonstrating their adaptability in distinguishing relevant archaeological materials from extraneous elements such as the natural background or modern additions. Attempts to similarly refine the performance of state models, however, seem to

fall short, resulting in models that do not surpass the baseline model's (**S1**) effectiveness in real-world applications. Lastly, while typological models are currently unsuitable for practical use, they show enough improvement to suggest that further refinements could lead to viable models. The following sections will further analyze their performance through quantitative metrics.

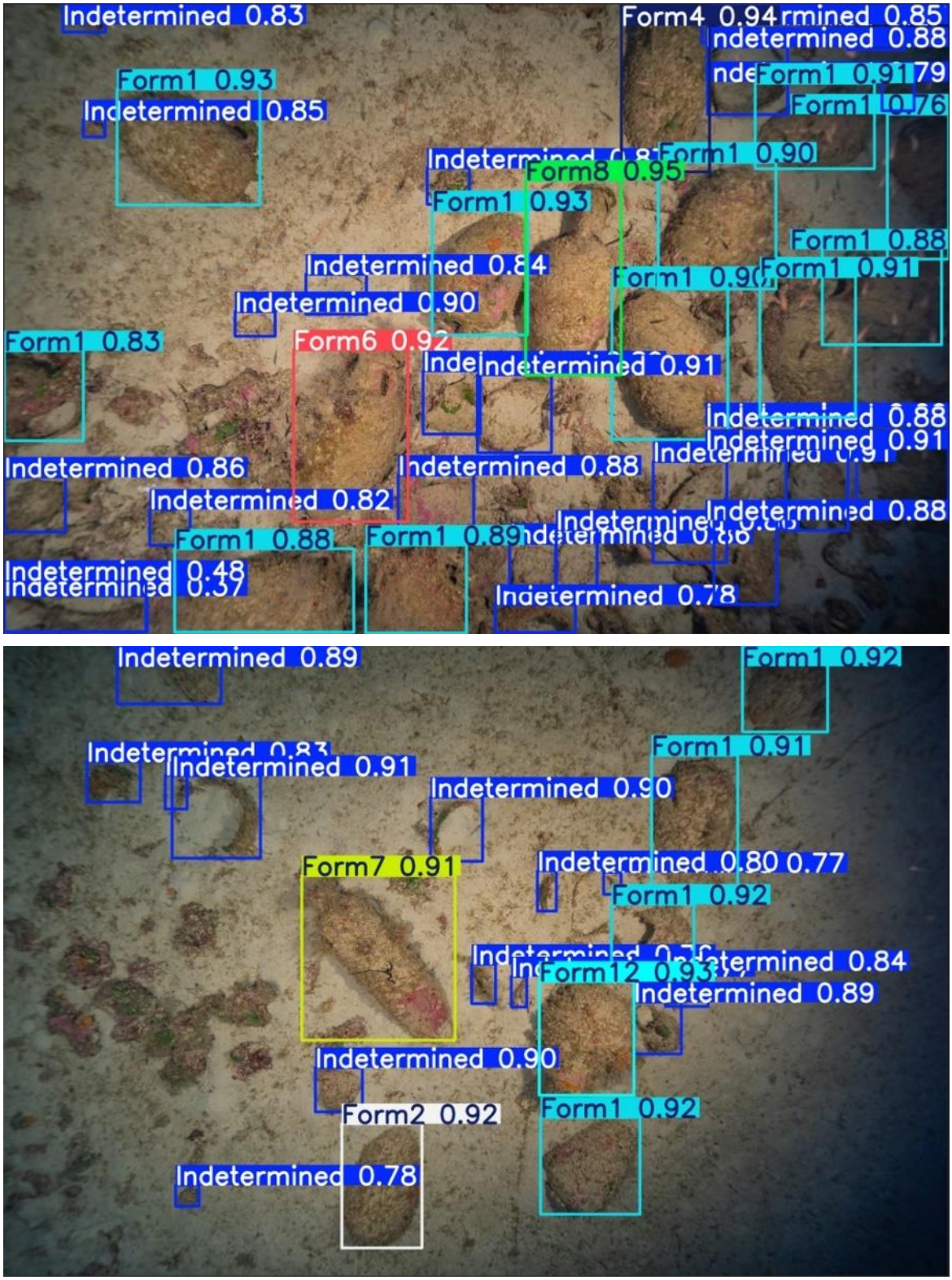


Figure 35. Examples of T1 model predictions on unseen data.

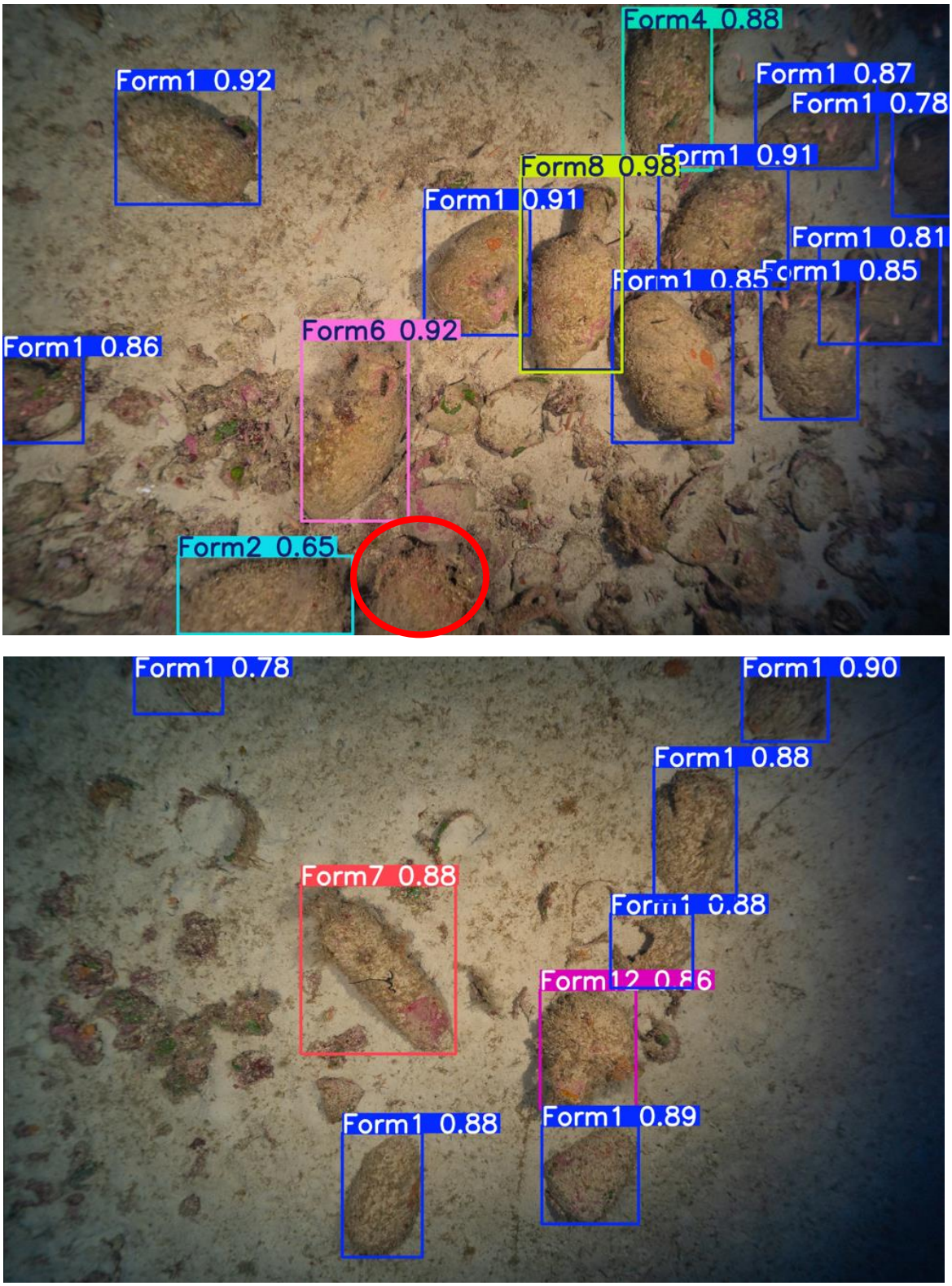


Figure 36. Examples of T2 model predictions on unseen data. Notice the absence of the 'indetermined' class, the only difference with T1 models and how it helps to reduce clout at the expense of some precision. Notice the increase in false negatives from T1 (red circle).

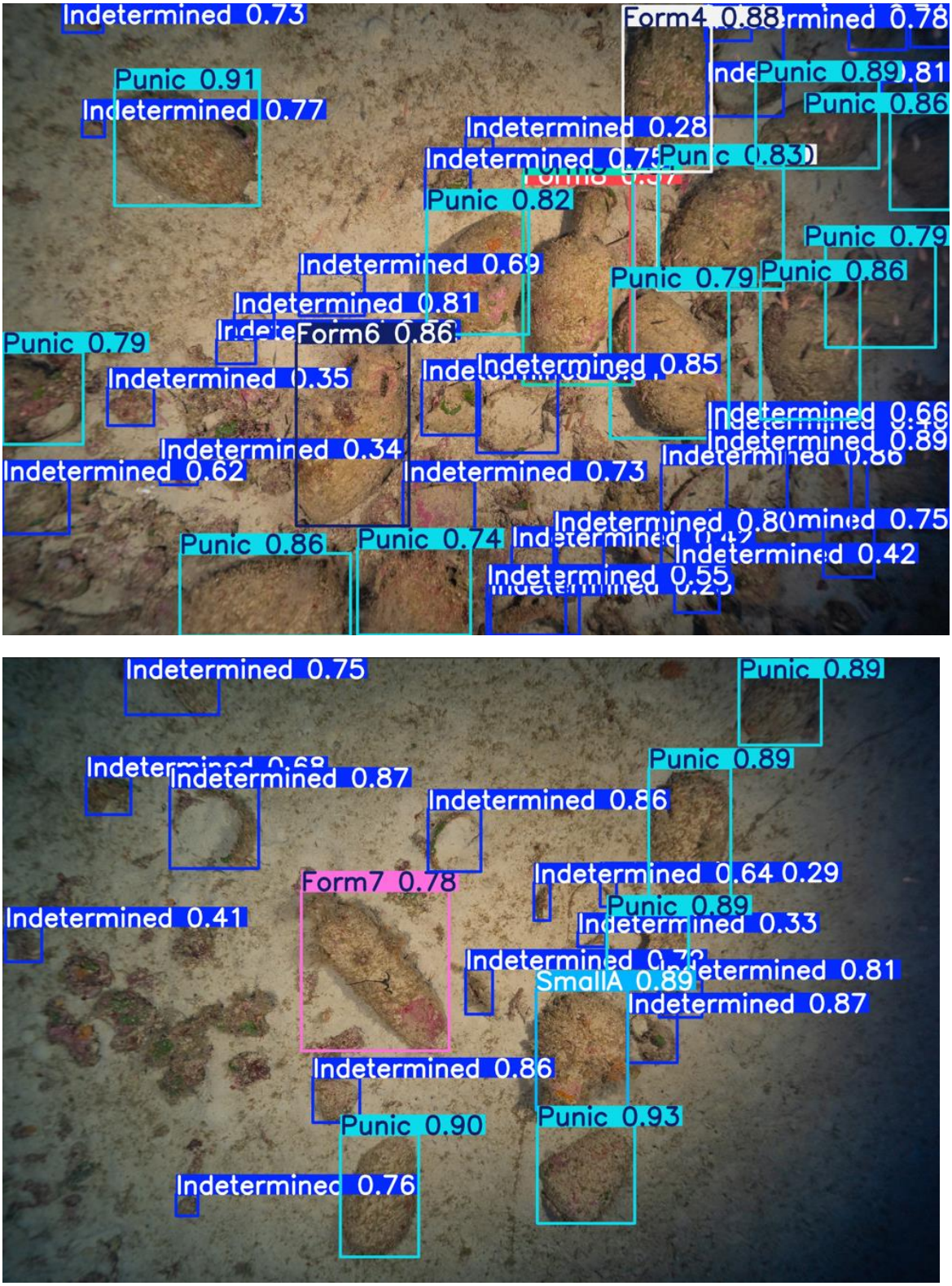


Figure 37. Examples of T3 model predictions on unseen data. Notice the high precision on Punic materials and how it differentiates them from the small amphorae.

Figure 38. Examples of T4 model predictions on unseen data. Notice the high precision of identification across the board.

6.2 Evaluation of Metrics Results

All 72 trained models developed in this dissertation underwent practical testing on a set of images from a photogrammetric dataset removed from the training area. The resulting metrics from each test are presented here through tables that highlight the best YOLO versions for each PAI.

Given the large volume of metrics and cross-referenced data generated during testing, displaying all of them in full would likely be more confusing than insightful. Therefore, while some of the analysis will be accompanied by examples of visual metrics for clarifying purposes, most of the metrics used in this section align with those provided by Paraskevas et al. (2023:4), the only other practical study on underwater assemblages using YOLO models.

In addition to the comprehensive review of all metrics provided on Appendix IV,[94] an additional summary tailored to the practical application of the metrics displayed on the result tables is provided here for quick reference:

-Model: This column specifies the size of the YOLO model used for training across two different versions. Different versions affect the data differently, as does the size.[95] In practical terms, for the different sizes used here (nano, small, and medium) the models are theoretically more accurate at the cost of significant speed. In the tables presented, the best-performing model for each PAI is highlighted in green, while the worst performer is highlighted in orange.[96]

-mAP50-Class (true number of objects): This metric represents the trained model's average precision across all classes, considering only objects detected with a confidence level of 50% or higher.

-mAP50:95: A stricter metric that evaluates model performance at progressively higher confidence thresholds.

-Average Recall: Metric that rates the average of objects missed by the model across all classes.

[94] Appendix IV (p.209).
[95] Size (p.62).
[96] In this comparison, the models are valued individually and rated based on the sum of their metrics for the purpose of achieving a more archaeologically valid model.

Metrics N1			
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)
			Anthropic (4578)
Yolo8n	0.462	0.679	0.766
Yolo8s	0.477	0.689	0.769
Yolo8m	0.488	0.663	0.777
Yolo11n	0.458	0.664	0.752
Yolo11s	0.467	0.659	0.755
Yolo11m	0.479	0.654	0.754

a)

Metrics N2			
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)
			Anthropic (1929)
Yolo8n	0.513	0.734	0.797
Yolo8s	0.547	0.72	0.813
Yolo8m	0.552	0.709	0.804
Yolo11n	0.534	0.745	0.81
Yolo11s	0.491	0.687	0.756
Yolo11m	0.521	0.709	0.783

b)

Metrics N3			
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)
			Anthropic (1528)
Yolo8n	0.571	0.763	0.845
Yolo8s	0.586	0.788	0.831
Yolo8m	0.574	0.726	0.817
Yolo11n	0.577	0.773	0.849
Yolo11s	0.558	0.769	0.819
Yolo11m	0.59	0.752	0.824

c)

Metrics N4					
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)		
			Avg.	Ceramic (4601)	Litter(3)
Yolo8n	0.641	0.769	0.871	0.747	0.995
Yolo8s	0.62	0.777	0.868	0.741	0.995
Yolo8m	0.595	0.794	0.873	0.752	0.995
Yolo11n	0.427	0.586	0.707	0.744	0.67
Yolo11s	0.643	0.792	0.878	0.761	0.995
Yolo11m	0.642	0.797	0.878	0.761	0.995

d)

Table 5. Testing metrics of nature models N1(a), N2(b), N3(c) and N4(d). Notice the highlight on the most (green) and least (orange) efficient architecture.

Metrics S1						
Model	mAP50-95	Avg.Recall	mAP50-Class(true n° of objects)			
			Avg. (4578)	Complete (528)	Buried (542)	Broken (3509)
Yolo8n	0.442	0.654	0.663	0.788	0.647	0.553
Yolo8s	0.464	0.696	0.679	0.808	0.575	0.655
Yolo8m	0.454	0.664	0.651	0.772	0.616	0.566
Yolo11n	0.457	0.668	0.681	0.802	0.67	0.571
Yolo11s	0.475	0.686	0.692	0.805	0.689	0.583
Yolo11m	0.475	0.67	0.681	0.805	0.654	0.582

a)

Metrics S2						
Model	mAP50-95	Avg.Recall	mAP50-Class(true n° of objects)			
			Avg. (1883)	Complete (500)	Buried (573)	Broken (810)
Yolo8n	0.413	0.641	0.623	0.8	0.723	0.449
Yolo8s	0.408	0.593	0.607	0.775	0.552	0.495
Yolo8m	0.408	0.584	0.595	0.775	0.57	0.438
Yolo11n	0.38	0.62	0.583	0.778	0.484	0.487
Yolo11s	0.401	0.616	0.606	0.738	0.549	0.532
Yolo11m	0.395	0.606	0.582	0.764	0.492	0.49

b)

Metrics S3						
Model	mAP50-95	Avg.Recall	mAP50-Class(true n° of objects)			
			Avg. (1528)	Complete (474)	Buried(663)	Broken(391)
Yolo8n	0.383	0.606	0.573	0.69	0.527	0.502
Yolo8s	0.398	0.564	0.565	0.694	0.522	0.479
Yolo8m	0.4	0.55	0.56	0.722	0.478	0.48
Yolo11n	0.387	0.59	0.578	0.704	0.541	0.489
Yolo11s	0.41	0.626	0.598	0.73	0.546	0.516
Yolo11m	0.354	0.566	0.526	0.663	0.49	0.424

c)

Metrics S4					
Model	mAP50-95	Avg.Recall	mAP50-Class(true n° of objects)		
			Avg. (4578)	Complete (527)	Buried-Inc. (4051)
Yolo8n	0.509	0.721	0.738	0.824	0.653
Yolo8s	0.502	0.694	0.725	0.791	0.654
Yolo8m	0.527	0.728	0.749	0.812	0.686
Yolo11n	0.423	0.634	0.655	0.785	0.525
Yolo11s	0.524	0.726	0.752	0.82	0.684
Yolo11m	0.518	0.729	0.753	0.823	0.682

d)

Table 6. Testing metrics of state models S1(a), S2(b), S3(c) and S4(d). Notice the highlight on the most (green) and least (orange) efficient architecture.

Metrics T1																			
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)																
			Avg	Indet.	Form1 (3636)	Form2 (65)	Form2b (24)	Form3 (12)	Form4 (9)	Form5 (41)	Form6 (2)	Form7 (6)	Form8 (21)	Form9 (26)	Form10 (3)	Form11 (62)	Form12 (13)	Form13 (11)	Form14 (2)
Yolo8n	0.221	0.317	0.282	0.531	0.656	0.144	0.15	0.643	0.0623	0.255	0.995	0.203	0.03	0.08	0.336	0.129	0.09	0.05	0.175
Yolo8s	0.31	0.309	0.383	0.571	0.699	0.264	0.164	0.723	0.14	0.267	0.995	0.343	0.011	0.222	0.336	0.189	0.253	0.379	0.566
Yolo8m	0.336	0.339	0.384	0.575	0.705	0.36	0.21	0.749	0.133	0.255	0.995	0.179	0.02	0.271	0.337	0.216	0.281	0.353	0.501
Yolo11n	0.259	0.303	0.323	0.563	0.716	0.177	0.153	0.272	0.124	0.202	0.995	0.369	0.0268	0.147	0.336	0.187	0.2	0.203	0.497
Yolo11s	0.287	0.346	0.356	0.552	0.693	0.227	0.172	0.559	0.099	0.222	0.995	0.198	0.0573	0.329	0.336	0.218	0.218	0.315	0.498
Yolo11m	0.293	0.325	0.348	0.582	0.692	0.294	0.186	0.393	0.076	0.163	0.995	0.323	0.0536	0.269	0.337	0.217	0.342	0.392	0.251

a)

Metrics T2																		
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)															
			Avg.	Form1 (657)	Form2 (65)	Form2b (24)	Form3 (12)	Form4 (9)	Form5 (41)	Form6 (2)	Form7 (6)	Form8 (21)	Form9 (26)	Form10 (3)	Form11 (62)	Form12 (13)	Form13 (11)	Form14 (2)
Yolo8n	0.227	0.282	0.286	0.712	0.251	0.176	0.473	0	0.137	0.995	0.174	0	0.109	0.333	0.189	0.111	0.179	0.5
Yolo8s	0.263	0.265	0.313	0.679	0.251	0.104	0.551	0	0.232	0.995	0.173	0	0.159	0.338	0.19	0.286	0.215	0.501
Yolo8m	0.258	0.262	0.313	0.714	0.261	0.0868	0.526	0.015	0.276	0.995	0.209	0.0127	0.107	0.338	0.173	0.238	0.232	0.513
Yolo11n	0.217	0.242	0.272	0.732	0.171	0.143	0.396	0.058	0.128	0.995	0.044	0.0401	0.0617	0.341	0.159	0.119	0.194	0.5
Yolo11s	0.255	0.291	0.309	0.723	0.237	0.0931	0.608	0.0587	0.225	0.995	0.175	0.0183	0.135	0.337	0.222	0.157	0.113	0.545
Yolo11m	0.254	0.241	0.297	0.703	0.221	0.0816	0.636	0.0162	0.212	0.995	0.249	0.039	0.0993	0.337	0.191	0.16	0.0142	0.5

b)

Metrics T3.																
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)													
			Avg. (4592)	Indet. (3636)	Punic (757)	Form3 (12)	Form4 (8)	Form5 (36)	Form6 (2)	Form7 (7)	Form8 (21)	Form9 (25)	Form10 (3)	SmallA (70)	Form13 (11)	Form14 (2)
Yolo8n	0.265	0.36	0.351	0.55	0.777	0.464	0.007	0.191	0.995	0.332	0	0.101	0.336	0.256	0.048	0.496
Yolo8s	0.278	0.305	0.359	0.557	0.791	0.528	0	0.225	0.995	0.164	0	0	0.333	0.201	0.217	0.497
Yolo8m	0.294	0.311	0.367	0.565	0.786	0.705	0	0.233	0.995	0.18	0	0	0.333	0.168	0.166	0.502
Yolo11n	0.241	0.364	0.326	0.527	0.789	0.492	0.149	0.142	0.995	0.174	0	0.1	0.18	0.15	0	0.496
Yolo11s	0.269	0.355	0.34	0.541	0.772	0.364	0	0.147	0.995	0.17	0	0	0.336	0.143	0.27	0.517
Yolo11m	0.296	0.376	0.374	0.594	0.801	0.57	0	0.254	0.995	0.182	0	0.146	0.337	0.2	0.217	0.497

c)

Metrics T4.								
Model	mAP50:95	Avg. Recall	mAP50-Class(true n° of objects)					
			Avg. (4591)	Indet. (3637)	Punic (774)	Italic (15)	Local (163)	nAfric. (2)
Yolo8n	0.416	0.538	0.57	0.559	0.809	0.159	0.326	0.995
Yolo8s	0.431	0.546	0.579	0.581	0.806	0.139	0.375	0.995
Yolo8m	0.457	0.49	0.587	0.565	0.785	0.151	0.439	0.995
Yolo11n	0.384	0.518	0.541	0.561	0.78	0.06	0.303	0.995
Yolo11s	0.431	0.552	0.587	0.592	0.813	0.146	0.391	0.995
Yolo11m	0.478	0.611	0.611	0.581	0.84	0.213	0.428	0.995

d)

Table 7. Testing metrics of state models T1(a), T2(b), T3(c) and T4(d). Notice the highlight on the most (green) and least (orange) efficient architecture.

6.2.1 How models compare at PCI level

The PCI was introduced as a factor to differentiate trained models based on the level of complexity and the amount of archaeological information incorporated into their training. According to this framework, it was expected that as model PCI changed from nature to state and to typological, model performance would decline. For this reason, the comparison at this level can be done only by looking at the baseline models from each PAI where baseline models are defined as the most obvious implementation of the PCI characteristics through a PAI that also considers **all** objects.

Technically, the first application of a detection model to underwater assemblages was presented by Paraskevas et al. (2023). Under our classification, their model would qualify as an **N1** model in terms of PAI. In fact, our N1 was designed both to establish a baseline for nature models and to replicate and be compared to Parasevkas et al's experiments. In this regard, Paraskevas et al. achieved a maximum **mAP50 of 0.755 (75.5% mean accuracy) and 0.6 AR (average recall)** using a YOLOv8 small frame (Table 8).[97] Our **N1** models (Table 5a) demonstrated superior performance when deployed on a YOLOv8 medium frame, achieving **76.9% mAP50 and AR of 0.689.** Because this nature PAI considers all material present at the site, it serves as a general-purpose benchmark for evaluating the performance of nature models.

Model	mAP50	AR	Speed (ms)
Nano	0.724	0.587	3.6
Small	**0.755**	**0.6**	**6.5**
Medium	0.727	0.53	14.7
Large	0.747	0.53	22.5

Table 8. Metrics table from Paraskevas et al. experiment (Paraskevas et al., 2023:4).

The next PCI group in this comparison is the one constituted by state models (Table 6).[98] Of these, **S1** models form the baseline because, like **N1**, they consider all archaeological materials. Regarding their metrics, **S1** models achieved **a mAP50 of 69.2% and an AR of 0.686**, which is a

[97] p.88 for quick reference of mAP50 and AR (Appendix IV for full breakdown, p. 209).
[98] This project marks the first instance in maritime archaeology where a detection model has been trained to automatically assess a complex concept like the state of preservation, with no direct parallel for comparison.

sizeable decrease compared to the **76.9% mAP50 and 0.689 AR** shown by **N1** models. This decline reflects the expected impact of increasing PCI, as transitioning from nature to state models introduces greater complexity.

The third group in this comparison are the typological models (Table 7). Among them, **T1** models are the most complex, as they classify all ceramic types at the site across multiple categories. They serve as the baseline PAI for typological PCI because they incorporate the entire ceramic catalogue and leave no material unaccounted for. In terms of performance, their metrics dropped significantly compared to the baseline models in **N1** and **S1**, displaying a much lower **mAP50 of 38.4% and an AR of 0.339**. As per the visual evaluation, while most of the typological models are not yet suitable for real-world implementation, the results from the metrics of their baseline model align with the expected trend: An increase in PCI leads to a decline in performance.

Analysis: While typological models largely fail to be usable based on both their metrics and visual evaluation, they still help by establishing the boundaries of future work. The low performance of models incorporating full typologies suggests that, for now, they are impractical for real-world use without further adjustments. This is not the case for nature and state models, which achieve strong results, outperforming even the field-proven models presented in Table 8 despite this dissertation never aiming to create fully deployable models. This suggests that a dedicated project focused on refining them could achieve even higher performance.[99]

The results presented here support the central hypothesis of this comparison: as PCI increases, efficiency and precision decrease. The trend is evident in the performance of different model types from the baseline nature models (**76.9% mAP50**) to the state models (**68.2% mAP50**), and finally to the typological models (**38.4% mAP50)**. Since this decline is directly linked to the PCI, the metrics shown allow us to have an idea and perhaps define the limits within which future PCI groups can be developed. In archaeological terms, this establishes both the baseline (nature models) and the upper complexity threshold (typological models) for what detection models can realistically achieve with our versions and sizes of YOLO.

Beyond these performance insights, categorizing models based on how they consider archaeological materials—whether by their nature, state of preservation, or typology—opens the

[99] For example, by expanding the dataset and using large model architectures capable of processing our 2D data at full scale.

door for new research directions. Some examples of possible ideas for other PCI groups include models that focus on the marine biota present on archaeological sites, their ratio of growth, or the ratio of degradation of more modern wrecks.

More importantly, the (mostly) successful results first seen in this comparison highlight the significant advantages computer vision holds over traditional archaeological methods that rely on paper maps, at least when it comes to the possible research questions that our test detection models can address. Detection models provide real-time analysis and a level of data utility far beyond manual approaches thanks to their exportability, the digitalization of information like number of objects by class, and their nature as visual tools. In addition, detection data can be combined with other computer vision techniques for deeper insights—for example, projecting 2D detection results onto a 3D photogrammetric model to enable object classification and quantification (Zammit et al., 2024).

Also related to this topic is the major advantage of detection models against traditional models in terms of efficiency and reusability. While traditional site maps could still be used to study patterns of deposition based on material density, the sheer effort required to manually map all the objects at a site like Xlendi Archaeological Park makes this approach unrealistic. In contrast, during this dissertation, creating a nature model required less than 10 hours of labeling work. Moreover, once an initial model is labeled and trained, because labelling data can be recycled, subsequent models with similar PAI require significantly less effort. For instance, while creating the **N1** nature models took under 10 hours of labeling and from two to eight hours of training for each of the six versions of **N1**, the labeling time for **N2** models was reduced to under five hours, and that of **N3** and **N4** models less than two hours each.[100] The same principle applies to state and typological models. For typological models, rather than re-labeling everything for **T2**, **T3**, and **T4**, the data recorded during the **T1** models' creation allowed us to modify and reorganize classes with just a few lines of code. This heavily reduced labeling time for each successive typological model.

[100] This enormous time reduction was achieved by a combination of two factors. First, reusing the location of the boxes from previous labelling processes means we only must add or delete boxes based on the differences between PAI, as well classifying the entire set. The result of this is we have a significantly smaller number of work to do in terms of "clicking" and forcing our eyes. The second factor is our ability to generate computer code to rename and merge classes at will by manipulating the label files. For instance, if we have a state model that has three different classes (complete, buried and broken) and we want a different state model that wants to merge the last two and display only 'complete' and 'incomplete' elements, we can write some computer code to effectively merge the last two. With the help of generative AI, no specific coding knowledge is required for this step.

According to this overview, it is fair to say there is not a situation where the concept of PAI can be implemented in traditional methods. Doing so by manual identification and mapping would require the same exhaustive effort for each PAI, likely amounting to hundreds of hours of work that, on top of everything, would result in data that is less flexible, less useful, and not exportable. Ultimately, this comparison of models based on their PCI underscores the transformative potential of detection models in maritime archaeology, offering a scalable, efficient, and far more powerful alternative to conventional techniques.

6.2.2 How models compare at PAI level

Each PAI within the different PCI groups is designed to adjust its identification and classification parameters (excluding location) to serve a specific archaeological purpose or scenario. As such, any attempt to directly compare their metrics must be approached with a certain degree of relativism. This review has been conducted with this consideration in mind. For each PCI, we will first review the purpose behind the design of each of its PAI, its best performer, and some examples of how they might be used before comparing their metrics.

6.2.2.1 Nature models (N)

These include **N1**, **N2**, **N3**, and **N4** models. Classifying objects by nature, all of them are valid according to the visual evaluation.

-N1 models (Table 5a): Building on the experiments by Paraskevas et al. (2023), these models were designed as the baseline for archaeological detection, incorporating all archaeological materials present in the dataset. The best-performing one achieved a **mAP50 of 76.9% and AR of 0.689**, which in this context along with the visual evaluation means they work very well.

Their primary advantage lies in their speed and real-time usability. They can be integrated into camera feeds to assist in underwater site surveying. Additionally, when applied to an orthomosaic,[101] they provide both numerical data (NISP)[102] and visual insights into the site's density of materials.

[101] Orthomosaic (p.15).
[102] Number of Identified Specimens (p.53).

-N2 models (Table 5b): The best-performing **N2** model achieved a **mAP50 of 81% and an AR of 0.745**, which are improved numbers over **N1** models. **N2** models in general were developed from the PCI baseline framework but tailored for scenarios where researchers want to focus only on the most relevant archaeological materials rather than the entire assemblage. Also, they were designed to assess whether this selective approach impacts performance compared to **N1** models.

In practice, **N2** models provide similar interpretative capabilities and speed as **N1** while reducing the visual clutter caused by scattered ceramic sherds.

-N3 models (Table 5c): Following the success of **N2** models, **N3** models were designed to take the same concept one step further by ignoring all non-representative ceramic elements, thus fitting a scenario where the researcher wants to consider the density of **individual** vessels. The best-performing **N3** model achieved a **mAP50 of 84.9% and an AR of 0.773**, which is also a step above their predecessors.

When applied to an orthomosaic, **N3** models provide both numerical data (MNV)[103] and visual insights on the density of individual vessels on archaeological sites. This could be applied to a situation where a researcher is trying to count the number of individual items over an area, for example.

-N4 models (Table 5d): N4 models were designed to evaluate how the inclusion of non-ceramic elements impacts overall model performance. The best-performing **N4** model achieved a **mAP50 of 87.8% and an AR of 0.797.**

While this PAI is not particularly useful for Xlendi Archaeological Park—where non-ceramic elements are limited to a few plastic bottles and objects left by divers during photogrammetric surveys (e.g., scooters, measuring tools)—it has broader applications in other, shallower and therefore more exposed archaeological contexts. In these cases, for instance, **N4** models could be highly valuable for heritage management, including site monitoring, preservation efforts, and public engagement initiatives regarding the necessity for the preservation of underwater heritage.

Analysis: The first noticeable trend when comparing these models is the steady improvement in performance as out-of-context ceramic sherds are eliminated from the group of identifiable materials. From **N1** (**0.769 mAP50, 0.689 AR**) to **N2** (**0.81 mAP50, 0.745 AR**) and to **N3** (**0.849 mAP50, 0.773 AR**), nature models exhibit a total progressive increase of 10.4% in accuracy. This

[103] Minimum number of vessels (p.53).

progression follows the shift from counting the test area's NISP with **N1** models to estimating the area's MNV with **N3** models. Visual metrics, such as confusion matrices for all three groups, suggest that this improvement is primarily due to a reduction in background interference, particularly from smaller sherds (Figure 39).

N4 models were designed to assess whether introducing non-archaeological elements such as litter or scuba gear would affect performance. The results indicate that these additional elements did not significantly impact model accuracy, which was expected given their stark contrast they present with archaeological materials. While 'ceramic' materials were detected with **76%** accuracy in this model, the inclusion of an extra class with a 100% precision rate (but very low frequency) slightly skewed the mean accuracy, making it less representative of real-world performance.

Beyond the models developed in this study, future nature model designs (PAI) could incorporate strategic elements of the natural background such as reefs as distinct classes. Additionally, models could be expanded to identify a wider range of archaeological materials, including grinding stones, wooden elements, cannons, cannonballs, and anchors. These expanded capabilities, combined with the variations explored in this study, demonstrate the potential of nature models in both archaeological research and heritage management. Even in their current state, they can be effectively used for surveying, mission planning, analyzing material density, categorizing assemblages, site monitoring, preservation efforts, and public outreach initiatives.

Example of use: As an example of how nature PAI can be used to effectively apply and interpret a nature model, let's consider a scenario in which we, as archaeologists, aim to map the location of every individual vessel of an underwater site. Our goal is to analyze depositional patterns by assessing the density of vessels and visually plotting the site's MNV to help isolate potential shipwrecks within a widely scattered assemblage of materials.

To achieve this, we must carefully choose the appropriate trained model. State and typological models provide more detailed classifications, but since this data is not essential for our specific goal, using these models would unnecessarily reduce performance without adding relevant insights. Following this train of thought, while a nature model quickly becomes the best fit for our needs, we must further refine our selection. Since all available nature models identify archaeological materials and exclude non-archaeological elements, examining their PAI allows us to determine the most suitable option. This assessment will lead us to see that **N1** and **N4** models

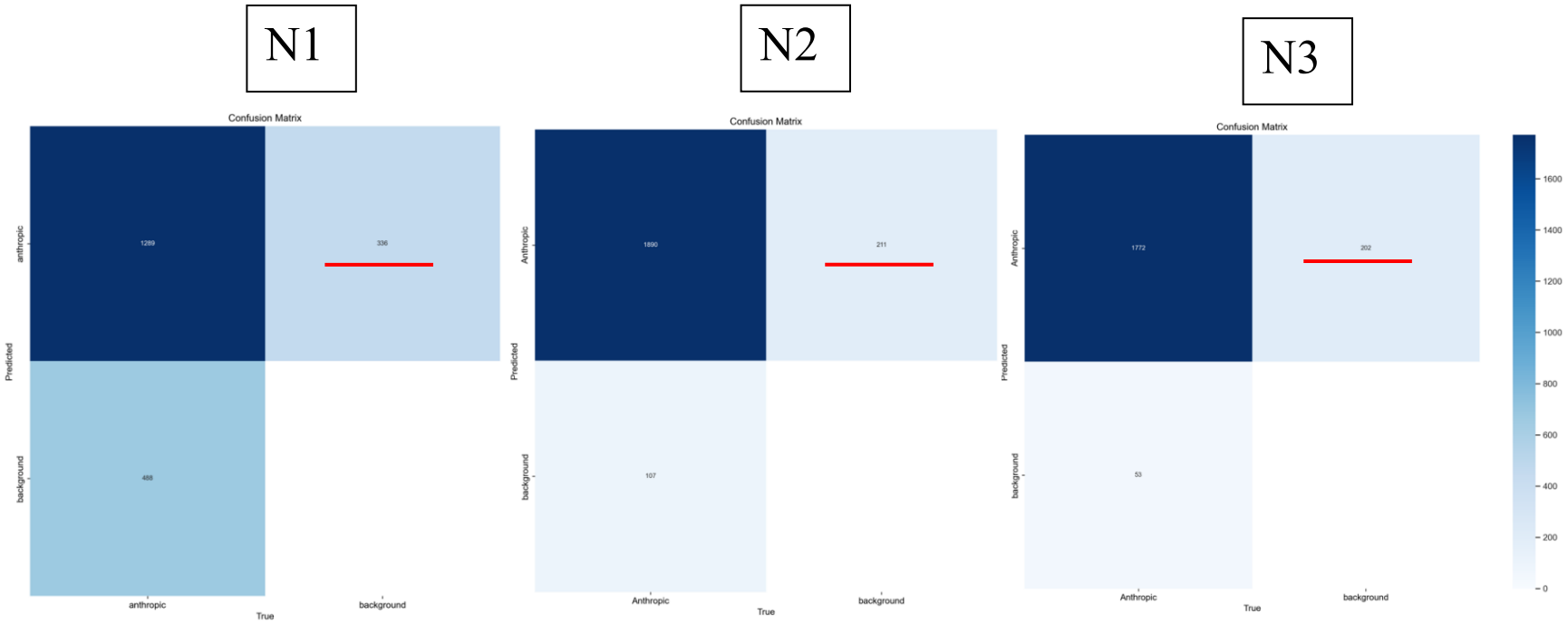


Figure 39. From left to right, confusion matrixes of N1, N2 and N3. These matrixes help understand how well a model is predicting different classes and why. This is done by comparing the model's predictions (X-axis) to the ground truth labels (Y-axis). In this case, if we look at the numbers of the square corresponding to the model's predictions on anthropic elements against the background (non-labelled) elements of the ground truth (top right square), we notice a progressive decrease from N1 to N3. This means that N2 is less prone than N1 to misidentify background elements like rocks as anthropic materials. In the same way, N3 is less prone to misidentify in this way than N2.

consider all anthropic elements and not only individual vessels. **N2** models work better for us, as they do not consider 'broken' ceramic sherds. However, in this case, **N3** models are the most suitable choice, as they focus only on elements truly representative of a complete vessel (Table 5c, Figure 29).

Therefore, by selecting **N3** in this example, we ensure that our model provides precise and relevant data, allowing us to answer our archaeological questions efficiently while maintaining optimal performance.

In summary, nature models improve as training parameters are refined to exclude out-of-context fragments and to focus only on representative elements. Additionally, non-ceramic elements can be incorporated without negatively affecting detection performance for archaeological materials. Most importantly, this comparison highlights the responsiveness of detection models to the strategic manipulation of the labeling process through archaeological insight, showcasing their vast potential for future applications.

6.2.2.2 State models (S)

These are **S1**, **S2**, **S3** and **S4**. Classifying ceramic materials based on their state of preservation, state models are all valid according to the visual evaluation, demonstrating precision levels sufficient for real-world applications.

-S1 models (Table 6a): These models were designed as a baseline for state models and to assess how a detection model performs when trained with a strong layer of subjective archaeological information. While additional classes could have been included to represent different states of preservation, three were deemed sufficient to produce meaningful results. The best-performing **S1** model achieved a **mAP50 of 69.2% and an AR of 0.686**, which aligns with the positive results of the visual evaluation.

From an archaeological perspective, **S1** models have broad applications that are similar to those of **N1** models, but with the benefit of being particularly valuable for understanding site morphology and site formation processes. When plotted on an orthomosaic, the output of trained **S1** models allows researchers to identify deposition patterns more easily and to infer their causes based on the state of preservation of the materials. For example, clusters of 'complete' amphorae

could indicate the presence of shipwrecks, while objects arranged by size around natural elements—resembling the effect of a sieve—suggest natural post-depositional processes. Similarly, variations in the density of 'buried' materials across different areas might signal the presence of additional, still-hidden archaeological materials, aiding in mission planning in a different way.

-S2 models (Table 6b): These models were developed based on the **S1** framework but modified for scenarios where researchers apply slightly different definition of what constitutes a 'complete' or a 'buried' amphora. In this version, an amphora must retain key elements such as handles, a rim, or a base to be considered 'complete'. Like **N2** models, **S2** models are designed to exclude small pottery sherds, making them suitable for studies focused on more archaeologically significant materials. By adopting this selective approach, **S2** models aim to assess whether filtering out minor fragments enhances performance in the same way it did for nature models.

The best-performing **S2** model achieved a **mAP50 of 62.3% and an AR of 0.641**, performing worse than state models that do not exclude any elements. However, **S2** models still provide similar interpretative capabilities to **S1** while reducing the visual and computational clutter caused by smaller broken fragments.

-S3 models (Table 6c): Following the performance decline observed from **S1** to **S2** models, **S3** models were designed to verify if this trend holds, and to explore the theoretical principles behind their PAI but with an added layer of complexity. The best-performing **S3** model achieved a **mAP50 of 59.8% and an AR of 0.626**, which maintains the declining trend in precision on state models.

In practical terms, **S3** models are identical to **N3** models in that they exclude non-representative ceramic elements to focus on the material density of individual vessels and the site's MNV. However, they also introduce an additional layer of subjective information regarding the state of preservation, and so their outcome trained models form a hybrid between nature and state models. This situation highlights how, depending on the specific archaeological question, it may be more efficient to use either a nature model like **N3** or one of the comparatively underperforming, more informative state models like **S2** or **S3** instead.

Nevertheless, **S3** models' hybrid nature could be particularly useful in scenarios where both the site's MNV and the state of preservation of each piece need to be quantified and visualized. With this model's classification system, researchers can determine the exact number of 'complete', 'buried', and 'broken' individual vessels within an area.

-S4 models (Table 6d): Given the declining performance observed in **S2** and **S3** models compared to the baseline **S1** model, **S4** models were designed to improve the performance of state models by simplifying the training to two classes ('complete' and 'incomplete-buried'). The best-performing **S4** model achieved a **mAP50 of 75.2% and an AR of 0.726**, which suggests that we succeeded in our goal of improving the performance of **S1** models while maintaining state of preservation considerations.

S4 models are ideal for scenarios focusing solely on 'complete' vessels. For example, they would be useful in studies aiming to identify clusters of complete amphorae that could indicate the presence of shipwrecks over large areas such as Xlendi Archaeological Park.

Analysis: Starting with the baseline **S1** models, a study of how different states of preservation are detected reveals that 'complete' materials, with an identification rate of **80%**, are the easiest in terms of feature extraction.[104] However, this precision is overshadowed in the general context by the metrics shown for 'buried' **(68.9%)** and 'broken' **(58.3%)** materials. In this regard, a precision-confidence graph for **S1** models (Figure 40) helps illustrate the discrepancies between these precision rates and model confidence, offering further insight into the factors ballasting the model's efficiency. With this in mind, and upon reviewing the confusion matrix and other visual metrics from **S1**, it becomes apparent that the model's ability to detect 'broken' materials is hindered by background elements such as reefs and vegetation (Figure 41).

We have established that **S2** and **S3** models were designed to enhance the performance of **S1** models by refining the interpretation of preservation states (through the PAI at class level). In this regard, the primary adjustment made to **S2** tried to accentuate the differences between 'complete' and 'buried' materials. Additionally, **S2** models ignored the smallest pieces of 'broken' materials, considering them inconsequential to the current archaeological context/question (Table 6b). **S3** models further adjusted their parameters with relatively slight differences for 'complete' and 'buried' materials but also enforced a stricter definition of 'broken' by only acknowledging representative vessel fragments. During design, this approach logically followed the success from nature models, where disregarding situationally unimportant materials led to improved performance.

[104] Feature extraction (p.159).

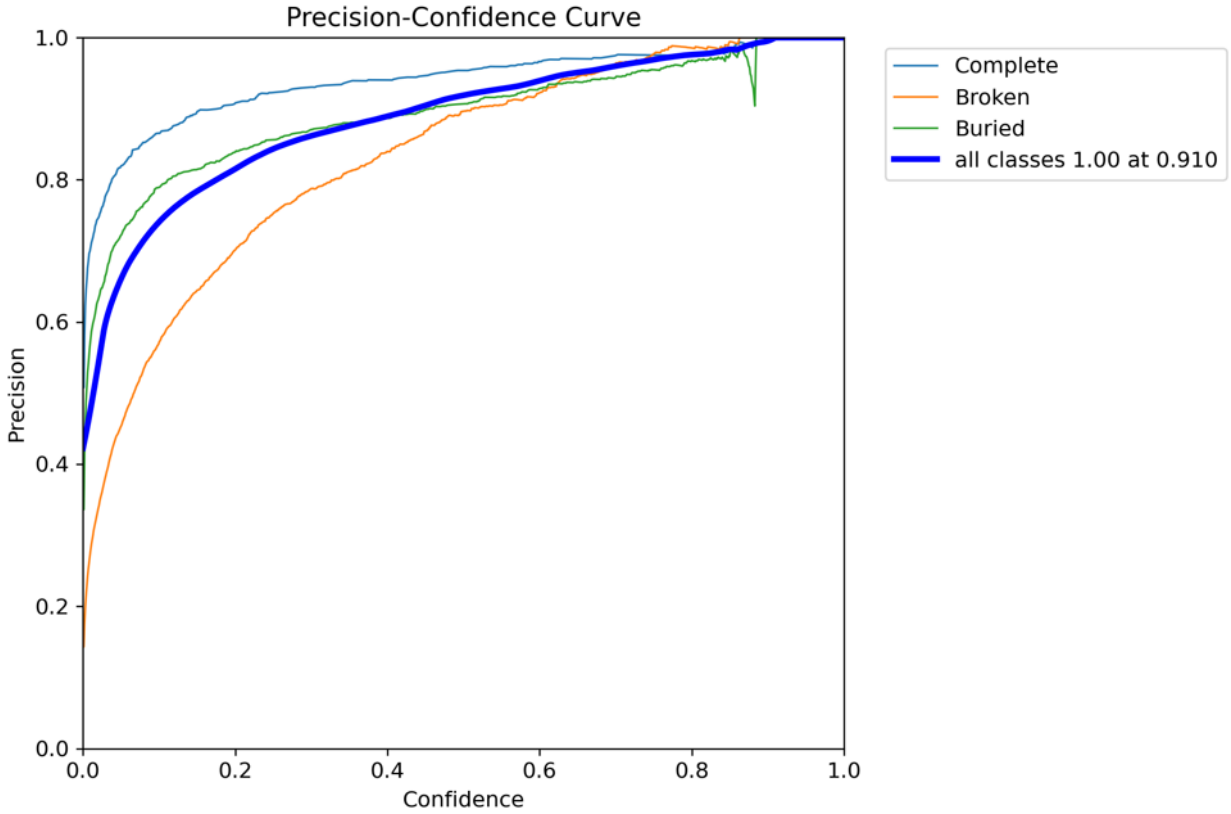


Figure 40. Precision over detection confidence curve for every class of S1 models.

Figure 41. From left to right, confusion matrixes of S1, S2 and S3. These matrixes help understand how well a model is predicting different classes and why. This is done by comparing the model's predictions (X-axis) to the ground truth labels (Y-axis). In this case, if we look at the evolution of the squares corresponding to the model's predictions on 'buried' and 'broken' elements (determined by color intensity on the right columns of each model) we notice how the main source of misclassification with the background gradually switch from being mainly on 'broken' materials in S1 (red circle) to 'buried' materials in S3 (green circle)

However, the results indicate that this strategy did not yield the same benefits for state models as it did for nature models. Performance metrics for **S2** (**0.623 mAP50, 0.641 AR**) and **S3** (**0.598 mAP50, 0.626 AR**) decreased compared to **S1**, with precision dropping not only in general, but also across the models' ability to correctly predict the rest of the classes:

- 'Complete': **S1 (80.5%), S2 (80%), S3 (73%).**
- 'Buried': **S1 (68.9%), S2 (72.3%), S3 (54.6%).**
- 'Broken': **S1 (58.3%), S2 (44.9%), S3 (51.6%).**

This experiment leads to two conclusions:

First, unlike nature models, reducing the number of ceramic sherds considered does not decrease model confusion in state models. Instead, it introduces additional misclassification, particularly between 'buried' materials and background elements. It's possible to observe this trend when analyzing visual metrics such as the normalized confusion matrices for the three PAI approaches (Figure 41). In this case, our interpretation of the situation is that while progressively removing out-of-context sherds reduces background interference, it also amplifies misclassification of 'buried' elements, as the model now faces an increased number of unidentified sherds in the background. Ultimately, strategically excluding certain 'broken' materials not only failed to improve detection accuracy but also decreased the model's overall confidence when identifying all other materials.

The second key conclusion is that among the three models designed for a three-class state of preservation classification, **S1** models emerge as the most effective for assessing preservation state—at least within the Xlendi Archaeological Park dataset. The results suggest that the initial premise behind the **S2** and **S3** experiments was flawed. Further experiments exploring alternative adjustments to the PAI at both general and class levels could potentially produce more effective models.

The success of the **S4** models confirms this point. Developed as an alternative to enhance performance, **S4** achieved the highest efficiency among all state models, with a **mAP50 of 75.2% and an AR of 0.726.** This improvement was made possible by merging the 'broken' and 'buried' categories before training, allowing the model to focus on distinguishing fully preserved artifacts from the rest of the assemblage. The benefits of this approach become evident when comparing the performance of different PAI at correctly identifying 'complete' materials (see Table 6). While

S1, **S2**, and **S3** models show **mAP50 values of 78.8%, 80%, and 73%,** respectively, for 'complete' items, **S4** models improve upon these results, achieving an **82%** accuracy rate.

Interestingly, the comparison of state models reveals that, contrary to what nature model metrics suggested, improving model performance through PAI manipulation is neither straightforward nor guaranteed. Instead, it highlights that correct model design requires a nuanced understanding of how detection models identify, classify, and differentiate objects. Fortunately, increasing the PCI from nature to state models also provides many more options for adjusting PAI to further refine results like, for instance, redefining state of preservation categories or modifying the number of classes to produce cleaner and more efficient models.

Also, regarding the possibilities of other state PAI, a few examples that come to mind are dedicated models designed to assess the preservation state of different material types, models designed to include a wider variety of ceramic elements like rims and handles, or perhaps models designed to analyze degradation patterns on modern shipwrecks.

All in all, integrating state of preservation data into nature models significantly expands their capabilities while slightly decreasing their precision. Using similar tools, state models can help reconstruct site formation and post-depositional processes in underwater assemblages. They also improve our ability to interpret these assemblages, aiding in the identification of shipwrecks and monitoring their preservation—particularly for modern wrecks. Furthermore, state models enhance underwater mission efficiency by generating archaeological priority maps and identifying optimal areas for material recovery.

Example of use: As an example of how state PAI can be applied in a practical archaeological scenario, let's consider a potential future study aimed at unraveling the site formation process of Xlendi Archaeological Park.

In this scenario, if we had to choose model design for this purpose, which one would it be? Without careful consideration, we might be tempted to collaborate with computer vision experts to create a model based on typological data or perhaps opt for the simplest approach by designing a nature model. However, upon deeper reflection, we realize that having our materials classified by typologies does not only not answer to preservation information in a useful way for this question, but also hinders us by virtue of typological models being inherently inaccurate, as even if we succeed in creating working typological models, they are still going to be less efficient than

the others. Nature models that plot all materials as a single category, on the other hand, would obscure depositional patterns based on preservation, which is the kind of information we would need to assess site formation processes. Hence, given our research goals, state models are the obvious choice.

Once we have decided on a model with a state PCI, we can refine our selection by examining their PAI (Table 2). For this scenario, **S2** and **S3** are of no use to use as they have been trained to ignore certain broken elements, which would skew our interpretation as we would not have the real number of broken elements. **S4**, on the other hand, makes no distinction between 'buried' and 'broken' elements, which is also detrimental to our goal. In this case, **S1** models (Figure 31, Table 6a) emerge as the best option, as they classify every instance of archaeological material based on its state of preservation, including the smallest 'broken' pieces. This ensures that we can accurately assess the spatial relationships between 'complete', 'buried', and 'broken' vessels, ultimately allowing us to reconstruct depositional patterns with greater accuracy.

In summary, while state models require a more nuanced approach to PAI design than nature models, they respond equally well to strategic training adjustments and to their use in real-time. Their broader range of PAI options offers significant potential, though refining them for archaeological use demands both an understanding of detection models and specialized archaeological knowledge. Nevertheless, the fact that they function effectively even in the test scenario proposed in this dissertation—and that they embed valuable archaeological insights—demonstrates that their significant advantages are already accessible to researchers today.

6.2.2.3 Typological models (T)

T1, **T2**, **T3**, and **T4** models were designed to identify and classify archaeological materials based on the typological information from Appendix III. After the visual evaluation, although it was evident that all models within this PCI group function as intended,[105] only **T4** models could be considered somewhat valid. However, given that the models from all four PAI work as intended and their limitations seem to stem from factors beyond their theoretical design (which may be

[105]The algorithms work as intended because they try to classify the ceramics following our exact design—implementing the typologies correctly. They are not functional or valid because they don't do it well enough.

addressed with time), we will still explore hypothetical scenarios regarding their potential applications.

Notably, the development of these models did not focus on adapting them to different archaeological scenarios like nature and state models. Instead, the progression from **T1** to **T4** reflects an effort to refine their performance with the goal of creating a reliable and accurate typological classification model. Some elements of the comparison are therefore included in the explanation of each model's result.

-T1 models (Table 7a): T1 models were trained in the most straightforward manner to establish a performance baseline for typological models. All the ceramic types from the catalogue were implemented as 15 classes along with an extra 'Indetermined' class for unidentifiable materials.

While they work as intended, the best-performing **T1** model achieved only a **mAP50 of 38.4% and an AR of 0.339**. The model seems to be more proficient at identifying and classifying the following classes:

- 'Indetermined' at **57.5%.**
- 'Form1' (Ramon 3.2.1.2) at **70.5%.**
- 'Form3' (Ramon 7.1.1.1) at **74.9%.**
- 'Form6' (Maña C) at **99%.**

Not coincidentally, as we can see in Table 7a, these classes are either some of the most common on the site in terms of numbers (giving the model more chances to learn and create more robust predictions) or have very particular shapes and therefore very recognizable physical features (and thus are more easily differentiated). The rest of the materials show very low precision ratios.

-T2 models (Table 7b): Building on the failure of **T1** models, **T2** models were developed to test whether training the model to focus only on objects with recognizable typology would improve overall performance. However, as shown by the performance metrics of the best-performing model (**mAP50 31.3% and AR 0.265**), the effect was quite the opposite. This drop in performance is not solely due to the removal of an 'Indetermined' class with a relatively high average precision—thus lowering the overall mean precision—but also to a general decline in the metric across almost all remaining classes. This suggests that excluding materials with no identifiable typology introduces additional confusion into the model, increasing both false positives and false negatives.

-T3 models (Table 7c): After unsuccessful attempts to improve typological model performance with the existing tools, **T3** models were developed to enhance the results through different means.

This was done in two ways: first, by reintroducing the 'Indeterminate' class, and second, by strategically merging typologies with similar archaeological significance before training. A detailed breakdown of these mergers can be found in Table 4a:

- 'Form1', 'Form2', and 'Form2b' were merged and renamed as 'Punic' since they represent variations of the same Punic shapes and hold similar significance within the assemblage.
- 'Form11' and 'Form12' were merged into a new class named 'SmallA' due to their shared characteristics as small, unidentified amphorae of local origin and similar features.

Ultimately, this approach proved unsuccessful as well. The best-performing **T3** model performed worse than the **T1** models with a **mAP50 of 37.4% and an AR of 0.376**. A breakdown of the different typologies displayed in Table 7 reveals the reason for this: while the newly created 'Punic' class achieved a higher average precision ratio (**80.1%**) than its individual components (Forms 1,2 and 2b) did in the **T1** and **T2** models, the slight reduction in typological options did not lead the models' parameters getting less confused with the other classes. In the end, overall performance decline is explainable by the fact that 'Punic' elements were already among the most precisely recognized typologies and, by merging them, we effectively removed their contribution from the mean average, which led to lower overall metrics.

-T4 models (Table 7d): **T4** models represent the final attempt at developing a functional typological model. They implement a more effective approach to the improvement of performance through adjustments on their PAI that are an evolution of those introduced for **T3** models. Rather than grouping classes based on typologies, the PAI of these models groups the classes at class level by geographical origin.[106] In total, these models have five classes that group all typologies at the site: 'Indetermined', 'Punic', 'Italic', 'Local' and 'NAfrican'. By shifting the focus from direct typological classification to vessel production origins, this approach capitalizes on the distinct physical characteristics of amphorae from different regions,[107] improving the model's ability to

[106] Geographically induced dysmorphism (p.58).

[107] The labels in this group are the biggest example of the idiosyncratic nature of archaeological model design. Physically (the only evaluable factor here), the geographical origin of amphorae can often be inferred by aspects of shape and form like size, overall shape, size, the specific shape of individual features like handles, neck or base…etc. These models (T4) are designed to use this very specific (typological) information to classify amphorae by geographical origin. For example, Punic amphorae on this site are often squat, egg-shaped, and with short or nonexistent necks, while North African ones are characteristically long, narrow and with short handles halfway through the long bodies.

Using this same principle and typological information, different model designs (PAI) could label and identify the objects in different ways. We could have the objects grouped and labelled using shape (e.g., ovoid, cylindrical) their

differentiate between them. These characteristics are being transmitted during the training process as follows:

- 'Punic': shape (ovoid), neck (very short), necks (very simple), handles (small, usually attached high on the body just before the sharp angle towards the neck). These are just some of the more obvious characteristics.

- 'North African': shape (long, cylindrical, narrow), handles (placed halfway through the long body), neck (marked, concave), rims (flaring outwardly).

- 'Italic': shape (generally defined by Graeco-Italics, Dressel and Lamboglia types), handles (big, looping), necks (usually long). These are just some of the more obvious characteristics.

- 'Local': More heterogenous group, but unique in their own way as well.D

While not flawless, the metrics of **T4** models are far more promising than those of **T1**, **T2**, and **T3**. Notable improvements include an average accuracy of **61.1%** (though this is heavily influenced by high-performing classes like 'NAfrican' and 'Punic'), an overall **AR of 0.611**, and a **42.8%** precision for 'Local' vessels, which is a sharp improvement compared to the earlier typological models. Still, challenges remain particularly with the 'Italic' forms, which show a precision of just **21%.** The limiting factors for these forms remain consistent with the issues identified in the typological breakdown of **T1**-**T3** models (which we will next analyze).

Due to the different frameworks used to identify materials, **T4** models stand apart from other typological models in their hypothetical applications. Unlike traditional typological models, which classify materials based on their specific typologies, **T4** models classify objects by their geographical origin. This shift in approach still relies on typological information but groups materials based on the region of production. This innovative concept not only finally proves the model's ability to differentiate between materials based on typological information but also opens the door for developing similar models that use typological data in a more flexible and context-driven way.

Analysis: In an underwater archaeological environment, many of the parameters that object detection models rely on to classify objects are distorted. Ceramic vessels already share many similarities to begin with, with typological distinctions often depending on subtle details like lip

date (e.g., 2nd century BC, 1st century BC, IV century AD) their use (e.g., kitchen ware, table ware, transport) and many others, all answering to models designed to answer archaeological questions related to those factors.

curvature or handle width. Underwater, these vessels appear in various states of preservation, further obscuring the distinguishing features needed for typologyical classification. Additionally, centuries of exposure, the diffused nature of underwater light, and the biological growth on artefacts result in uniform textures and colors across different materials. The challenge is further compounded by the usage in this project of small and fast versions of YOLO that reduce image resolution to optimize speed and memory usage. Given these limitations and our understanding of the mechanics involved in the classifying methods used by detection algorithms,[108] it was clear from the outset that training a detection model for typological classification would be difficult, especially for models intended for real-time application.

We have stated that **T1**, **T2** and **T3** models are unsuitable for field applications. If we follow the breakdown of individual typologies across their metrics (Tables 7a-7c), however, they can be considered useful in revealing clear patterns in what the detection model could and could not identify correctly, which can lead us to future improvements and model design ideas:

-Patterns in successfully identified types:

- 'Form6' (Maña C) was always correctly identified in the two instances it appeared in the test—though this was allegedly due to ideal positioning, preservation and lighting rather than a robust classification ability. These amphorae are a rare occurrence on the site, which limits our possibilities for more definitive conclusions.
- 'Form14' (Dressel 20) and 'Form3' (Ramon 7.1.1.1) also showed promising results, with precision rates above 50% for the first in most models and 70% for the second in **T1**. This can be attributed to their distinctive shapes: In this catalogue, Dressel 20 are characteristically globular and Ramon T-7.1.1.1 characteristically narrow and elongated. This also applies to the Maña C from the first point.

- Patterns in frequent forms that skew the average:

- The 'Indetermined' class accounted for a significant portion of identifications in **T1** and **T3** models (over 4,000 detections, just below **60%** precision in both groups).
- Materials form the 'Punic' class constitute **79.2%** of all identifiable types in the test dataset.

[108] Yolo balances an object's class using a probability-based class score achieved through a variety of techniques throughout its DL structure: Edges textures and shapes are extracted using a CNN. Multi-scale detection keeps track of the classes bounding box sizes. Also, the model learns color, spatial and texture patterns typical from each class. These are just some examples of those mechanisms (Terven et al., 2023: 1686-90). More information available on Appendix I (p.153).

Being also easily distinguishable, **T3** models achieved **80%** precision rate on predicting them when they were grouped into a single class (as opposed to **T1** models where 'Form1', the most common variation of Punic material in the site, managed a **70%** precision rate).

- Patterns in poorly identified types:
  - The models consistently failed to correctly classify 'Forms 4, 5, 7, 8, and 9'. These correspond to Graeco-Italic amphorae, Dressel 1, Sagona urns III-IV: 4a-b, Sagona urns III-IV: 3, and Malta 1 amphorae.
  - These types are similar in shape (elongated vessels with robust bodies, and defined handles) and infrequent at the site. Adding this to the size of the dataset used during this dissertation (when compared with the dataset we would have to use to develop typological models for real-world use) and the poor state of preservation of many materials, we have the reasons for the model's misclassification issues.
- Overall model performance:
  - Recall metrics for all three model groups were low (**T1: 0.339, T2: 0.265 and T3: 0.376**), indicating both high misclassification rates and a tendency to miss objects entirely.

As we can see, while the typological models developed in this project did not yield practical field-ready models, they revealed key trends that suggest future improvements are possible:

-Certain vessel types are inherently easier to classify than others due to their distinct shapes and frequent appearances in the dataset.

-Direct typological classification does not work well in current YOLO-based models **under the specific training conditions** used during this dissertation. With improved training data, refined detection algorithms, and more focused experiments, typological models could very well become viable in the future.

-Broad classification by geographical origin (**T4** models) proved more effective, opening the door for future improvements.

If these challenges can be addressed, the applications of typological models would extend far beyond those of nature and state models. From an archaeological perspective, a working typological model would allow us to instantly build an understanding of a site's ceramics in terms of characteristics, distribution, and state of preservation, all of which have applications on

underwater assemblages that feed directly into a site's ceramic study and all which that entails.[109] Numerous studies highlight the significance of specific ceramic typologies in underwater contexts (Malfitana, 2008; Sukkham et al., 2021; Fan and Li, 2021). A detection model capable of recognizing ceramic typologies would not only streamline these studies by accelerating the identification process but also introduce new dimensions of interpretation. For instance, it could enable automated counting of each ceramic type and provide a visual representation of the entire assemblage categorized by typology, which would be particularly useful for research analyzing the spatial distribution of ceramics in ship cargoes, exploring deposition patterns to characterize potential multi-shipwreck sites like Xlendi Archaeological Park, or classifying objects based on chronology, geographical origin, or even by their contents.

Example of Use: As an example of how Typological PAI can be applied in a practical scenario, let us consider a research project focused on reconstructing ancient trade networks that connected Malta and Xlendi to the wider Mediterranean area. This study would rely on the ceramic assemblage at Xlendi Bay as key evidence.

To achieve this, a detailed analysis of the site's pottery and its origins is required. While this could be done manually by examining and recovering amphorae, a detection model significantly streamlines the process. However, not all model types are suitable for this research goal: nature models and state models do not provide typological or geographic information, making them ineffective for analyzing trade networks. Typological models, in contrast, do offer that information, making them the best choice for this study.

Among typological models, **T4** models stand out as the optimal selection, as it classifies archaeological materials based on their geographical origin while specifying exactly which typologies were merged to form a class (Table 4b, Figure 38). Choosing a **T4** model over a model that categorizes ceramics by typology alone allows for instant visualization of the site's ceramic distribution by origin when projected onto an orthomosaic.[110] Additionally, it provides quantitative data on each ceramic type's origin, offering valuable insights into trade connections, making it the best model to answer our research question.

[109] Summary of how ceramic studies can be used in the context of maritime archaeology (p. 32, 67).
[110] Orthomosaic (p.15).

In summary, the test results for typological models offer valuable insights into the limitations that currently prevent their full implementation in real-world scenarios. It is important to note that these tests were conducted using model versions and sizes optimized for real-time performance by reducing input image resolution, which is not an essential constraint. A closer examination of the typological models reveals that they function as intended despite being unproductive, and future advancements—such as improving the PAI, refining training parameters, and adjusting model versions and sizes—could eventually lead to the development of a field-ready typological model.

6.2.3 How models compare at model version/size level

During this dissertation, we tested six different combinations of YOLO versions and sizes. The selection was based on previous work with underwater assemblage detection and the need to keep tests simple and replicable on lower-end hardware units. Because of this, it is worth noting that the more complex models defined here would achieve higher degrees of performance when used on larger sized versions of the YOLO architecture than the ones we use here. This could be especially relevant in cases where a project has access to powerful hardware and does not need real-time processing capabilities.

In addition, several conclusions and patterns emerge when comparing the performance metrics of different YOLO versions and sizes, as shown in Tables 5-7. Before delving into these, however, there are some contextual factors to consider regarding the particular versions and sizes used for this project:

-Versions (YOLOv8 and YOLOv11): YOLOv8 has been the most widely used version for underwater assemblages (Paraskevas et al., 2023; Kamal et al., 2024; Zammit et al., 2024). YOLOv11 is a newer version of this family of algorithms released in September 2024. Supposedly, it provides improvements across the board but for speed—which means it should improve performance according to official sources (Figure 42, Table 9).

-Sizes (nano, small, medium): There is a wide variety of available sizes for YOLO models. The main difference this makes to archaeological detection models is that larger models are slower than smaller ones, which limits their capability to be integrated in a real-time application like the one deployed by Paraskevas et al. (2023). They are also able to handle larger images and show overall

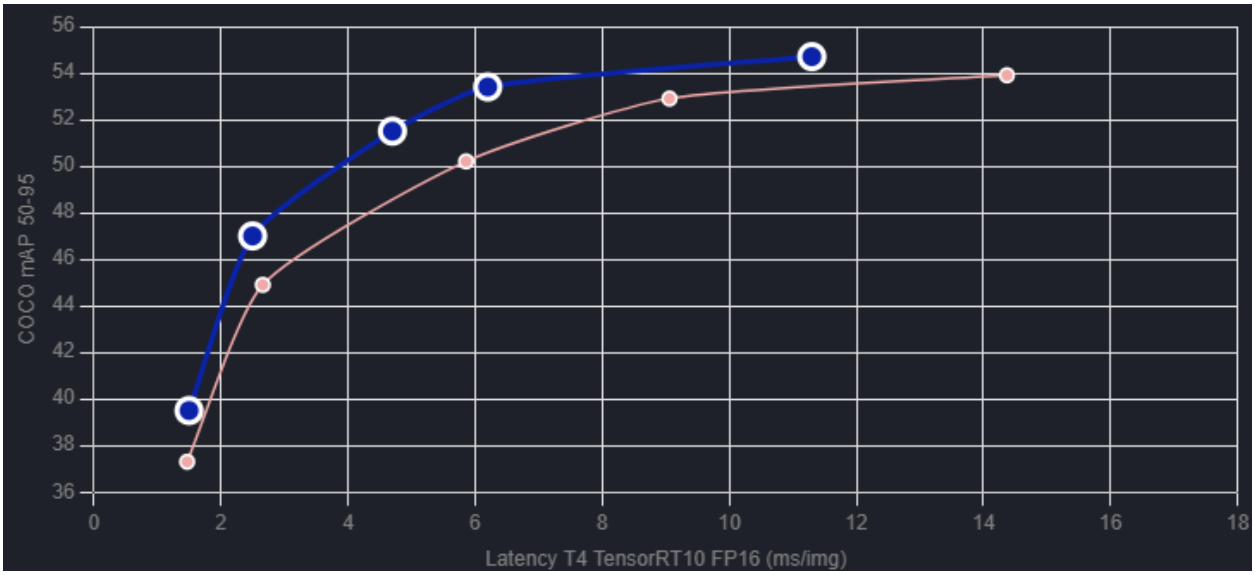


Figure 42. Official efficiency/speed comparison between YOLOv11 (blue) and Yolov8 (pink).

Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

a)

Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed T4 TensorRT10 (ms)	params (M)	FLOPs (B)
YOLO11n	640	39.5	56.1 ± 0.8	1.5 ± 0.0	2.6	6.5
YOLO11s	640	47.0	90.0 ± 1.2	2.5 ± 0.0	9.4	21.5
YOLO11m	640	51.5	183.2 ± 2.0	4.7 ± 0.1	20.1	68.0
YOLO11l	640	53.4	238.6 ± 1.4	6.2 ± 0.1	25.3	86.9
YOLO11x	640	54.7	462.8 ± 6.7	11.3 ± 0.2	56.9	194.9

b)

Table 9. a) Official YOLOv8 performance metrics. B) Official YOLOv11 performance metrics.

better prediction performance. We can see this relation in the official metric tables for each of the versions used during the tests (see Tables 9a, 9b).

When evaluating the performance of different YOLO model architectures during our tests, a key question arises from the perspective of maritime archaeology: Is there a specific architecture that stands out as the optimal choice for archaeological detection?

•**Nature models (N):** Looking at nature models first, we observe enough divergence between the best and worst performers to answer negatively to the question. Interestingly, **N1** models showed the best performance with YOLOv8small, which aligns with results from previous tests that inspired their design (Table 8). However, for other nature models (**N2** and **N3**), YOLOv11nano emerged as the best performer. It is important to note that **N2** and **N3** differ from **N1** and **N4** models in that they involve fewer objects to identify (1929 and 1528 for **N2** and **N3**, compared to roughly 4600 for **N1** and **N4**), but this difference alone is insufficient to establish a clear pattern establishing that the smaller models work best because of this reduction in object density—though it does seem a pretty fair assessment.

Ultimately, for nature models and looking not only at the best and worst performers, YOLOv11 models slightly outperformed YOLOv8 models across various metrics. Moreover, since YOLOv11 is a bit slower than YOLOv8 (Table 9), this difference may be exacerbated in scenarios where performance is prioritized over real-time processing constraints through the use of a larger YOLO model.

•**State models (S):** The performance of state models in terms of version and size differs from that of nature models. While there is still some variability across all PAI groups, three out of PAI (**S1**, **S3**, and **S4** models) performed best with YOLOv11small. However, **S2**, **S3**, and **S4** models also present a different size of YOLOv11 as the worst-performing option. Other than this, no clear preference emerged for one version or architecture size.

•**Typological models (T):** For typological models, the analysis is more complicated due to the large number of classes involved as, in some cases, a specific model might perform better at recognizing certain typologies based on their size. However, the experiments for these PCI groups were not extensive enough to draw definitive conclusions on this. Superficially, we could say that

the PAI groups with more classes (**T1** and **T2** models) tend to yield better results with YOLOv8, while **T3** and **T4** models perform better with YOLOv11.

In general, no significant difference can be observed across typological PAI regarding architecture size but for one notable exception: **T4** models. For **T4**, performance consistently improved across all metrics as the model size increased whether using YOLOv8 or YOLOv11. This trend is especially pronounced with YOLOv11, which shows substantial performance gains in all metrics, including mAP50:95 (nano: **38.4%**, small: **43.1%** and medium: **47.8%**), average recall (nano: **0.518**, small: **0.552**, and medium: **0.611**) and mAP50 (nano: **54.1%**, small: **58.7%**, and medium: **61.1%**). This pattern is made significative by the fact that **T4** models are the only typological PAI that shows promise for field implementation based on the preliminary and very improvable experimental tests conducted during this project.

The following conclusions can be made for the version/size of comparison of the project's experimental models:

-Overall, while no single model architecture is the clear best option across all PAI groups, YOLOv11 medium slightly outperformed YOLOv8, making it more suitable for when real-time processing is not a constraint.

-Nature and state models did not show a clear preference for specific architectures, but **N4** models consistently improved in performance as model size increased.

-Typological models displayed a notable pattern: PAI groups with more classes performed better with YOLOv8, whereas those with fewer classes favored YOLOv11. **T4** models stood out as the only typological PAI that showed a clear performance improvement across all metrics as model size increased.

- When it comes to detection models in underwater archaeological assemblages, larger models can be expected to produce even better results in some cases, suggesting that future experiments with larger architectures than the ones used here could enhance detection performance beyond the scope of this project.

7. DISCUSSION

Contents

7.1 Summary

During this research project, a total of 72 detection models were trained to identify archaeological materials from Xlendi Archaeological Park, an underwater site west of Xlendi, Gozo. These models were differentiated based on three factors.

Two of these factors were specifically designed to align with archaeological methodologies. First, the PCI categorized the 72 models into three groups based on complexity: Group "N" focused on the nature of the materials present on the site, Group "S" addressed their state of preservation, and Group "T" examined their typology.

The second factor, the PAI, further subdivided these groups into four categories. In this way, group "N" was divided into N1, N2, N3, and N4, each comprising models tailored to test specific archaeological hypotheses or provide comparative insights. The same structure applied to groups "S" and "T."

The final factor distinguished these groups into sets of six models that used different variations of version and size.

Additionally, a typological chart of Xlendi Archaeological Park was developed using site images and incorporated into the models. Type identification was conducted through comparative shape analysis, referencing relevant publications and emphasizing physical properties deemed critical for computer vision techniques.

The predictive capabilities of the trained models were assessed using unseen data from a sector of the site not included in the training stage. This evaluation involved testing the models on new sets of photos and a video. On a general level, baseline models detecting the presence of ceramic materials on the seabed performed with near-flawless accuracy; models assessing object preservation states were slightly less precise but still reliable; and typology-focused models, while showing some success, demonstrated significant room for improvement.

Further evaluation involved comparative analysis across the model groups to validate the observed trends using quantitative metrics. The results confirmed that detection models can be designed to address specific archaeological inquiries effectively. Moreover, the designs used during this dissertation highlighted some of the ways in which automated object detection can be useful as a tool for the study of underwater assemblages, one that may be used without the support of complicated techniques reliant on site-specific situations.

This chapter aims to interpret these results in relation to the project's objectives and to contextualize their applications (the ones mentioned) within the broader study of underwater assemblages and maritime archaeology.

7.2 How the Results Relate to the Aims

The first aim of this research was to define automated object detection for use in maritime archaeology. While extensive maritime archaeological literature exists on common computer vision techniques like photogrammetry (Drap et al., 2015; Rosic et al., 2019), automated object detection being applied to assemblages is often mentioned only in passing within broader projects or as part of new methodological approaches to specific sites (Pasquet et al., 2017; Paraskevas et al., 2023; Yang et al., 2023; Kamal et al., 2024; Zammit et al., 2024). Moreover, most discussions on adapting detection techniques to archaeology have focused on terrestrial contexts; and while the fundamental principles of object detection remain the same for both land and underwater archaeology, their applications and effectiveness vary significantly due to the challenges posed by the underwater environment. In this research we underlined these differences, situating automated detection within the broader history of AI implementation in maritime archaeology. In addition to this, we explained automated detection techniques from a theoretical and practical perspective, ensuring accessibility for readers without expertise in AI. Overall, without serving as a manual,

this study presents a clear and practical approach to implementing object detection, aiming to facilitate its adoption in future projects and stimulate further discussion about its potential.

The second aim of this project was to evaluate automated detection as a tool for analyzing and interpreting underwater assemblages. This involved more than just theorizing over the methodology to determine its effectiveness or perhaps using it in conjunction with other methods. In our opinion, fulfilling the proposed aim required us to apply the technique independently to a real-world scenario without relying on additional support: the evaluation of detection had to consist of an assessment of its limitations, potential applications, and potential adaptability to other sites while also considering the learning curve associated with its implementation. We achieved this by testing the methodology on the underwater assemblage of Xlendi Archaeological Park. By analytically comparing the results from the different detection models that show promise for reiterative application in maritime archaeology, this study provided a contextual understanding of the technique's capabilities, ultimately contributing to a broader discourse on its potential as well as areas for further experimentation.

The third aim was to test the hypothesis that object detection models are most effective for underwater assemblages when designed, implemented, and interpreted from an archaeological perspective. We demonstrated this throughout the project. More concretely, we approached it by developing a theoretical framework that shows how models can be structured and trained to address specific archaeological questions. According to this framework, two archaeological factors to differentiate trained models on an archaeological level were successfully introduced. These factors played a key role in validating the hypothesis. This is evidenced by the fact that the models constructed with them, when applied to an underwater assemblage, returned results confirming that detection models can be purposefully designed to incorporate archaeological information, thereby producing models whose outputs are informed by archaeological reasoning.

7.3 Model Design. Introducing Subjective Information into a Mathematical Model

In this project, we have demonstrated that training an object detection model like YOLO for submerged archaeological materials follows the same fundamental steps as any other detection task would.[111]

The first stage involves constructing a dataset that is both large and diverse enough to ensure robust training and reliable testing on unseen data. The second stage is the labeling process, the only phase where the user directly influences the trained model. Here, the user defines the ground truth—establishing the patterns the model will learn through bounding boxes and classification. The final stage, training, is fully automated. It is in this stage that the algorithm refines its predictive parameters based on the ground truth, identifying objects using features such as color, texture, shape, and spatial relationships.

In most object detection applications, these patterns are purely mathematical. The integration of subjective information into detection models is rare, occurring primarily in fields such as marketing analysis (e.g., detecting emotions like happiness or anger), food quality assessment, or Esports AI coaching. Even in these cases, subjective information is only detectable when physical patterns from the image's objects provide a measurable correlation to the information.

The inspiration for this project stemmed from the realization that while archaeological interpretation parameters are inherently subjective, they are often grounded in tangible, physical patterns that a detection algorithm can identify. For example, although definitions of "state of preservation" may vary among archaeologists, differences between materials in this regard and on an underwater assemblage can often be inferred from their general size, shape, and edge texture.

According to the training process outlined above, the only point where archaeological subjectivity can be introduced into the model is during the labeling stage, where the user teaches the algorithm what to recognize. To test this approach, we needed a controlled method for gradually incorporating archaeological information into the labeling process. In this regard, the classification system used in this project served a dual purpose: structuring the labeling stage and providing a standardized way to name models, facilitating comparative analysis (Figure 23).

[111] Whether identifying basketball players in live broadcasts, human figures in CCTV footage, or defects in computer motherboards, the core process remains consistent.

From this foundation, we systematically increased the complexity of introduced archaeological information until we reached the limitations of smaller YOLO models. Each trained model was designed with specific archaeological applications in mind:

-N1 models identify every piece of archaeological material in the field, effectively providing a visual representation of each image's NISP.[112] When projected onto a 3D orthomosaic (Zammit et al., 2024),[113] these models offer site-wide NISP analysis and are useful for assessing material density or large-scale surveys using cameras to detect anthropic evidence. S1 models, for their part, maintain the benefits of N1 models while also categorizing materials by state of preservation. This allows for NISP-based site analysis without significant loss in recall and with the added benefits of state models, ensuring comprehensive detection with minimal false negatives.

-N3 models disregard fragments that are not individually representative of a single vessel. Instead, they focus on counting materials with identifiable elements or specific sizes, effectively providing MNV[114] data across the site when projected onto a 3D orthomosaic. This makes S3 models particularly useful for studying deposition patterns of complete vessels rather than dispersed sherds affected by underwater currents. They serve as the counterpart to N3 models, maintaining MNV-focused analysis while also categorizing materials by state of preservation.

-N4 models are designed for surveying and ecological purposes. These models can classify archaeological remains while also identifying and quantifying other human-related elements such as plastics and bottles. With further development, such an idea could lead to models that classify unexploded ordnance (UXO), marine biota, and other types of debris in relation to archaeological materials.

-State models in general provide applications for studies requiring archaeological materials to be categorized by state of preservation—an inherently subjective concept. By adjusting labeling parameters based on observable physical properties, the models can distinguish between a 'broken' and a 'buried' vessel using differences in shape, edge characteristics, and size. Refining these classification parameters led to measurable changes in model predictions and performance metrics, demonstrating the influence of labeling decisions on detection outcomes. S4 in particular is

[112] Number of identified specimens (p.53).
[113] Orthomosaic (p.15).
[114] Minimum number of vessels (p.53).

optimized for scenarios where the focus is solely on complete materials, excelling at isolating intact artefacts from background noise and other anthropic elements while improving precision over other state models. In general, state models are valuable for understanding site morphology and site formation processes as well as assisting during mission planning.

-Finally, typological models, while not fully functional in real-world applications, were designed to address specific and complex archaeological questions. Despite their limited success, we observed measurable improvements in both output and metrics across various PAI that were critical to confirm the possibility of having functional typological models in the future. On this note, larger datasets and YOLO architectures would likely improve our chances at increasing typological model performance.[115]

Successfully implementing such models could enable automated typology detection, facilitating instant site characterization and dating. In cases like Xlendi Archaeological Park, these models could help answer unresolved questions about site formation by mapping deposition patterns based on artefact typology. Ultimately, T4 models show the true scope of possibilities available for typological models as well. These can be designed to detect and classify materials based on their geographical origin, chronological differences, or even the nature of imported and exported commerce goods.

The results of all these trained models confirm that it is possible to integrate subjective archaeological information into detection models. Performance metrics further support these observations. N1 and N3 models achieved 77% and 85% mAP50 respectively when tested on unseen data.[116] As the amount of embedded archaeological information increased, while the models remained functional, overall model accuracy declined—dropping to 69% for S1 models and 38.5% for T1 models. In the case of typological models, however, analyzing class interpretation patterns revealed a direct link between decreasing precision and dataset limitations, highlighting the need for more comprehensive data to refine the models' ability to recognize subtle typological differences because, notably, typologies with clearer physical distinctions consistently showed better precision rates.

[115] Possible influence of model version and size on the results (p.113).
[116] See Appendix IV for metrics (p.209).

In essence, it could be argued that this research has gone beyond the use of conventional automated object detection. It has pioneered an approach specifically tailored to the needs of maritime archaeology, demonstrating that subjectivity in archaeological interpretation can be systematically integrated into automated detection DL models.[117] This is automated **archaeological** object detection.

7.4 Exportability.

7.4.1 Rigid Detection vs. Flexible Detection.

In maritime archaeology, object detection is primarily a data management tool. While it enhances scientific thought through visualization, pattern recognition, and statistical data generation, its main value lies in its ability to process data automatically. This is particularly evident in the past two decades, during which detection has been used mainly to locate and identify archaeological sites on the seabed. At that stage, however, one might argue that the technique's productivity was limited to the enhancement of underwater surveying.

Let's consider most underwater assemblages with ceramic remains—typically, single shipwrecks with large, localized cargoes of amphorae concentrated around the hull, with some additional materials scattered nearby. In these cases, a detection model could be useful, for example, when implemented on a surveying ROV.[118] However, the likelihood of a detection model ever being trained for such a project is low. The reason is simple: unless a site is of exceptional importance and uniqueness, maritime archaeology is always on a race to gather information as efficiently and economically as possible, and detection methods do require specialized knowledge not so readily available. To use them, an archaeologist must either invest time in learning or rely on specialists willing to collaborate. In addition, since in maritime archaeology object detection is at its core a data management tool, it has historically not been an efficient choice for teams studying a single shipwreck to develop complex models when the data they produce could just as easily be processed manually without significant difficulty.

[117] Deep learning (p.11).
[118] ROV (p.8).

Because of this, so far detection models have been specifically developed in only two types of scenarios. The first involves sites with vast assemblages such as the Xlendi Archaeological Park (Figure 15) or the Ming Dynasty Slope I and II shipwrecks[119] where the need to process such tremendous amounts of data justifies the investment in automation. The second applies to highly significant sites where well-funded archaeological teams push technological boundaries for comprehensive data collection such as the Phoenician shipwreck off the coast of Gozo, Malta (Kamal et al., 2024). This does not mean that smaller sites would not benefit from object detection, particularly models capable of identifying MNV, assessing preservation states, or classifying typologies. Rather, it means that developing detection models solely for the benefit of these smaller sites is not practical. In fact, in many cases, it is not feasible either: This is because detection models are extremely data-hungry; and thus they perform best when trained on large and varied datasets, allowing them to generalize across different conditions.[120] As a result, a model trained on a single, small shipwreck—where images are captured under limited lighting conditions, with minimal variation in shape, potential obstructions like reefs, or differences in photogrammetric runs—would lack the flexibility to function in even slightly different scenarios. In other words, a model trained on a small site would be rigid, with no practical applications beyond that site.

If a detection model is to be truly exportable to other sites, it must be trained on one that provides the foundations for the creation of robust models capable of testing successfully on multiple others across the Mediterranean—or any other sea. Xlendi Archaeological Park is one of those sites.[121] In such a scenario, many of the limitations that previously hindered the deployment of detection models on small research projects would no longer apply. Once trained on one of these ideal sites, the resulting model could be applied almost instantly to similar others, thus eliminating the need for archaeologists to develop new models from scratch—which would make readily available trained models a tremendous asset. In this regard, it is particularly interesting to consider the concept of implementing transfer learning techniques.[122] Conceptually, these would be useful when dealing, for instance, with the exportability of typological models as follows: On a

[119] Appendix VI (p.215).
[120] Explanations for "data hunger" (p.8) and model robustness (p.9).
[121] The models developed in this research are inherently limited in flexibility due to the dataset's constrained size. Since this project serves as a testing ground for multiple models rather than an actual research-driven initiative, creating a large and varied dataset was neither necessary nor practical. However, in theory, Xlendi Archaeological Park is the perfect scenario for such a thing. See p. 36 for Xlendi's Value for Automated Object Detection.
[122] Transfer learning (p.15).

typological level, not every site will contain the same types of vessels. A Punic shipwreck in Malta may feature fifteen types, while a similar wreck in Tunisia may share only ten of those types but introduce five entirely new ones. In such cases, with a bit of effort and if the original training data were provided alongside the model (transfer learning), a more complete model could be refined by incorporating these new elements from a site not large enough to produce its own versions.

To highlight the theoretical exportability of the models trained for this dissertation, Appendix VI[123] considers three past studies that correspond to typical underwater site archetypes as examples of how **our limited** testing models could be applied.[124] From the previous points and by analyzing these case studies, we can better understand the adaptability of our trained models across different underwater archaeological environments. Each site presents unique challenges, requiring varying degrees of model modification or redevelopment. They also highlight a few key factors regarding the exportability of trained detection models.

The most important one is that automated detection methodology cannot be understood as a general point-and-click tool from the moment a trained model is created. Instead, the idea of creating practical and exportable trained models relies on them being considered more as links within a constantly evolving and expanding network of applicability. For models to be easily exportable, they would need to be trained in sites that lend themselves for training within a particular geography or chronology, which would in turn make the trained models suitable to be used in those same chronological, cultural and geographical spheres. We could call these base training sites. For instance, Xlendi Archaeological Park could be a base training site for spread Punic assemblages with a chronology between 5th century BC and 2nd century AD. In this case and as shown in the examples from Appendix VI[125], the percentage of models trained in Xlendi Archaeological Park that would be exportable without modifications would decrease as the importing site's characteristics differed more and more from those of Xlendi Archaeological Park. Also, the examples show that when it comes to exportability, we must consider that environmental differences also present a challenge, as a model trained in the Mediterranean would likely struggle

[123] Appendix VI (p.215).
[124] Since mentioning every way in which we could integrate detection models into these very nuanced sites is unrealistic, we will simply give some examples of how our models could be applied while mentioning some ideas for new PAI.
[125] Appendix VI (p.216).

to perform in the Baltic due to vastly different underwater conditions. In this way, each base training site would work as an area of influence for detection methods, forming a network.

The second is a conclusion that stems from all the above: while the creation of fully exportable archaeological models remains a medium-term goal, their development appears inevitable.

7.4.1 Ambiental Factors.

Because of the way YOLO algorithms function—recognizing patterns based on the spatial distribution of pixels—any disruption that affects lighting or visibility during data acquisition can significantly reduce the model's accuracy and reliability. These disruptions often stem from issues during the data collection stage. For instance, as it happens on this dissertation, if the photographs were taken manually, they will show inconsistent lighting, unusual angles, or disturbances caused by divers or their camera setups. These issues are generally correctable with improved data-gathering techniques.

Environmental (ambiental) factors, on the other hand, are not as easily controlled and often play a secondary—but sometimes more impactful—role in determining the exportability and generalizability of a trained model, even more so than the issues of training robustness discussed above. Some examples of these factors are visibility, depth or the visual effect of marine biota. Different seas, and geographic regions in general, present differences reflected in these factors. Depth, for example, affects how much light penetrates the water—thereby influencing image quality. Water transparency and the presence of vegetation have similar influences on the ratio in which they can obscure critical visual information.

As a result, a model trained under one set of environmental conditions is unlikely to perform well under significantly different ones. The drop in performance will often be proportional to the difference in environmental characteristics. For example, a model trained to detect underwater archaeological assemblages in Xlendi, Malta—at a depth of 115 meters with clear water and minimal marine growth—would likely struggle when applied to a 30-meter-deep site in the Baltic Sea, where visibility is lower and the site may be heavily covered in algae.

To address this, new models must be trained for specific marine environments according to the subjective needs of the researcher. The unique characteristics of each environment should be documented—ideally in a standardized format akin to a the PAI, or embedded directly in the

model's own metadata. These considerations reflect the full extent to which environmental factors affect the exportability of underwater archaeological detection models.

7.5 Automated Detection's Projection as a Tool

When considering the general applications of automated object detection in the study of underwater assemblages as mentioned in the previous points, two distinct avenues of use emerge. The first is the approach explored in this project: using detection models independently without considering supplementary techniques and their integration into specialized methods. The second involves object detection as exactly that: a supporting tool within more complex, interdisciplinary research projects that make use of additional techniques. Aside from this, it is important to note that the contextual applications of detection are so extensive that it is best to present their possible uses as roles they could play in a research project, based on the characteristics of their output format. Having mentioned many of them during the project, they can be outlined as follows:

The primary advantage of automated object detection in underwater archaeology lies in its efficiency. By automating data processing, researchers can save valuable time that would otherwise be spent on manual identification and classification. In addition, automation inherently reduces human error, which is inevitable in repetitive tasks, ensuring greater consistency in data collection. This advantage applies to both independent and integrated uses of detection models.

Another key application of trained detection models is their role in visualization and pattern recognition, particularly as a visual aid. As demonstrated by Paraskevas et al. (2023: 7), DL techniques show great promise when incorporated into real-time surveying tools, such as the video feeds of ROVs. Paraskevas et al. also suggested that detection models could classify different types and instances of pottery sherds (Paraskevas et al., 2023:7). This hypothesis has been confirmed in this project, further enhancing the potential of detection both as a visual aid for manual archaeological investigations and for remote, noninvasive study of overarching aspects of submerged assemblages.

In addition, object detection can contribute significantly to understanding archaeological sites in terms of morphology, formation, preservation, relation with the marine environment and public outreach. A model trained to classify amphorae by state of preservation, for example, can help identify patterns of material distribution and provide insights into post-depositional processes.

Figure 43 illustrates this point, showing how distribution patterns align with theories attributing the formation of Xlendi Archaeological Park to post-depositional movement rather than multiple shipwrecks or ritual deposits. Similarly, detection models can be of great use during mission planning, giving a team the ability to quickly assess the distribution of materials, thus allowing for more efficient recovery operations. This reduces diver time on the seabed, conserving resources and minimizing exposure risks.

As for the second avenue of use, detection techniques also have the potential to be integrated with a wide array of techniques and advanced computer vision methods, leading to an entirely new level of applications. The specific benefits in this case depend on the characteristics of the site and the ingenuity of those implementing the technology. For instance, Kamal et al. (2024) developed a 3D instance segmentation methodology that projected 2D object detection data from a YOLO model into a more complex 3D model. This approach provided a non-invasive technique for visualizing stratigraphic relationships, offering an unparalleled tool for analyzing and interpreting the Xlendi Phoenician shipwreck. Other studies, such as Zammit et al. (2024) have explored the integration of object detection with precision-enhancing techniques, using trained models to map object locations onto a photogrammetric orthomosaic. This method enables automated documentation of an archaeological site by including material counts by class, which provides data in the estimation of cluster distributions and enhanced long-term site monitoring among many other potential advantages in terms of archaeological research.

Despite the vast potential of the technique, the application of object detection in maritime archaeology remains an underdeveloped field. Attempting to list all possible uses at this stage is impractical, as the full extent of its capabilities is still unknown. What is certain, however, is that object detection has applications far beyond assemblage analysis. It could be used to improve our understanding of ship construction by analyzing features on hull planks, or to enhance our study of the relationship between marine biota and wrecks by detecting, locating, and classifying biological growth on submerged artefacts and wrecks. The possibilities are endless.

In truth, this dissertation set out to explore the limits of conventional detection methods when used without additional support, only to discover that these limits are not defined by the technology itself, but by our imagination and the inevitable future development of new detection models.

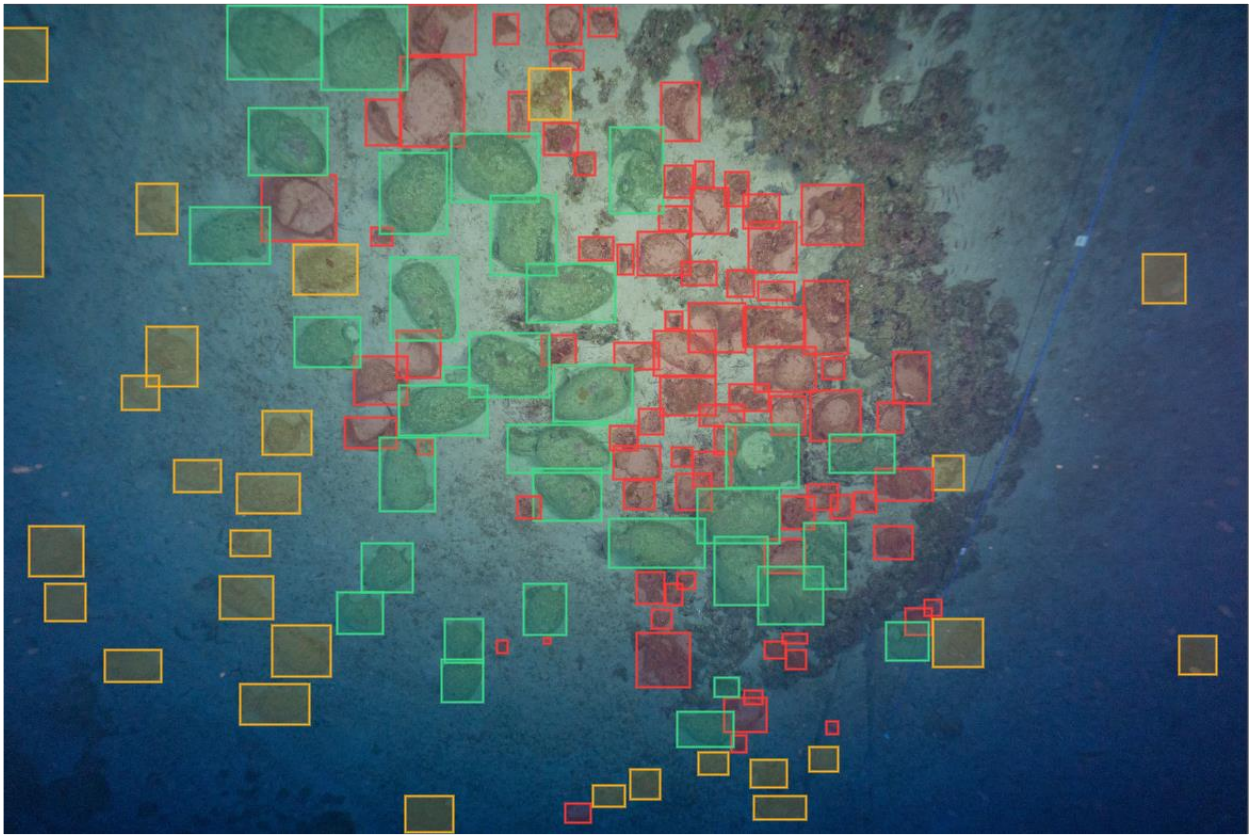


Figure 43. Capture of an image from the training process of a state model (S1). The location of 'complete' (green), buried' (yellow) and 'broken' (red) materials around the rocky reef evinces the presence of a pattern based on size and state of preservation, which is consistent with the theories that ascribe postdeposition pattern as the main cause for the formation of Xlendi Archaeological Park instead of it being the result of multiple shipwrecks or ritual overboard loss.

7.6 Learning Curve. Addressing Potential Concerns Regarding the Use of AI

It is fair to say that the use of AI techniques presents several challenges and concerns. We will specify them here before addressing them individually for the purpose of computer vision and detection methodology. In our opinion, the challenges can be summarized as follows:

-Loss of human expertise: This refers to the notions that AI systems lack the nuanced understanding and contextual knowledge that an archaeologist brings to the table, and that leaving the AI to make decisions might mean overlooking critical factors not so easily perceived by it such as project contextual history, environment, and culture.

In our opinion, this concern is less relevant when it comes to computer vision. Unlike generative AI models like ChatGPT, which create content autonomously from whatever source they see fit, computer vision techniques simply automate processes defined by the human expert that wants to use them—or by another expert's parameters, which are explained through the trained PAI if a model has been exported. This is to say that detection models do not decide what is archaeologically important. Instead, they generate predictions based on what the user deemed important through the labelling process.[126]

Another concern at this level is the potential over-reliance on detection models. While questions about whether AI could diminish traditional fieldwork skills are very fair, these do not apply to detection techniques. This is because training a model with just 5% of a dataset—saving hundreds of hours of manual work—does not eliminate the need for archaeologists to develop data-processing skills. On the contrary, for the model to function effectively, the data must be processed with more precision than usual, reinforcing the importance of human expertise in the workflow. Having said this, it is important to point out that detection models should never be seen as shortcuts, but as tools for aiding us to address archaeological questions.

[126] Throughout this dissertation, we have demonstrated how responsive these models are to user input. If an archaeologist integrates their knowledge of cultural and environmental factors into their workflow, the AI's predictions will inherently reflect that context. If such factors are not considered, it is not a shortcoming of the AI, but rather an omission in the archaeologist's own approach.

-Data bias and model accuracy: This refers to the notions that biased or inaccurate datasets and labelling process could lead to flawed interpretations.

In this regard, the only thing to say is that computer vision is no different from any other archaeological method.[127] If bias is introduced during training, the results will reflect that bias. This, however, would be a matter of human oversight, not a flaw inherent to computer vision itself: The accuracy and reliability of detection models depend entirely on the archaeologist's discretion in curating and labeling the dataset as well as in interpreting the results.

-Technical and resource-related challenges: Implementing AI in archaeology requires technical expertise, computational resources, and proper funding. Most archaeological teams operate with limited budgets and may not have access to the necessary support to effectively deploy AI-based methods considering all those aspects of it. This is the fairest of all the concerns regarding the use of detection models being handled by non-AI experts, and was taken very into account throughout the whole project while explaining some of the methodological choices

If we accept the hypothesis that detection models are most effective when deployed by an archaeological expert, their true value lies in whether their data-processing capabilities justify the steep learning curve they present to those without prior experience. From the outset of this project, we considered the challenge of defining that learning curve, as it directly impacts the method's exportability. Two key observations emerge from this analysis: First, utilizing pre-trained detection models and interpreting their output requires minimal technical knowledge. Beyond installing free software on a CPU, the process is straightforward and can be explained in just a few simple instructions. Second, designing and training archaeological detection models does demand a foundational understanding of how detection algorithms extract features from images. While programming expertise is not required for any of these steps with today's tools, a working knowledge of detection principles is essential. This level of expertise, references included, aligns with the knowledge provided throughout this dissertation.

In summary, implementing and interpreting trained models is relatively simple, much like other common archaeological techniques. The main challenge lies in designing and training the models, which does not imply that AI experts are necessary for the process. This dissertation, in fact,

[127] More on bias in the context of the training of detection models (p.52).

demonstrates that archaeologists with no prior experience in AI can successfully engage with these tools.[128]

[128] It's important to note that this dissertation should not be seen as a "how-to guide." Future projects may better serve this purpose by breaking down the process in a clear, step-by-step manner for non-experts in AI.

8. CONCLUSION

Contents

8.1 Significance

This dissertation contextualizes the past and present use of AI in non-intrusive intervention techniques in maritime archaeology. It explores how and why the underwater environment has posed significant challenges to the development of reliable detection techniques and turned the first quarter of the 21st century into a period of struggle to overcome these issues. The research highlights how the YOLO family of algorithms has emerged as a viable solution with significant potential in maritime archaeology beyond site detection. Thanks to a combination of speed, accuracy, and simplicity, these algorithms offer accessibility to non-AI experts, opening the possibility for their broader application within the discipline.

Beyond merely discussing computer vision techniques, this dissertation provides a comprehensive understanding of their mechanics and the necessary tools for developing trained models. While many maritime archaeological projects mention AI-based detection techniques as part of their toolkit for achieving specific results, we took a broader perspective. Rather than limiting the discussion to their immediate applications, which can still be done in the future in sites like Xlendi Archaeological Park, we sought to assess, demonstrate and expose the true potential of these techniques within the field.

The research takes advantage of the ideal testing environment provided by Xlendi's Underwater Archaeological Park to explore these possibilities. Through a series of comparative tests, it examines the capabilities of AI-driven detection models when embedded with layers of subjective archaeological information through a simple theoretical framework designed to track and evaluate how the trained models' outputs changed in response to the archaeological information they were taught. To push the limits of the models' ability to absorb and interpret archaeological data, a complete typological chart of the visible surface of Xlendi Archaeological Park was created.

Notable resulting models such as N1, N3, S1, S4, and T4 underscore the potential of these techniques—not only as tools for data management but also as generators of valuable archaeological insights obtained through non-intrusive methods that preserve the integrity of underwater assemblages.

This dissertation also underlines the necessity of a fundamental framework when it comes to the integration of detection in archaeology—both independently and in conjunction with other technologies such as instance segmentation and photogrammetry—as a tool to enhance our understanding and efficiency in studying underwater assemblages. Although the theoretical framework proposed here is open to improvement, it provides a starting point for advancing archaeological object detection as a concept and promoting further progress in the field of computer vision techniques applied to maritime archaeology.

Lastly and more broadly speaking, the dissertation initiates discussions on critical topics that include the method's exportability, particularly interesting applications, the inherent challenges of using AI techniques, and the ceramic analysis of Xlendi Archaeological Park.

8.1 Future Directions

While this research was successful in achieving our objectives, we took a broad approach to many of the topics we touched upon, leaving ample room for further exploration. Several avenues of investigation have emerged from this project, each warranting deeper study. (i.e. the addressment of concerns and challenges archaeologists may have about relying on AI, the analysis of the cost-benefit relation between detection techniques and traditional methods, the creation of a guide explaining the step-by-step mechanics of installing and using the necessary software,[129] and the application of detection models contextualizing the relationship between marine biota and archaeological materials, among others).

Some studies, like those regarding Xlendi Archaeological Park in terms of ceramic and site analysis, are particularly necessary. The typological chart developed here was not specifically designed for studying Xlendi Archaeological Park, but as a tool for detection models, placing its standalone significance beyond the scope of this dissertation. Moreover, the trained models were

[129] CUDA, Pytorch, Pycharm, Makesense and the Ultralytics repository.

not applied to the site itself. Now that these models are available, it would be remiss not to utilize them in a separate study focused on the site itself.

Regarding the process of training and developing models, this research represents the work of an archaeologist venturing into an unfamiliar field. As a result, many methodological steps were necessarily tentative and may appear clumsy from a computer vision perspective. Future projects will undoubtedly refine and improve these processes based on the knowledge gained here. As this research progressed, we discovered new and improved approaches to the task at hand that had to be set aside for the sake of maintaining focus on the aims and the original idea. The benefit of this restraint is the wealth of possibilities now available for future expansion on the topic. The 72 trained models represent only a fraction of the variations tested and considered during this project, and many opportunities in both PCI and PAI classification remain unexplored.

The experiments presented here offer several areas for refinement and improvement. First, every model in this study was trained for testing purposes and would benefit from further development on a larger and more diverse dataset. This is especially true for typological models, where increasing the number of training images—ideally doubling or tripling the dataset while maintaining its density—could lead to significant performance improvements. In addition, while no single model architecture emerged as the definitive best across all PAI groups, YOLOv11 medium models consistently demonstrated superior performance when real-time processing was not a constraint. Future research should explore the impact of model size on accuracy, particularly for large-scale archaeological datasets where real-time application is not essential. Finally, T4 models should be tested using the largest available YOLO versions to assess whether the performance improvements observed in smaller versions hold in that scenario.

Beyond model refinement, future research should focus on the exportability and broader application of detection techniques. This includes developing more robust models for sites beyond Xlendi Archaeological Park and creating a framework for distributing these models along with detailed usage guidelines to interested researchers. Additionally, integrating the baseline for archaeological object detection established in this research with other computer vision techniques could further enhance detection capabilities in the underwater archaeological context. For example, combining model outputs with 3D photogrammetry could enable the creation of fully detected site reconstructions, where each class instance is quantified—something not possible with detection on overlapping photogrammetric images alone.

Improving underwater image acquisition techniques is another crucial area for future research. This could involve experimenting with different lighting conditions, enhancing image quality, utilizing submersible-integrated cameras for consistent and wide-site coverage, and optimizing supporting methodologies to streamline detection.

Together, the ideas presented here offer a promising path forward. By building on the foundations laid in this study, we can expand and refine our practical understanding of automated archaeological object detection and promote its use in the field of maritime archaeology

LIST OF SOURCES

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., & Ahmed, A. A., 2020. Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, 32, 101827.

Akkaynak, D. and Treibitz, T., 2018. 'A revised underwater image formation model'. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6723–6732.

Al-Masni, M.A., Al-Antari, M.A., Park, J.M., Gi, G., Kim, T.Y., Rivera, P., Valarezo, E., Choi, M.T., Han, S.M. and Kim, T.S., 2018. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer methods and programs in biomedicine*, 157, pp.85-94.

Almeida, R., & Dhanasekar, M., 1995. Automated detection of oil pipeline faults using image processing techniques. *Journal of Petroleum Science and Engineering*, 14(4), 253-262.

Anastasi, M. (2019). Pottery from Roman Malta. Oxford: Archaeopress.

Anichini, F., Dershowitz, N., Dubbini, N., Gattiglia, G., Itkin, B. and Wolf, L., 2021 'The automatic recognition of ceramics from only one photo: The ArchAIDE app', *Journal of Archaeological Science: Reports* **36**, 102788.

Atallah, L., Shang, C. and Bates, R., 2005. Object detection at different resolution in archaeological side-scan sonar images. In *Europe Oceans 2005* (Vol. 1, pp. 287-292). IEEE.

Atauz, A., McManamon, J., 2000. Underwater survey of Malta: the reconnaissance season of 2000, in *Institute of Nautical Archaeology Quarterly:* 2811. 22-28.

Atauz, A.D., 2004. Trade, Piracy and Naval Wmfare in the Central Mediterranean: the Maritime History and Archaeology of Malta. Unpublished PhD Dissertation, Texas A&M.

Argyrou, A. and Agapiou, A., 2022. 'A review of artificial intelligence and remote sensing for archaeological research', *Remote Sensing* **14**(23), 6000.

Aubet, M., & Barthelemy, M., 2000. *Actas del IV congreso internacional de estudios fenicios y púnicos Vol. 3*. Universidad de Cádiz.

Aubet, M. E., 2001. The Phoenicians and the West; Politcs, Colonies, and Trade. 2nd ed. Cambridge: Cambridge University Press.

Aull, A.M. and Gabel, R.A., 1989. Machine intelligence applied to radar image understanding. In: International Conference on Acoustics, Speech, and Signal Processing, 1989, IEEE, pp. 1791-1794.

Azzopardi, E., 2006. *The Xlendi Bay Shipwrecks: An Archaeological Study*. Master's dissertation. University of Malta, L-Università ta' Malta.

Azzopardi, E., 2013. The Shipwrecks of Xlendi Bay, Gozo, Malta. *International Journal of Nautical Archaeology*, 42(2), pp. 286-295.

Barngrover, C., Kastner, R. and Belongie, S., 2014. Semisynthetic versus real-world sonar training data for the classification of mine-like objects. *IEEE Journal of Oceanic Engineering*, 40(1), pp. 48-56.

Bass, G.F., 1986. A Bronze Age Shipwreck at Ulu Burun (Kas): 1984 Campaign. *American Journal of Archaeology*, 90(3), pp. 269–296.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, A., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp. 82-115.

Bechtold, B., 2010. The pottery repertoire from late 6th-mid 2nd century BC Carthage: observations based on the big messaouda excavations. Ghent University, Department of Archaeology and Ancient History of Europe.

Bechtold, B. (2012). Amphorae Production in Punic Sicily (7th–3rd/2nd Centuries BCE): An Overview. FACEM (version 06/12/2012).

Bechtold, B., 2018. La distribuzione di anfore da trasporto maltesi fuori dall'arcipelago: nuovi dati. In: *The Lure of the Antique: Essays on Malta and Mediterranean Archaeology in Honour of Anthony Bonanno*. Peeters, Leuven, pp. 257-273.

Beltrán Lloris, M., 1970. *Las ánforas romanas en España*. Zaragoza: Institución "Fernando el Católico".

Ben Jerbania, I., 2017. La production des amphores ovoïdes de type «Africaine ancienne» à Utique. *Antiquités africaines: L'Afrique du Nord de la protohistoire à la conquête arabe*, 53, pp. 175-192.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021. On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 1 March 2021, New York, NY, USA. Association for Computing Machinery, pp. 610-623.

Benhabiles, H. and Tabia, H., 2017. Convolutional neural network for pottery retrieval. *Journal of Electronic Imaging*, 26(1), pp. 011005-011005.

Bickler, S.H., 2021. Machine learning arrives in archaeology. *Advances in Archaeological Practice*, 9(2), pp. 186-191.

Bochkovskiy, A., 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. Available at: https://arxiv.org/abs/2004.10934 [Accessed 18 January 2025].

Boulinguez, D. and Quinquis, A., 1999. Underwater buried object recognition using wavelet packets and Fourier descriptors. In: Proceedings of the 10th International Conference on Image Analysis and Processing, IEEE, pp. 478–483.

Brandsen, A. and Lippok, F., 2021. A burning question - using an intelligent grey literature search engine to change our views on early medieval burial practices in the Netherlands. Journal of Archaeological Science, 133, 105456.

Brown, C.J., Beaudoin, J., Brissette, M. and Gazzola, V., 2019. Multispectral multibeam echo sounder backscatter as a tool for improved seafloor characterization. Geosciences, 9(3), p.126.

Bruno, B. and Capelli, C., 2000. Nuovi tipi di anfore da trasporto a Malta. In: D'Amico, C. and Tampellini, C., eds. *6a Giornata: Le scienze della terra e l'archeometria*. Este: Grafica Atestina, pp. 59–65.

Bruno, B., 2004. *L'arcipelago maltese in età romana e bizantina: attività economiche e scambi al centro del Mediterraneo*. Bari: Edipuglia.

Buonamici, F., Carfagni, M., Furferi, R., Volpe, Y. and Governi, L., 2020. Generative design: an explorative study. *Computer-Aided Design and Applications,* 18(1), pp. 144-155.

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, pp. 36–43.

Camara, A., de Almeida, A., Caçador, D. and Oliveira, J., 2023. Automated methods for image detection of cultural heritage: Overviews and perspectives. *Archaeological Prospection*, 30(2), pp. 153-169.

Casilli, A., 2019. *En Attendant les Robots: Enquête sur le Travail du Clic*. Paris: Seuil.

Casini, L., Roccetti, M., Delnevo, G., Marchetti, N., and Orrù, V., 2021. The Barrier of meaning in archaeological data science. Department of Computer Science and Engineering, Universidad of Bologna.

Casini, L., Marchetti, N., Montanucci, A., Orrù, V. and Roccetti, M., 2023. A human-AI collaboration workflow for archaeological sites detection. *Scientific Reports*, 13(1), 8699.

Carmichael, D., 1998. Image processing techniques for the analysis of sidescan sonar survey data. IEEE Colloquium on Underwater Applications of Image Processing (Ref. No. 1998/217), pp. 1-7.

Caspari, G. and Crespo, P., 2019. Convolutional neural networks for archaeological site detection – Finding "princely" tombs. *Journal of Archaeological Science,* 110, 104998.

Cerri, S.A., Landini, P. and Leoncini, M., 1987. Cooperative agents for knowledge-based information systems: dialogue about the archeology of Rome. *Applied Artificial Intelligence: An International Journal*, 1(1), pp. 1-24.

Character, L., Ortiz Jr, A., Beach, T., and Luzzadder-Beach, S., 2021. Archaeologic machine learning for shipwreck detection using lidar and sonar. *Remote Sensing*, 13(9), 1759.

Chen, R.C., 2019. Automatic License Plate Recognition via sliding-window darknet-YOLO deep learning. *Image and Vision Computing,* 87, pp. 47-56.

Chen, L., Zhou, F., Wang, S., Dong, J., Li, N., Ma, H., et al., 2020. SWIPENET: Object detection in noisy underwater images. arXiv preprint arXiv:2010.10006. Available at: https://arxiv.org/abs/2010.10006 [Accessed 24 March 2025].

Chen, X.Q., Xia, K., Hu, W., Cao, M., Deng, K. and Fang, S., 2022. Extraction of underwater fragile artefacts: Research status and prospect. *Heritage Science,* 10(1), 9.

Cheng, L., Li, J., Duan, P. and Wang, M., 2021. A small attentional YOLO model for landslide detection from satellite remote sensing images. *Landslides*, 18(8), pp. 2751-2765.

Chen, J., Zhu, S. and Luo, W., 2024. Instance Segmentation of Underwater Images by Using Deep Learning. *Electronics*, 13(2), 274.

Chew, A.L., Tong, P.B. and Chia, C.S., 2007. Automatic detection and classification of man-made targets in side scan sonar images. In: *2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies*, pp. 126-132. IEEE.

Chollet, F., 2021. *Deep Learning with Python*. Simon and Schuster, Shelter Island, New York.

Clarke, D.L., (1968) 2014. Analytical archaeology. London and New York: Routledge.

Clavert, F. and Gensburger, S., 2023. Is artificial intelligence the future of collective memory? Bridging AI scholarship and Memory Studies. Call for Papers for the Second Volume of the Memory Studies Review, Brill Publishing, 2024.

Cobb, P., 2023. Large Language Models and Generative AI, Oh My!: Archaeology in the Time of ChatGPT, Midjourney, and Beyond. *Advances in Archaeological Practice*, 11, pp. 363-369.

Cortes, C. and Vapnik, V., 1995. Support vector machine. *Machine Learning*, 20(3), pp. 273-297

Cowgill, G., 1967. Computer applications in archaeology. *Computers in the Humanities*, 2, pp. 17–23.

Crawford, K., 2021. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press.

Crisp, J. and Watson, M., 2001. The application of object recognition systems for offshore oil rig inspection. *Journal of Offshore Technology*, 45(2), pp. 113-120.

Cutter, G., Stierhoff, K. and Zeng, J., 2015. Automated detection of rockfish in unconstrained underwater videos using Haar cascades and a new image dataset: Labeled fishes in the wild. In: 2015 IEEE Winter Applications and Computer Vision Workshops, pp. 57-62.

Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. *In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'05), vol. 1, pp. 886-893. IEEE.

Davis, D. (2020) 'Defining what we study: The contribution of machine automation in archaeological research', *Digital Applications in Archaeology and Cultural Heritage* **18**, e00152.

Daniel, S., Le Léannec, F., Roux, C., Soliman, B. and Maillard, E.P., (1998) Side-scan sonar image matching. *IEEE Journal of Oceanic Engineering*, 23(3), pp.245-259.

De Lucca, D., 1990. The built environment in Gozo: a historical review. In: C. Cini, ed. Gozo: *The Roots of an Island*, pp. 121–160. Malta.

Dewi, C., Chen, R.C., Jiang, X. and Yu, H., 2022. Deep convolutional neural network for enhancing traffic sign recognition developed on Yolo V4. *Multimedia Tools and Applications*, 81(26), pp. 37821-37845.

Diaz-Ramirez, V.H., Trujillo, L. and Pinto-Fernandez, S., 2012. Advances in adaptive composite filters for object recognition. In: *Advances in Object Recognition Systems,* pp.91-110.

Diodorus Siculus, Fischer, C.T., Vogel, F. and Dindorf, L. (eds.), 1985. Bibliotheca historica. Vol. 5. Bibliotheca scriptorum Graecorum et Romanorum Teubneriana. Wiesbaden: Vieweg+Teubner Verlag.

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

Doran, J., 1970. Systems theory, computer simulations and archaeology. *World Archaeology*, 1(3), pp.289-298.

Doran, J.E. and Hodson, F.R., 1975. *Mathematics and Computers in Archaeology.* Cambridge, MA: Harvard University Press.

Doran, J., 1977. Knowledge representation for archaeological inference. *Machine Intelligence*, 8, pp.433-454.

Doran, J., 1988. Expert systems and archaeology: what lies ahead? *Department of Computer Science,* University of Essex.

Dos Reis, D.H., Welfer, D., De Souza Leite Cuadros, M.A. and Gamarra, D.F.T., 2019. Mobile robot navigation using an object recognition software with RGBD images and the YOLO algorithm. *Applied Artificial Intelligence,* 33(14), pp.1290-1305.

Drap, P. and Long, L., 2001. Towards a digital excavation data management system: the "Grand Ribaud F" Etruscan deep-water wreck. *In: Proceedings of the 2001 Conference on Virtual Reality, Archaeology, and Cultural Heritage*, pp.17-26.

Drap, P., Merad, D., Hijazi, B., Gaoua, L., Nawaf, M.M., Saccone, M., Chemisky, B., Seinturier, J., Sourisseau, J.C., Gambin, T. and Castro, F., 2015. Underwater photogrammetry and object modeling: a case study of Xlendi Wreck in Malta. *Sensors,* 15(12), pp.30351-30384.

Drap, P., Papini, O., Merad, D., Pasquet, J., Royer, J.P., Motasem Nawaf, M., ... and Castro, F., 2019. Deepwater archaeological survey: an interdisciplinary and complex process. 3*D Recording and Interpretation for Maritime Archaeology,* pp.135-153.

Dressel, H., 1899. Über alte Funde von römischen Ton-Gefäßen mit Stempel-Aufschriften. Teubner Verlag, Wiesbaden.

Du, Y., Pan, N., Xu, Z., Deng, F., Shen, Y. and Kang, H., 2021. Pavement distress detection and classification based on YOLO network. *International Journal of Pavement Engineering*, 22(13), pp.1659-1672.

Ehrenreich, R.M., 1995. Archaeometry into archaeology. *Journal of Archaeological Method and Theory,* pp.1-6.

El Rejal, A.A., Pester, A. and Nagaty, K., 2023. Tiny machine learning for underwater image enhancement: pruning and quantization approach. In: *2023 International Conference on Computer and Applications* (ICCA), pp.1-6. IEEE.

Fan, J. and Li, H., 2021. On-demand maritime trade: a case study on the loading of cargo and the packaged goods of the Sinan shipwreck. *Journal of Maritime Archaeology*, 16, pp.163-186.

Fan, C.L., 2025. Evaluation Model for Crack Detection with Deep Learning: Improved Confusion Matrix Based on Linear Features. *Journal of Construction Engineering and Management*, 151(3), p.04024210.

Fayaz, S., Parah, S.A. and Qureshi, G.J., 2022. Underwater object detection: architectures and algorithms–a comprehensive review. *Multimedia Tools and Applications,* 81(15), pp.20871-20916.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D., 2009-10. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 32(9), pp.1627-1645.

Ferentinos, G., Fakiris, E., Christodoulou, D., Geraga, M., Dimas, X., Georgiou, N., Kordella, S., Papatheodorou, G., Prevenios, M. and Sotiropoulos, M., 2020. Optimal sidescan sonar and subbottom profiler surveying of ancient wrecks: The 'Fiskardo' wreck, Kefallinia Island, Ionian Sea. *Journal of Archaeological Science*, 113, p.105032.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A. and James, S., 2020. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters,* 133, pp.102-108.

Fisher, M., Fradley, M., Flohr, P., Rouhani, B. and Simi, F., 2021. Ethical considerations for remote sensing and open data in relation to the Endangered Archaeology in the Middle East and North Africa project. *Archaeological Prospection,* 28(3), pp.279-292.

Freed, J., 1996. Early Roman amphoras in the collection of the Museum of Carthage. *Echos du Monde Classique/Classical Views*, XL, pp.119-155.

Frendo, A.J., 1988. H.J. Franken's method of ceramic typology: an appreciation. *Palestine Exploration Quarterly,* 120(2), pp.108-129.

Freund, Y. and Schapire, R.E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In: *European Conference on Computational Learning Theory* (COLT), pp.23-37.

Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), pp.119-139.

Gal, D., Saaroni, H. and Cvikel, D., 2023. Mappings of potential sailing mobility in the Mediterranean during Antiquity. *Journal of Archaeological Method and Theory*, 30(2), pp.397-448.

Gallwey, J., Eyre, M., Tonkins, M. and Coggan, J., 2019. Bringing lunar LiDAR back down to Earth: Mapping our industrial heritage through deep transfer learning. *Remote Sensing*, 11(17), p.1994.

Gambin, T., 2005. The harbours of ancient Gozo.

Gambin, T. & Bonanno, A., 2006. Underwater Archaeological Heritage in Malta: Management and Conservation. *Journal of Maritime Archaeology*, 1(1), pp. 65-79.

Gambin, T., 2014. Managing Malta's Underwater Cultural Heritage. In*: S.A. Kingsley & G. Coupland, eds. Underwater Cultural Heritage in Europe.* Oxford: Archaeopress, pp. 78-89.

Gambin, T., 2015. A Phoenician shipwreck off Gozo, Malta.

Gambin, T., Drap, P., Chemisky, B., Hyttinen, K. & Kozak, G., 2018. Exploring the Phoenician shipwreck off Xlendi bay, Gozo: A report on methodologies used for the study of a deep-water site. *Underwater Technology,* 35(3), 71-86.

Gambin, T., 2020. Malta: Submerged landscapes and early navigation. In: G. Bailey, N. Galanidous, H.P. Jöns & M. Mennenga, eds. *The archaeology of Europe's drowned landscapes.* Springer, pp. 341-346.

Gambin, T. & Sauskemat, M., 2021. Tower Wreck Project. Report for 2021 Season, Heritage Malta, November 2021.

Gambin, T., Sourisseau, J.C. & Anastasi, M., 2021. The cargo of the Phoenician shipwreck off Xlendi Bay, Gozo: Analysis of the objects recovered between 2014–2017 and their historical contexts. *International Journal of Nautical Archaeology*, 50(1), pp. 3-18.

Gardin, J.C. & Laurière, J.L., 1987. Systèmes experts et sciences humaines: le cas de l'archéologie. Eyrolles, Paris.

Gattiglia, G., 2022. A postphenomenological perspective on digital and algorithmic archaeology. *Archeologia e Calcolatori*, 33(2), pp. 319-334.

Gauci, M & Grima, R. 1993. *The Sunday Times,* May 9th 1993.

Gehrig, J. & Becker, J., 2004. Artificial intelligence in oil and gas exploration: A new frontier for autonomous robots. In: *Proceedings of the 7th European Conference on Artificial Intelligence in Robotics,* pp. 233–240.

Gibbins, D. & Adams, J., 2001. Shipwrecks and maritime archaeology. *World Archaeology,* 32(3), pp. 279-291.

Girshick, R., Donahue, J., Darrell, T. & Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 580–587.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 7–13 December 2015, Santiago, Chile. IEEE, pp. 1440-1448.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. 2014. Generative adversarial networks. Advances in Neural Information Processing Systems, 27.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Gonzalez, P., 2024. Artificial Intelligence in the Field of Archaeology. Bachelor's dissertation. University of Gothenburg.

Grima, R., 1993. *Survey by Griffon off Gozo*. Unpublished report, National Museum of Archaeology, Malta.

Gualandi, M.L., Scopigno, R., Wolf, L., Richards, J., Buxeda i Garrigós, J., Heinzelmann, M., Hervás, M.A., Vila, L. and Zallocco, M., 2016. ArchAIDE - Archaeological Automatic Interpretation and Documentation of Ceramics. In: Eurographics Workshop on Graphics and Cultural Heritage, 2016, Eurographics Association, pp. 1-8.

Guyot, A., Lennon, M. and Hubert-Moy, L., 2021. Objective comparison of relief visualization techniques with deep CNN for archaeology. *Journal of Archaeological Science: Reports*, 38, p.103027.

Haenlein, M. & Kaplan, A., 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review,* 61(4), pp. 5-14.

Heenkenda, H.M.S.C.R. & Fernando, T.G.I., 2020. Approaches used to recognise and decipher ancient inscriptions: A review. *Vidyodaya Journal of Science*, 23(02).

Hein, I., Rojas-Domínguez, A., Ornelas, M., D'Ercole, G. & Peloschek, L., 2018. Automated classification of archaeological ceramic materials by means of texture measures. Journal of Archaeological Science: Reports, 21, pp. 921-928.

Horden, P. and Purcell, N., 2000. The Corrupting Sea: A Study of Mediterranean History. Oxford: Blackwell Publishers.

Hu, K., Weng, C., Zhang, Y., Jin, J. & Xia, Q., 2022. An overview of underwater vision enhancement: From traditional methods to recent deep learning. *Journal of Marine Science and Engineering, 10*(2), p. 241.

Huggett, J. & Baker, K., 1985. The computerised archaeologist: The development of expert systems. Science and Archaeology, (27), pp. 3-7.

Itkin, B., Wolf, L., & Dershowitz, N. (2019). Computational ceramicology. *arXiv preprint arXiv:*1911.09960. [Accessed 24 March 2025].

Jaklič, A., Erič, M., Mihajlović, I., Stopinšek, Ž. & Solina, F., 2015. Volumetric models from 3D point clouds: The case study of sarcophagi cargo from a 2nd/3rd century AD Roman shipwreck near Sutivan on island Brač, Croatia. *Journal of Archaeological Science,* 62, pp. 143-152.

Jamil, A.H., Yakub, F., Azizan, A., Roslan, S.A., Zaki, S.A. & Ahmad, S.A., 2022. A review on Deep Learning application for detection of archaeological structures. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 26(1), pp. 7-14.

Johnson, S.G. and Deaett, M.A., 1994. The application of automated recognition techniques to side-scan sonar imagery. *IEEE journal of Oceanic Engineering,* 19(1), pp.138-144.

Johri, P., Khatri, S.K., Al-Taani, A.T., Sabharwal, M., Suvanov, S. and Kumar, A., 2021. Natural language processing: History, evolution, application, and future work. In*: Proceedings of the 3rd International Conference on Computing Informatics and Networks*: ICCIN 2020, 2020, Singapore. Springer Singapore, pp. 365-375.

Jones, A., 2004. Archaeometry and materiality: materials-based analysis in theory and practice. *Archaeometry*, 46(3), pp. 327-338.

Kamal Al-anni, M. and Drap, P., 2024. Efficient 3D Instance Segmentation for Archaeological Sites Using 2D Object Detection and Tracking. *International Journal of Computing and Digital Systems*, 15(1), pp.1333-1342.

Kolar, Z., Chen, H., & Luo, X. 2018. Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. *Automation in Construction*, *89*, 58-70.

Kulik, S.D. and Shtanko, A.N., 2020. Experiments with neural net object detection system YOLO on small training datasets for intelligent robotics. In: Advanced Technologies in Robotics and Intelligent Systems: Proceedings of ITR 2019. *Springer International Publishing*, pp. 157-162.

Kumar, P., Narasimha Swamy, S., Kumar, P., Purohit, G. and Raju, K.S., 2021. Real-time, YOLO-based intelligent surveillance and monitoring system using jetson TX2. In: *Data analytics and management: proceedings of ICDAM*. Springer Singapore, pp. 461-471.

Kypraios, I. ed., (2012) Advances in Object Recognition Systems. BoD–Books on Demand Norderstedt, Germany.

Lagrange, M.S. and Renaud, M., 1985. Intelligent knowledge-based systems in archaeology: A computerized simulation of reasoning by means of an expert system. *Computers and the Humanities*, 19(1), pp. 37-52.

Lake, M. W., 2014. "Trends in Archaeological Simulation." *Journal of Archaeological Method and Theory*, 21(2), pp. 258-287.

de Lapérouse, J.F., 2020. Ceramic musealization: how ceramics are conserved and the implications for research. *Archaeological and Anthropological Sciences*, 12(8), p. 166.

LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp. 436-444.

Lei, F., Tang, F., Li, S., 2022. Underwater target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.*, 10, p. 310.

Lienhart, R. and Maydt, J., 2002. September. An extended set of Haar-like features for rapid object detection. In: *Proceedings. International conference on image processing*. Vol. 1, pp. I-I. IEEE.

Liritzis, I., Laskaris, N., Vafiadou, A., Karapanagiotis, I., Volonakis, P., Papageorgopoulou, C. and Bratitsi, M., 2020. Archaeometry: An Overview. *Scientific Culture*, 6(1).

Li, J., Su, Z., Geng, J., and Yin, Y., 2018. Real-time detection of steel strip surface defects based on improved YOLO detection network. *IFAC-PapersOnLine*, 51(21), pp. 76-81.

Lippi, M., Bonucci, N., Carpio, R.F., Contarini, M., Speranza, S. and Gasparri, A., 2021. June. A YOLO-based pest detection system for precision agriculture. In: *2021 29th Mediterranean Conference on Control and Automation (MED)*, pp. 342-347. IEEE.

Lowe, D.G., 1999. September. Object recognition from local scale-invariant features. In: *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2, pp. 1150-1157.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, pp. 91-110.

Mañá López, A., 1951. 'Las ánforas romanas en España', *Archivo Español de Arqueología*, 34, pp. 57-112.

Magnoni, A., Stanton, T.W., Barth, N., Fernandez-Diaz, J.C., León, J.F.O., Ruíz, F.P. and Wheeler, J.A., 2016. Detection thresholds of archaeological features in airborne LiDAR data from Central Yucatán. *Advances in Archaeological Practice*, 4(3), pp. 232-248.

Malfitana, D., 2008. Roman Sicily Project (RSP): Ceramics and Trade: A Multidisciplinary Approach to the Study of Material Culture Assemblages: First Overview: The Transport Amphorae Evidence. *Facta: A Journal of Roman Material Culture Studies*, 2, pp. 1000-1066.

Marshall, F. and Pilgram, T., 1993. NISP vs. MNI in quantification of body-part representation. *American Antiquity*, 58(2), pp. 261-269.

Masita, K.L., Hasan, A.N., & Shongwe, T., 2020. Deep learning in object detection: A review. In: *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1-11. IEEE.

McCarthy, J., 2007. *What is Artificial Intelligence?* Stanford University, Stanford, CA.

McCarthy, J., Benjamin, J., Winton, T. and Van Duivenvoorde, W., 2019. The rise of 3D in maritime archaeology. In: *3D Recording and Interpretation for Maritime Archaeology*, pp. 1-10.

Menna, F., Agrafiotis, P., & Georgopoulos, A., 2018. State of the art and applications in archaeological underwater 3D recording and mapping. *Journal of Cultural Heritage*, 33, pp. 231-248.

Minsky, M. and Papert, S., 1969. An introduction to computational geometry. *Cambridge Tiass.*, HIT, 479(480), p. 104.

Mitchell, M., 2009. *Complexity: A Guided Tour*. Oxford University Press.

Molua, C.O., 2024. Advanced Image Processing for Archaeological Site Identification, Management, and Conservation.

Moniruzzaman, M., Islam, S.M.S., Bennamoun, M., & Lavery, P., 2017. Deep learning on underwater marine object detection: A survey. In: *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017*, Antwerp, Belgium, September 18-21, 2017, Proceedings 18, pp. 150-160. Springer International Publishing.

Muckelroy, K., 1978. *Maritime Archaeology*. Cambridge: Cambridge University Press.

Murray, L., 2018. Shipwreck capital of the world. *Engineering & Technology*, 13(7/8), pp. 54-58.

Nau, A., 1983. Expert computer systems. *Computer*, 16(2), pp. 63-85.

Nayak, N., Nara, M., Gambin, T., Wood, Z., and Clark, C.M., 2021. Machine learning techniques for AUV side-scan sonar data feature extraction as applied to intelligent search for underwater archaeological sites. In: *Field and Service Robotics: Results of the 12th International Conference*, pp. 219-233. Springer Singapore.

Neubeck, A. and Van Gool, L., 2006. Efficient non-maximum suppression. In: *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3, pp. 850-855. IEEE.

Nie, Y., Sommella, P., O'Nils, M., Liguori, C. and Lundgren, J. 2019 'Automatic detection of melanoma with YOLO deep convolutional neural networks', in *2019 E-Health and Bioengineering Conference (EHB)*, November, pp. 1–4. IEEE.

Orengo, H.A. and Garcia-Molsosa, A. 2019. 'A brave new world for archaeological survey: automated machine learning-based potsherd detection using high-resolution drone imagery', *Journal of Archaeological Science*, **112**.

Orengo, H.A., Conesa, F.C., Garcia-Molsosa, A., Lobo, A., Green, A.S., Madella, M. and Petrie, C.A. 2020. 'Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data', *Proceedings of the National Academy of Sciences*, **117**(31), pp. 18240–18250.

Orton, C., Tyers, P. and Vince, A. 1993. *Pottery in Archaeology*. Cambridge: Cambridge University Press.

Paraskevas, K., Mariolis, I., Giouvanis, G., Peleka, G., Zampokas, G. and Tzovaras, D. 2023. 'Underwater detection of ancient pottery sherds using deep learning', *International Journal on Cybernetics & Informatics (IJCI)*, **12**(12), p.1.

Paraskevas, K. and Kavallieratou, E. 2023. 'Detecting holes in fish farming nets: A two–method approach', in *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*, pp. 1–7. IEEE.

Pasquet, J., Demesticha, S., Skarlatos, D., Merad, D. and Drap, P. 2017. 'Amphora detection based on a gradient weighted error in a convolution neuronal network', in *IMEKO International Conference on Metrology for Archaeology and Cultural Heritage (MetroArchaeo 2017)*.

Peacock, D. 1977. 'Roman amphorae: typology, fabric and origins', *Publications de l'École Française de Rome*, **32**(1), pp. 261–278.

Peacock, D.P.S. and Williams, D.F. 1986. *Amphorae and the Roman Economy: An Introductory Guide*. London: Longman.

Pham, M.T., Courtrai, L., Friguet, C., Lefèvre, S. and Baussard, A. 2020. 'YOLO-Fine: One-stage detector of small objects under various backgrounds in remote sensing images', *Remote Sensing*, **12**(15), p. 2501.

Pitts, W. and McCulloch, W.S. 1947. 'How we know universals: the perception of auditory and visual forms', *The Bulletin of Mathematical Biophysics*, **9**, pp. 127–147.

Pulak, C. 1988. 'The Bronze Age Shipwreck at Ulu Burun, Turkey: 1985 Campaign', *American Journal of Archaeology*, **92**(1), pp. 1–37.

Pulak, C. 1998. 'The Uluburun Shipwreck: An Overview', *The International Journal of Nautical Archaeology*, **27**(3), pp. 188–224.

Raavi, S., Chandu, P.B. and SudalaiMuthu, T. 2023. 'Automated recognition of underwater objects using deep learning', in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1055–1059. IEEE.

Radić Rossi, I., Casabán, J., Yamafune, K., Torres, R. and Batur, K. 2019. 'Systematic photogrammetric recording of the Gnalić shipwreck hull remains and artefacts', *3D Recording and Interpretation for Maritime Archaeology*, pp. 45–65.

Ramon Torres, J. 1995. *Las ánforas fenicio-púnicas del Mediterráneo central y occidental*. Aix-en-Provence: Barcelona, Spain.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. 2016. 'You only look once: Unified, real-time object detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788.

Redmon, J. and Farhadi, A. 2017. 'YOLO9000: Better, faster, stronger', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263–7271.

Ren, S., He, K., Girshick, R. and Sun, J. 2016. 'Faster R-CNN: Towards real-time object detection with region proposal networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(6), pp. 1137–1149.

Ren, X.M., Qi, Y.R., Zhang, D.Y. and He, Z.Z. 1996. 'Photoacoustic detection of the underwater object', *Progress in Natural Science-Beijing*, **6**, pp. S-760.

Renfrew, C. and Bahn, P.G. 2011. *Arqueología: Teoría, métodos y prácticas*. España: Ediciones Akal.

Rice, P.M. 1987. *Pottery Analysis: A Sourcebook*. Chicago: University of Chicago Press.

Ringle, W.M., Gallareta Negrón, T., May Ciau, R., Seligson, K.E., Fernandez-Diaz, J.C. and Ortegón Zapata, D. 2021. 'Lidar survey of ancient Maya settlement in the Puuc region of Yucatán, Mexico', *PLOS ONE*, **16**(4), p. e0249314.

Rosenblatt, F. 1958. 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review*, **65**(6), p. 386.

Rousaki, A., Moens, L. and Vandenabeele, P. 2018. 'Archaeological investigations (archaeometry)', *Physical Sciences Reviews*, **3**(9), p. 20170048.

Roy, A.M., Bhaduri, J., Kumar, T. and Raj, K. 2023. 'WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection', *Ecological Informatics*, **75**, p. 101919.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. 1986. 'Learning representations by back-propagating errors', *Nature*, **323**(6088), pp. 533–536.

Sagona, C. 2002. *The Archaeology of Punic Malta (Vol. 9)*. Peeters.

Schweizer, P.F. and Petlevich, W.J. 1989. 'Automatic target detection and cuing system for an autonomous underwater vehicle (AUV)', in *Proceedings of the 6th International Symposium on Unmanned Untethered Submersible Technology*, June, pp. 359–371. IEEE.

Schweizer, P.F., Petlevich, W.J., Haley, P.H. and Oravec, J.J. 1994. 'Image processing architecture for AUV mine-hunters', *Sea Technology*, **33**(4), pp. 55–61.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y. 2013. 'Overfeat: Integrated recognition, localization and detection using convolutional networks.'

Shin, F.B., Kil, D.H. and Dobeck, G.J. 1997. 'Integrated approach to bandwidth reduction and mine detection in shallow water with reduced-dimension image compression and automatic target recognition algorithms.' In *Detection and Remediation Technologies for Mines and Minelike Targets II*, Vol. 3079, pp. 203-212. SPIE.

Shinde, S., Kothari, A. and Gupta, V. 2018. 'YOLO based human action recognition and localization.' *Procedia Computer Science*, **133**, pp. 831-838.

Shortis, M. 2019. 'Camera calibration techniques for accurate measurement underwater.' In *3D Recording and Interpretation for Maritime Archaeology*, pp. 11-27.

Spahn, C., Gómez-de-Mariscal, E., Laine, R.F. et al. 2022. 'DeepBacs for multi-task bacterial image analysis using open-source deep learning approaches.' *Communications Biology*, **5**, p. 688.

Song, P., Li, P., Dai, L., Wang, T. and Chen, Z. 2023. 'Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection.' *Neurocomputing*, **530**, pp. 150-164.

Spennemann, D.H. 2024. 'Generative artificial intelligence, human agency and the future of cultural heritage.' *SSRN Electronic Journal.*

Sukkham, A. 2021. 'Ceramic assemblages from shipwrecks in Southeast Asia from the last half of the eighteenth to the early twentieth centuries.' *Journal of Maritime Archaeology*, **16**, pp. 155–186.

Sun, H., Liu, M., Li, L., Yan, L., Zhou, Y. and Feng, X. 2020. 'A new classification method of ancient Chinese ceramics based on machine learning and component analysis.' *Ceramics International*, **46**(6), pp. 8104-8110.

Tan, C.F., Wahidin, L.S., Khalil, S.N., Tamaldin, N., Hu, J. and Rauterberg, G.W.M. 2016. 'The application of expert system: A review of research and applications.' *ARPN Journal of Engineering and Applied Sciences*, **11**(4), pp. 2448-2453.

Tan, L., Huangfu, T., Wu, L. and Chen, W. 2021. 'Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identification.' *BMC Medical Informatics and Decision Making*, **21**, pp. 1-11.

Tchernia, A. 1986. *Le vin de l'Italie romaine. Essai d'histoire économique d'après les amphores.* Persée-Portail des revues scientifiques en SHS.

Terven, J. and Cordova-Esparza, D. 2023. 'A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond.' *arXiv preprint arXiv:2304.00501.* [Accessed 18 January 2025].

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E. and Liang, Z. 2019. 'Apple detection during different growth stages in orchards using the improved YOLO-V3 model.' *Computers and Electronics in Agriculture*, **157**, pp. 417-426.

Tomasella, N., Flenghi, G. and Rosati, L. 2024. 'Between Image and Text: Automatic Image Processing for Character Recognition in Historical Inscriptions.' In *Advances in Representation: New AI-and XR-Driven Transdisciplinarity*, pp. 93-106. Cham: Springer Nature Switzerland.

Tuomi, I. 2019. *The Impact of Artificial Intelligence on Learning, Teaching, and Education: Policies for the Future.* JRC Science for Policy Report. European Commission.

Turing, A. 1950. 'Computing machinery and intelligence.' *Mind*, **49**, pp. 433–460.

Tyukin, I., Sofeikov, K., Levesley, J., Gorban, A.N., Allison, P. and Cooper, N.J. 2018. 'Exploring automated pottery identification [Arch-I-Scan].' *Internet Archaeology*, **50**. Available at: https://intarch.ac.uk/journal/issue50/11/index.html.

Twede, D. 2002. 'Commercial amphoras: The earliest consumer packages?.' *Journal of Macromarketing*, **22**(1), pp. 98-108.

Ünver, H.M. and Ayan, E. 2019. 'Skin lesion segmentation in dermoscopic images with combination of YOLO and grabcut algorithm.' *Diagnostics*, **9**(3), p. 72.

Vandermersch, C. 1994. *Vins et amphores de Grande Grèce et de Sicile.* Naples: Publications du Centre Jean Bérard.

Vasilev, I. 2019. *Advanced Deep Learning with Python: Design and Implement Advanced Next-Generation AI Solutions Using TensorFlow and PyTorch.* 1st edn. Birmingham: Packt Publishing, Limited.

Vella, N.C. 2002. 'The lie of the land: Ptolemy's temple of Hercules in Malta.' *Ancient Near Eastern Studies*, **XXXIX**, pp. 83–112.

Vella, G. 2006. 'Securing Gozo's ancient gateways.' *The Sunday Times*, 28 May, p. 59. London.

Verschoof-van der Vaart, W.B., Lambers, K., Kowalczyk, W. and Bourgeois, Q.P. 2020. 'Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from the Netherlands.' *ISPRS International Journal of Geo-Information*, **9**(5), p. 293.

Vitali, V. 1989. 'Archaeometric provenance studies: An expert system approach.' *Journal of Archaeological Science*, **16**(4), pp. 383-391.

Vidal Gonzalez, P. 1996. *La Isla de Malta en Época Fenicia y Púnica.* BAR International Series 653. Tempus Reparatum, Oxford.

Wang, C.C., Samani, H. and Yang, C.Y. 2019. 'Object detection with deep learning for underwater environment.' In *2019 4th International Conference on Information Technology Research (ICITR)*, pp. 1-6. IEEE.

Wen, X., Yuling, W. and Weiqing, Z. 1995. 'Sonar image processing system for an autonomous underwater vehicle (AUV).' In *Challenges of Our Changing Global Environment: Conference Proceedings. OCEANS '95 MTS/IEEE*, Vol. 3, pp. 1883-1886. IEEE.

Will, E.L. 1982. 'Greco-italic amphoras.' *Hesperia: The Journal of the American School of Classical Studies at Athens*, **51**(3), pp. 338-356.

Will, E. 1987. 'The Roman amphoras.' In *The Roman Port and Fishery of Cosa, a Centre of Ancient Trade*, edited by A.M. McCann, pp. 171-222. Princeton: Princeton University Press.

Willett, M., 2019. *A Computational Approach to Cultural Resource Management: Autodetecting Archaeological Features in Satellite Imagery with Convolutional Neural Networks.* University of California Press, Berkeley, CA.

Williams, D.F., 1986. Amphorae and the Roman economy: An introductory guide. Longman.

Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. 2018. CBAM: Convolutional Block Attention Module. European Conference on Computer Vision (ECCV), pp. 3–19.

Woods, J.D., 1962. The Malta expedition, 1961. *Exploration Review*, February 1962, Imperial College Exploration Society, pp. 11-36.

Wu, D., Lv, S., Jiang, M. and Song, H., 2020. Using channel pruning-based YOLOv4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Computers and Electronics in Agriculture*, 178, p.105742.

Yan, J., Zhou, Z., Zhou, D., Su, B., Xuanyuan, Z., Tang, J. and Liang, W., 2022. Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling. *Frontiers in Marine Science*, 9, 1056300.

Yang, W. and Jiachun, Z., 2018. Real-time face detection based on YOLO. In *2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII)*, pp. 221-224. IEEE.

Yang, Y., Liang, W., Zhou, D., Zhang, Y. and Xu, G., 2023. Object detection for underwater cultural artefacts based on deep aggregation network with deformation convolution. *Journal of Marine Science and Engineering*, 11(12), 2228.

Yardımcı, E., 2008. Uluburun Shipwreck: An overview of the underwater archaeological survey and excavations of the shipwreck. In: *Proceedings of the 9th International Symposium on Boat and Ship Archaeology*, 9th ed., Koç University Press, Istanbul, Turkey, pp. 55-66.

Zammit, E., Seychell, D., Debono, C.J., Gambin, T. and Wood, J., 2024. Underwater archaeological object detection through bidirectional photogrammetric fusion. In *2024 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, pp. 4122-4126. IEEE.

Zeng, L., Sun, B. and Zhu, D., 2021. Underwater object detection based on Faster R-CNN and adversarial occlusion network. *Engineering Applications of Artificial Intelligence*, 100, p.104190.

Zerr, B., 1991. Automatic image comparison using artificial intelligence. *Ocean Space Advanced Technology European Show*.

Zhang, W., Zhuang, P., Sun, H.H., Li, G., Kwong, S. and Li, C., 2022. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Transactions on Image Processing*, 31, pp. 3997-4010.

Zhang, C., Zhang, G., Li, H., Liu, H., Tan, J. and Xue, X., 2023. Underwater target detection algorithm based on improved YOLOv4 with SemiDSConv and FIoU loss function. *Frontiers in Marine Science*, 10, 1153416.

Zhang, M., Gao, F., Yang, W. and Zhang, H., 2023. Wildlife object detection method applying segmentation gradient flow and feature dimensionality reduction. *Electronics*, 12(2), 377.

Zhao, Z.Q., Zheng, P., Xu, S.T. and Wu, X., 2019. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), pp. 3212-3232.

Zheng, Y. and Zhang, H., 2022. Video analysis in sports by lightweight object detection network under the background of sports industry development. *Computational Intelligence and Neuroscience*, 2022(1), p.3844770.

Zhu, B., Wang, X., Chu, Z., Yang, Y. and Shi, J., 2019. Active learning for recognition of shipwreck target in side-scan sonar image. *Remote Sensing*, 11(3), 243.

Appendix I

Automated Object Detection. Background in relation to Archaeological Science and Operations

Contents

1 Background

On the practical side, the idea of AI was first developed by Alan Turing in the 1930s and 1940s. Turing, working as a leading cryptanalyst at the Government Code and Cypher School in Bletchley Park, Buckinghamshire, England, described an abstract computing machine that featured limitless memory and a scanner capable of moving back and forth, reading symbols and progressing the script. These ideas laid the foundation for modern ML[130] and led him to write his seminal article in 1950. In *Computer Machinery and Intelligence*, Turing outlined how to create intelligent machines and introduced the Turing Test, which established that if a human cannot distinguish between interacting with another human and a machine, the machine can be considered intelligent (Turing, 1950).

The term "artificial intelligence" itself was coined six years later, in 1956, by computer scientists Marvin Minsky and John McCarthy, who hosted the eight-week Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) at Dartmouth College in New Hampshire, US (Haenlein and Kaplan, 2019: 7). This workshop, convened ten years after the end of World War II, aimed to unite researchers from various fields to create a new area of study that would fulfill Turing's vision and build machines capable of simulating human intelligence. From this point forward, the field of AI experienced multiple rises and falls in interest and development, driven by

[130] Machine learning (p.1).

the efficiency of new hardware and software and the evolving concept of what artificial intelligence should encompass (Figure 44).

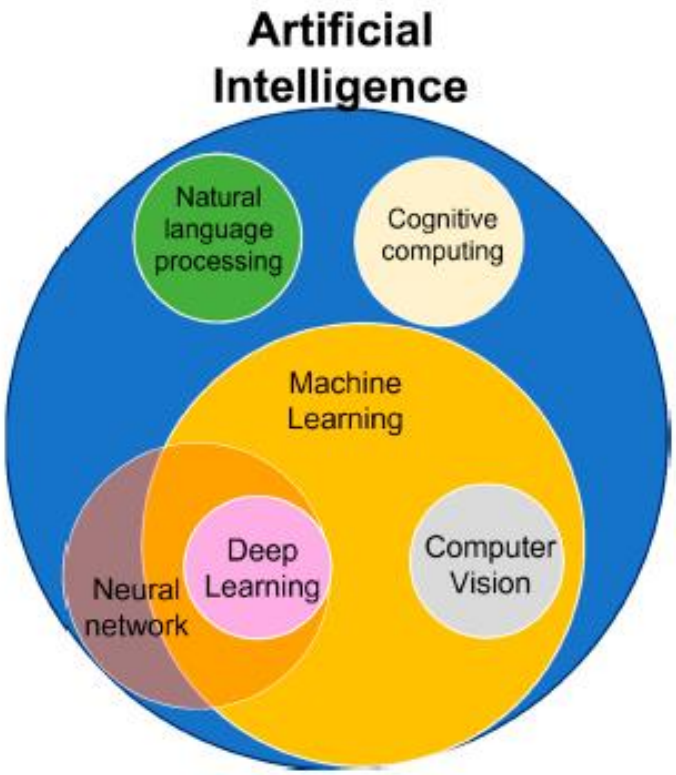


Figure 44. Key components of AI (in Argyrou, A. and Agapiou, A., 2022: 13).

1.1 Expert Systems in Archaeology

Following the DSRPAI and the ensuing rush of AI research, various avenues for simulating true human intelligence began to emerge (Haenlein and Kaplan, 2019: 8). One particular approach gained immediate success and significant media attention during the 1970s: symbolic AI, which appeared in the form of expert systems. Still in use today, expert or knowledge-based systems are problem-solving and decision-making programs that simulate the performance of human experts within a specific domain and without relying on complex procedural code (Doran, 1988: 237; Vitali, 1989: 385; Tan et al., 2016: 2448).[131]

[131] **Expert systems** are programs that work by functioning within a rigid framework of three key features: a knowledge base, an inference engine, and a user interface. The knowledge base consists of information or rules structured using expert knowledge on a specific problem within a domain. The inference engine applies formalized rules (such as "if-then" statements) and reasoning techniques to derive conclusions and decisions from the database (Nau, 1983: 63; Tan et al., 2016: 2448; Haenlein and Kaplan, 2019: 8). The user interface allows users to interact with the system by

In archaeology, variations of these expert systems began to emerge around 1970 (Clarke, 1968; Doran, 1970, 1977, 1988; Doran and Hudson, 1975; Lagrange and Renaud, 1985; Gardin et al., 1987; Cerri et al., 1987). Their functions were thoroughly studied by Cowgill (1967), who demonstrated how computers and expert systems could be used as database management tools. These tools served a dual purpose: first, replicating human decision-making abilities to retrieve and analyze vast amounts of information quickly; and second, contributing to simulations that helped archaeologists understand and analyze ancient societies' cultural and behavioral patterns. These early approaches reached their zenith with Vanda Vitali's archaeometry provenance studies (Vitali, 1989). Vitali's work investigated the role of archaeometric information in the interpretive phase of past reconstruction, using expert systems to assess their value in archaeological research and warning against confining computer science to technical departments without archaeologists' involvement in the model-training process (Vitali, 1987: 389).

Despite their impressive performance in rigid scenarios, expert systems failed to perform in more flexible or less structured situations. They were unable to distinguish between images of people and vehicles, could not adapt or learn from their mistakes, and lacked the flexibility and adaptability of the human brain—they did not fulfill the true vision of AI that Turing had proposed.

By the 1980s, these limitations led to widespread dissatisfaction, reduced funding, and discontinuation of AI research for several decades on a period often referred to as the "winter of AI." In archaeological science, this situation was illustrated by Doran's 1988 publication, *Expert Systems and Archaeology: What Lies Ahead?,* in which the author expressed a pessimistic view regarding the future of expert systems within the discipline. Soon after, AI-based methodologies would fall into obscurity again, regarded as raw and ineffective tools for archaeological research.

---

inputting queries and receiving solutions. Together, these components allow expert systems to use non-numerical knowledge to reason through problems, explain their reasoning, identify missing or contradictory information, and request further details (Huggett and Baker, 1985).

One well-known example of such a system is IBM's Deep Blue, developed in 1989. Deep Blue was a chess-playing expert system capable of processing 200 million possible moves per second and optimizing its next move by looking 20 turns ahead through a tree of "if-then" propositions. It famously defeated chess world champion Garry Kasparov, with the system's moves executed on a physical board by a human following the computer's instructions (user interface) after receiving each of Kasparov's moves as input. Other notable expert systems include the General Problem Solver, developed in 1957 by Allen Newell and Herbert A. Simon to solve simple problems like the Towers of Hanoi, and the ELIZA program, created in 1966 by Joseph Weizenbaum as a language processing tool capable of conversing with humans.

1.2 Automated Neural Networks

From the beginning, other approaches to achieve the creation of true AI were being discussed alongside expert systems. Of these, the product of the concept of Automated Neural Networks (ANN) would eventually be successful as the basis of most modern AI-based technologies. The origins of this method can be traced back to Canadian psychologist Donald Hebb's 1943 theory of Hebbian Learning.[132] In essence, ANN are a type of AI system inspired by the human brain and designed to recognize patterns and predict and process complex data through a frame of interconnected nodes that act as neurons.

Despite these advances, however, the study of neural networks seemingly stalled as well shortly after their creation. In 1969, their limitations were put into perspective when it was shown that computers of the time lacked the processing power necessary to make proper use of them (Minsky and Papert, 1969). This period of disinterest and lack of innovation lasted more than 15 years. It was not until 1986, with the invention of backpropagation, that a new era of research and innovation began.[133]

Following the advent of backpropagation and advancements in both software and hardware, numerous breakthroughs in neutral networks during the 1990s and early 2000s eventually led to the rise of DL[134] in 2012 (Figure 45). At this point, neural networks had evolved from simple neuron models to the sophisticated systems driving many of modern AI technologies (Figure 44).

[132] In his theory, Hebb proposed the idea of recreating the functioning of human neurons. That same year, Warren McCulloch and Walter Pitts published the first model of a neuron, expressing, through mathematical means, a rudimentary version of how neurons process information in biological systems. This model was expanded and put into practical use in 1958 by Frank Rosenblatt with his introduction of the *Perceptron*—an algorithm designed for supervised learning of binary classifiers. The Perceptron functioned as a simple computational unit modeled after a neuron's structure. It could take one or more inputs of information and produce a single output (Haenlein and Kaplan, 2019).

[133] **Backpropagation** is a ML technique essential to the optimization of training multi-layer neural networks by minimizing their errors (Rumelhart et al., 1986: 534).

[134] Deep learning (p.11).

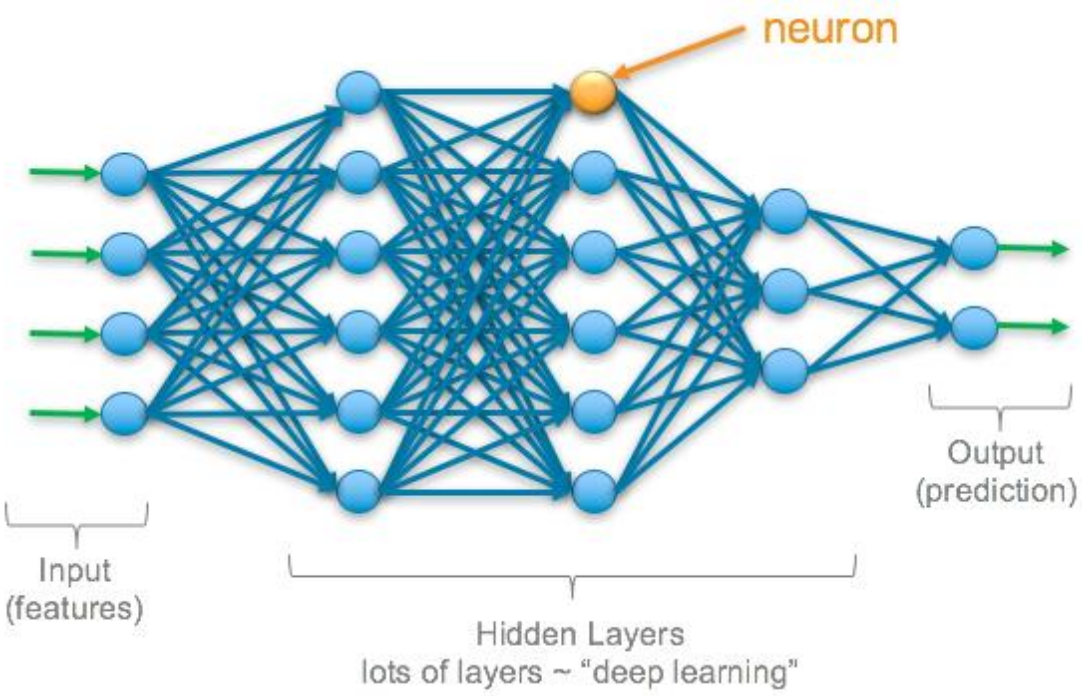


Figure 45. Diagram of the workings of DL. From: Ronaghan, S. (2018) Deep learning. [Online image]. Available at: https://srnghn.medium.com/deep-learning-overview-of-neurons-and-activation-functions-1d98286cf1e4 (Accessed November 5, 2024).

1.3 Modern Artificial Intelligence

But what exactly is modern artificial intelligence?

If you ask this question on the street, you are likely to hear practical, everyday examples of its use, such as YouTube's recommendation algorithm, smartphone applications, or the latest generation of robots. If you ask experts in the field, their answer will be more technical, explaining how neural networks are structured in layers that receive, organize, and classify data by assigning weights and thresholds. A professor, when explaining it, might frame it this way: "AI is the attempt to understand and build intelligent entities."

The reality is that the answer to this question will vary depending on who you ask, and most likely, each response will be correct.

Each definition of artificial intelligence sets a different context for how it should be understood, measured, and governed. In her *Atlas of Artificial Intelligence*, Crawford (2021: 8-9) introduces

AI as an ambiguous, political concept that she argues cannot be fully understood from a narrow technical perspective. She also acknowledges why this narrow viewpoint is often the chosen path.

As archaeologists looking to apply AI-based systems, we need to be aware of the ethical challenges and other implications of using AI. However, our primary goal is not to study AI itself but to understand how its technologies can benefit archaeological research and how we can leverage them in our field. Consequently, for us, a basic definition will suffice: AI refers to the collection of systems and software that can perform tasks typically associated with human intelligence.

These systems are categorized based on their capabilities, with narrow AI being the most basic. Narrow AI is designed to perform specific tasks only. Everyday examples include voice assistants like Siri or Alexa, translation tools, video game non-playable characters (NPCs), and recommendation systems used in streaming services and web browsers. These programs are part of larger AI technologies that are increasingly being applied across industries, transforming them by automating, personalizing, and scaling various repetitive processes. Some of the largest groups or technological packages within AI are shown in Figure 44. They include natural language processing (NLP),[135] generative AI,[136] autonomous systems, robotic process automation (RPA), machine learning (ML), and computer vision. The latter two, for instance, are most commonly used in maritime archaeology through automated object detection.

2. Automated Object Detection Models

To understand an image is not merely to be able to tell the difference between different images, but to accurately estimate the characteristics and locations of the objects contained within them (Zhao et al., 2019: 3212). This process is known as object detection. As one of the foundational topics in computer vision, object detection plays a vital role in analyzing all types of images and videos. However, while the concept may seem straightforward, numerous challenges hinder the creation of a single system that can flawlessly understand images in every possible scenario (Girshick et al., 2014: 582; Redmon et al., 2016: 780).

[135] **Natural language processing** is a package or field within AI focused on enabling computers to understand, interpret and generate human language (Johri et al., 2021: 366).
[136] **Generative AI** are those DL models able to generate data based on their training parameters. This data could be text, sound, video, or all sorts of other kind of content (Buonamici et al., 2020: 144, Spennemann, 2024: 3601).

There are many detection models available, and while each employs different methodologies, they all ultimately address two core tasks: determining the location of objects within an image and classifying each object into a specific category (Diaz-Ramirez et al., 2012: 93). Traditional detection models typically follow a three-stage pipeline,[137] wherein each step performs a distinct operation, with the output of one stage serving as the input for the next. This chain-like structure can lead to cumulative processing time across stages. Although some models have streamlined this process into a two-stage approach (Bochkovskiy et al., 2020: 2), all detection models must still complete three essential tasks to generate automatic predictions from images. These are informative region selection, feature extraction, and classification (Zhao et al., 2019: 3212).

-Informative region selection: Since objects of different sizes can appear anywhere in an image, the model's algorithm must scan the entire image. One of the most common approaches for this is the computationally expensive multiscale sliding window technique. This method systematically moves windows of various sizes across the image, performing computations at each step (Redmon et al., 2016: 780-781).

-Feature extraction: This step involves the algorithm extracting visual features that provide a semantic and robust interpretation of the objects in the image (Lowe, 2004: 93-95). The challenge lies in the variation in object appearances, lighting conditions, and backgrounds, as it is difficult to develop a feature extraction descriptor that works equally well across all possible objects (Zhao et al., 2019: 3212). Some examples of feature extractors that appear across different detection models include Scale-Invariant Feature Transform algorithms (SIFT)[138] (Lowe, 2004), Histograms of Oriented Gradients (HOG)[139] (Dalal and Triggs, 2005), and Haar-like features (Lienhart and Maydt, 2002).[140]

-Classification: The final step is classification, where the algorithm categorizes detected objects, making the representation of these objects more informative. Common classifiers include Support

[137] The term **pipeline** in this case refers to the structured sequence of steps through which data flows and executed within a program (Redmon et al., 2016:779).
[138] **SIFT** repeatedly blurs and downsizes images to create different scales, making it useful for resized images that display variations in lighting and color contrast (Lowe, 2004).
[139] **HOG** works by systematically dividing images into small cells like YOLO, but analyzes each cell separately after grouping sets together in defined blocks instead of all at once like YOLO does (Dalal and Triggs, 2005).
[140] **Haar-like features** analyze the images using rectangular windows to detect different edges, lines and textures. It has low computational time and is commonly used in face recognition (Lienhart and Maydt, 2002).

Vector Machines (SVM) [141] (Cortes and Vapnik, 1995), the Deformable Parts Model (DPM) [142] (Felzenszwalb et al., 2009), and AdaBoost (Freund and Schapire, 1997). [143]

Regarding all of these, it is worth nothing that many classifiers work better with specific feature extractors, and vice versa.

Nevertheless, it is not crucial for us to understand exactly how each of these tasks is accomplished. Instead, we have to recognize that for an algorithm to successfully detect objects, each part of the process must handle the unique challenges posed by each image. For underwater images, in particular, we must consider the inherent difficulties introduced by the maritime environment discussed in Chapter 2 (e.g., dataset size and variety, image distortion, object scaling, occlusion, and overlapping).[144]

According to the previous overview, given how the algorithm's structure works, it is clear how difficult, and perhaps even pointless, it would be to develop a detection algorithm that functions seamlessly across every underwater archaeological site. For instance, an algorithm might struggle with blurry photos during the classification task but perform well in other stages, or vice versa. This scenario helps explain the publication trends in the early 21st century, when experts in maritime disciplines began presenting different detection methods, each making slight adjustments to various points in the pipeline to address the specific challenges of the underwater environment. It also underlines the advantages of reducing the number of stages in a detection pipeline. Finally, the situation emphasizes the value of the YOLO model, which we will explore further shortly.

2.1 Operation and architecture

In Chapter 2 we saw why YOLO models, among all other families of detection models, have emerged as the model of choice for the application of detection in archaeological science.[145] Here,

[141] **SVMs** are a type of supervised learning algorithm used for classification and regression tasks particularly useful in high-dimensional datasets. They works by finding the optimal hyperplane to separate data points belonging to different classes (Cortes and Vapnik, 1995).
[142] **DPMs** are detection mechanisms that represent objects as dynamic collections of parts around a reference central position. Its advantages lay in their ability to handle variations of pose, which leads to more robust models able to handle partial occlusion and articulated elements (Felzenswalb et al., 2010).
[143] **AdaBoost** is a boosting algorithm that improves classifying performance. It increases model accuracy and reduces overfitting issues in small datasets (Freund and Schapire, 1995:28).
[144] Difficulties introduced by the maritime environment (p.8).
[145] Different detection models (p.10-13).

we will explain the operational processes of the model, which explain the reasons behind it being so simple, fast, accurate, efficient and versatile as it is.[146]

Since 2017, the YOLO family has released more than seventeen versions, each with its own features. One key difference between all YOLO models and other detection algorithms is in the way they process images. Traditional object detection models, like DPM and R-CNN,[147] classify objects and draw bounding boxes around detected objects in images. However, these models use separate stages in their pipeline, which slows down the process and complicates optimization, particularly when applied to new domains. These models typically process images regionally.[148] YOLO, on the other hand, was introduced in 2016 (Redmon et al., 2016) as a fast, unified object detector that considers the entire image at once during both the training and detection process. Hence its name: You Only Look Once.

Instead of analyzing images in sections or regions like most models (using sliding windows or region-proposal methods), YOLO evaluates the entire image at the same time. It predicts bounding boxes for all classes simultaneously, saving time while maintaining high average precision (Redmon et al., 2016: 780). In practical terms, the model achieves this by dividing each image into an S x S grid, where each cell is responsible for detecting objects whose center falls within them. The number of grid cells is adjustable, with a larger number resulting in higher computational demand.[149] Each grid cell predicts B bounding boxes, along with confidence scores for the likelihood of these boxes containing an object (Figure 46). Moreover, YOLO combines predictions from multiple feature maps at different resolutions, which helps it effectively detect objects of varying sizes. This feature makes YOLO highly efficient in a wide range of contexts.

Apart from the key difference constituted by this way of looking at 2D data, YOLO models also

[146] For additional information on the technical aspects of YOLO detection and its versions, check: Redmon et al., 2016; Redcmon and Farhadi, 2017; Bochkovskiy et al., 2020 and Terven et al., 2023.
[147] R-CNN (p.11).
[148] Informative region selection (p.159).
[149] The number of cells corresponds with model size A larger sized YOLO architecture means the grid is divided in more and more cells—increasing precision and ability to handle clustered objects at the cost of speed. More information on the influence of model size can be found in p.62.

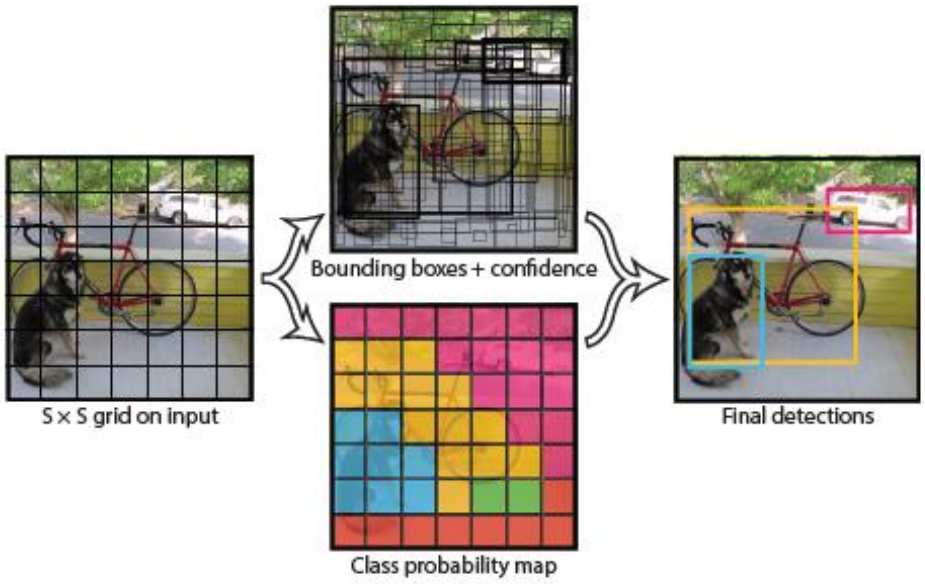


Figure 46. The system turns detection into a regression problem. It divides the image into a *S* x *S* grid and for each cell it predicts *B* bounding boxes, a confidence rating, and a C assessment of probability for accuracy (Redmon et al., 2016: 780).

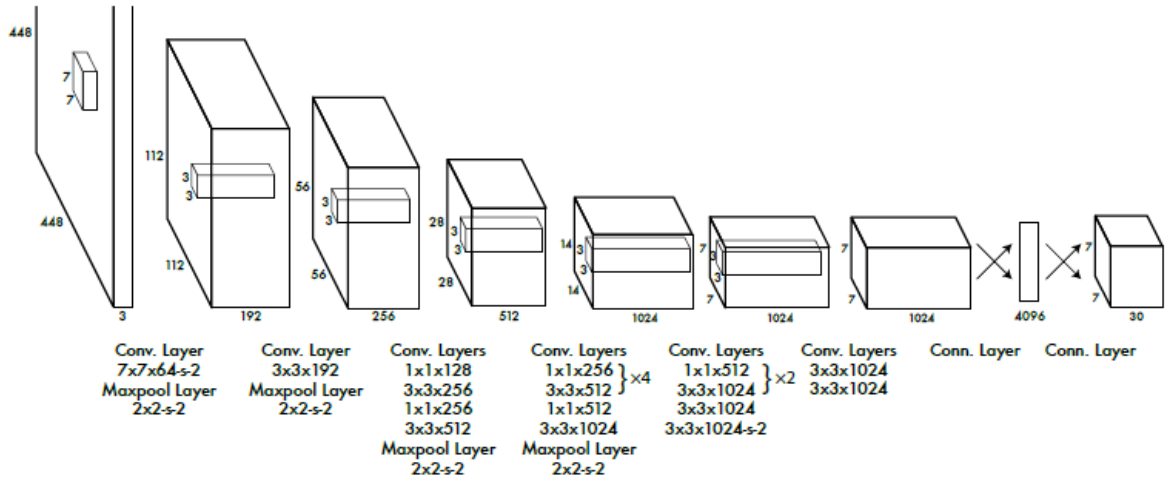


Figure 47. Basic architecture of YOLO. The detection network has 24 pattern recognition components followed by 2 fully connected layers (right). The pattern recognition components have local connectivity to one another and are well suited for identifying and learning patterns in images. The fully connected layers, on the other hand, have global connectivity to the previous layers. They are dense, and though can't capture spatial context like the previous ones, are used to help with decision-making and the production of the final output (Redmon et al., 2016: 781).

diverge from many traditional object detection systems on their basic architecture: Rather than splitting the detection process into separate stages as mentioned before (informative region selection, feature extraction, and classification), YOLO integrates all these stages into a single neural network.[150] The network consists of 24 pattern recognition components (convolutional layers)[151] followed by 2 fully connected layers (Figure 47),[152] which makes it highly optimizable and reduces computational time (Redmon et al., 2016: 781). This network is divided into three main components (Figure 48): the backbone, the neck, and the head (Bochkovskiy et al., 2020: 2; Terven et al., 2023: 1692).

-Backbone: This component extracts features from the input image. In most versions of YOLO, the backbone is a convolutional neural network (CNN),[153] which captures hierarchical features at different scales. Lower layers extract basic features (e.g., edges and textures), while deeper layers capture more complex features (e.g., object parts and semantic information) (Terven et al., 2023: 1692).

-Neck: The neck connects the backbone to the head. It aggregates and refines features from the backbone, enhancing their spatial and semantic information. Optimizing this part is crucial for improving feature representation (Terven et al., 2023: 1692).

-Head: The head processes the features from the neck and generates predictions for each candidate object, filtering out the most confident ones.

The integration of these components into a single neural network ensures a streamlined and efficient detection process, making YOLO particularly well-suited for real-time applications, such as archaeological research.

As a detection tool, YOLO works with images as an input by using ML tools like regression, classification and clustering to produce an output in the form of bounding boxes with each box

[150] Automated Neural network (p.156).

[151] **Pattern recognition components** or convolutional layers are core component of DL models. With their capability to capture spatial hierarchies and patterns through filters that slide across the inputted (image) data, they are especially useful for image and spatial data analysis. When stacked in in multiples in architectures such as YOLO's, they can recognize increasingly complex features (Goodfellow et al., 2016: 326)

[152] **Fully connected layers** are component of DL networks that differs from pattern recognition components in the fact that they have global connectivity to the neurons of the previous layer. They are typically used at the end of systems after pattern recognition components (see Figure 48) to interpret the features learned by these and generate the appropriate output (Goodfellow et al., 2016: 163)

[153] Convolutional neural networks (p.11).

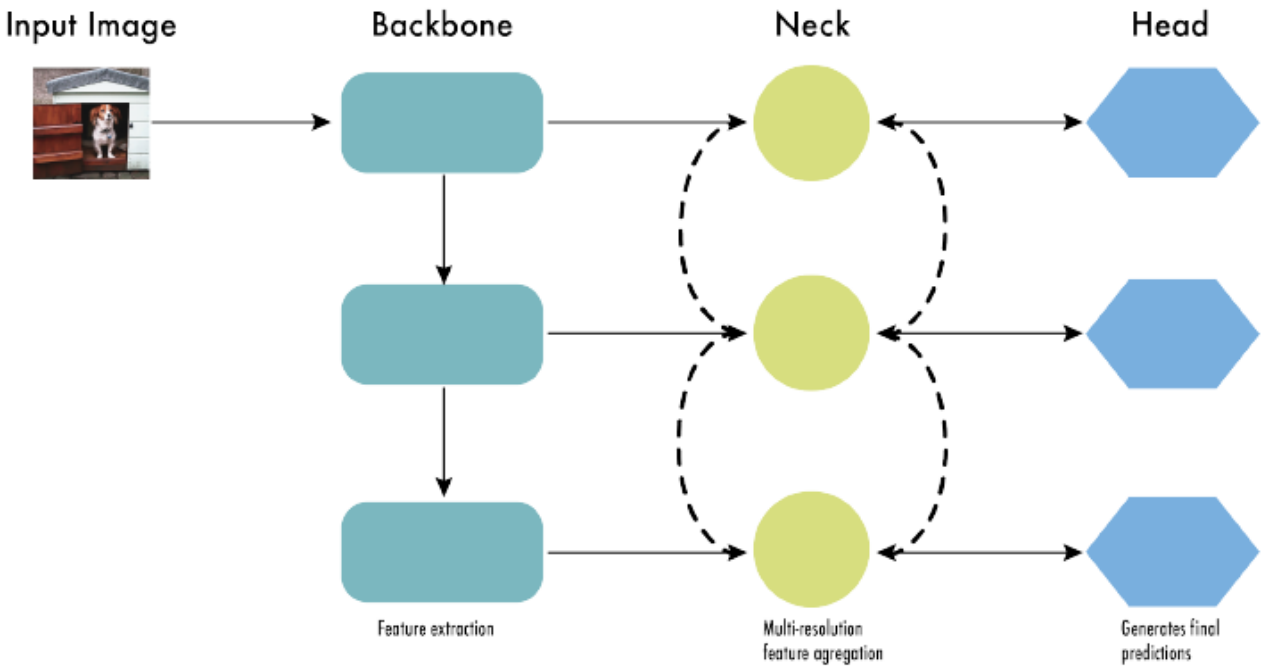


Figure 48. Most modern detectors have architecture that is often described as the backbone, neck and head. The backbone extracts vital features from the image. The neck refines the representation of the features, thus enabling the head to be able to make detections (in Terven et al., 2023: 169.

presenting a specific label (Redmon et al., 2016; Redmon and Farhadi, 2017; Bochkovskiy et al., 2020). In doing so, there are several problems that the models need to solve to produce an interpretable output, such as having to review a lot of bounding boxes for the correct fits, establishing a rule with which to classify them, handle all the redundancy of boxes on the final output, and more. While exploring exactly how these extra tasks are solved is outside the scope of the dissertation, we can explain the two main concepts/techniques involved in solving these issues, as they come up on the definition of the metrics presented in this dissertation (Appendix IV).

These are the concepts of Intersection over Union (IoU) and the postprocess technique of Non-Maximum Suppression (NMS).

-**Intersection over Union (IoU):** This is a metric that quantifies how well the model's prediction overlaps with the actual, human-labeled reference. It is defined as the ratio of the intersection area to the union area between the predicted bounding box and the ground truth box (Figure 49), and it

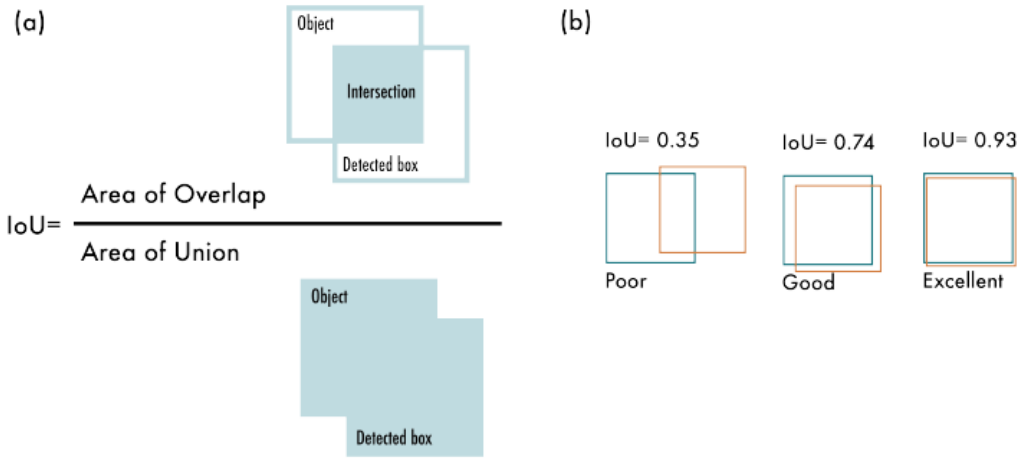


Figure 49. a) Inersection over Union (IoU) is calculated by dividing the intersection of two boxes by their union. b) examples of three different IoU values depending on box locations (Terven et al., 2023: 1683).

is used in average precision calculations to evaluate the accuracy of predicted bounding boxes (Terven et al., 2023: 1683).

-**Non-Maximum Suppression (NMS)**: Object detection models often generate multiple bounding boxes around the same object, including many with low confidence scores. NMS eliminates these low-confidence boxes, retaining only the most relevant, high-confidence detections as the final output. It is a post-processing technique used in object detection to reduce redundant bounding boxes and enhance accuracy (Neubeck and Van Gool, 2007; Terven et al., 2023: 1684). This step is crucial for improving the clarity and reliability of object detection results (Figure 50).

Figure 50. Typical output of a detection algorithm before (left) and after (right) the application of NMS (Terven et al., 2023:1684).

Appendix II

Xlendi Underwater Archaeological Park. Site Background

Contents

1 Site Background

1.1 Geographical Context

Xlendi Archaeological Park is located off the coast of Xlendi, a small town on the southwestern coast of Gozo, one of the islands forming the Maltese archipelago (Figure 51). The archipelago's strategic location in the Central Mediterranean and the morphology of its coastline have historically made it a crucial maritime hub. Its sheltered waters and central position facilitated extensive maritime activity, shaping the local culture into one deeply connected with the sea (Bonanno, 2005: 11).

Diodorus Siculus, in his *Bibliotheca Historica*, describes the importance of these islands in antiquity:

'For to the south of Sicily three islands lie out in the sea, and each of them possesses a city and harbors which can offer safety to ships in rough weather.' (Diodoro Siculus, Bibliotheca Historica: V, 12).

This was particularly true during the 8th century BC, when the Maltese islands became a Phoenician outpost, integrating them into a vast trade network spanning the Eastern and Western Mediterranean. This commercial activity persisted through the Punic and Roman periods (Bonanno, 2005: 20-21; Aubet, 2001: 138). Given these maritime connections, Xlendi was likely

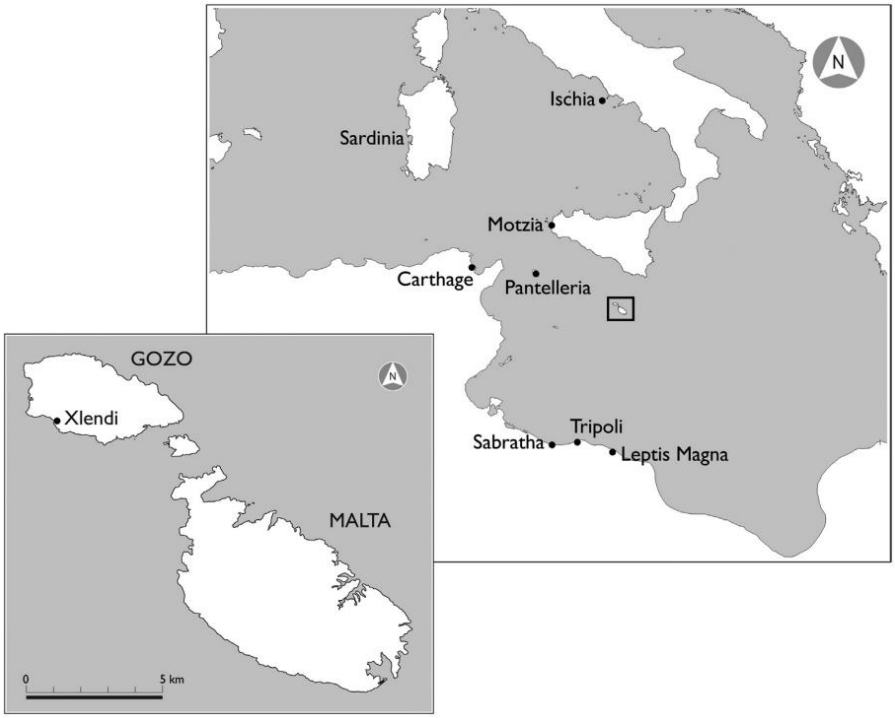


Figure 51. Position of Malta in the Mediterranean (Gambin et al., 2018:72).

one of the most significant harbors in the archipelago at the time (Gambin, 2005: 21).

Xlendi's harbor is situated within a deep, narrow inlet on Gozo's southwestern coast. Historically, its proximity to Rabat/Victoria, Gozo's capital, ensured its importance until the construction of the Mgarr breakwater in southeastern Gozo (Figure 52). Today, commercial trade no longer takes place in Xlendi, though the port remains an active fishing hub for local fishermen.

The bay extends inland through a deep valley, which—based on erosion patterns—historically reached 200 meters further inland before silting up (Gambin, 2005: 22; Azzopardi, 2006: 4, 2013). The entrance to Xlendi Bay is deceptively narrow due to the presence of two submerged reefs extending from the cliffs of Ras il-Bajda and Ras Maħrax, posing significant risks to navigation. Depths in the narrowest section range from 2m to 20m, increasing sharply beyond the cliffs, where the seabed rapidly drops to over 100 meters (Azzopardi, 2006: 7).

To the north, Ras il-Wardija rises 162 meters above sea level, serving as a key navigational landmark visible from 12 nautical miles away (Gambin, 2005: 20, Figures 53, 54). However, this headland also presents navigational challenges. The combination of strong winds, wind-opposing currents, wave effects from the reefs, and the difficulty of spotting the bay from a distance creates treacherous conditions for seafarers (Azzopardi, 2006: 8).

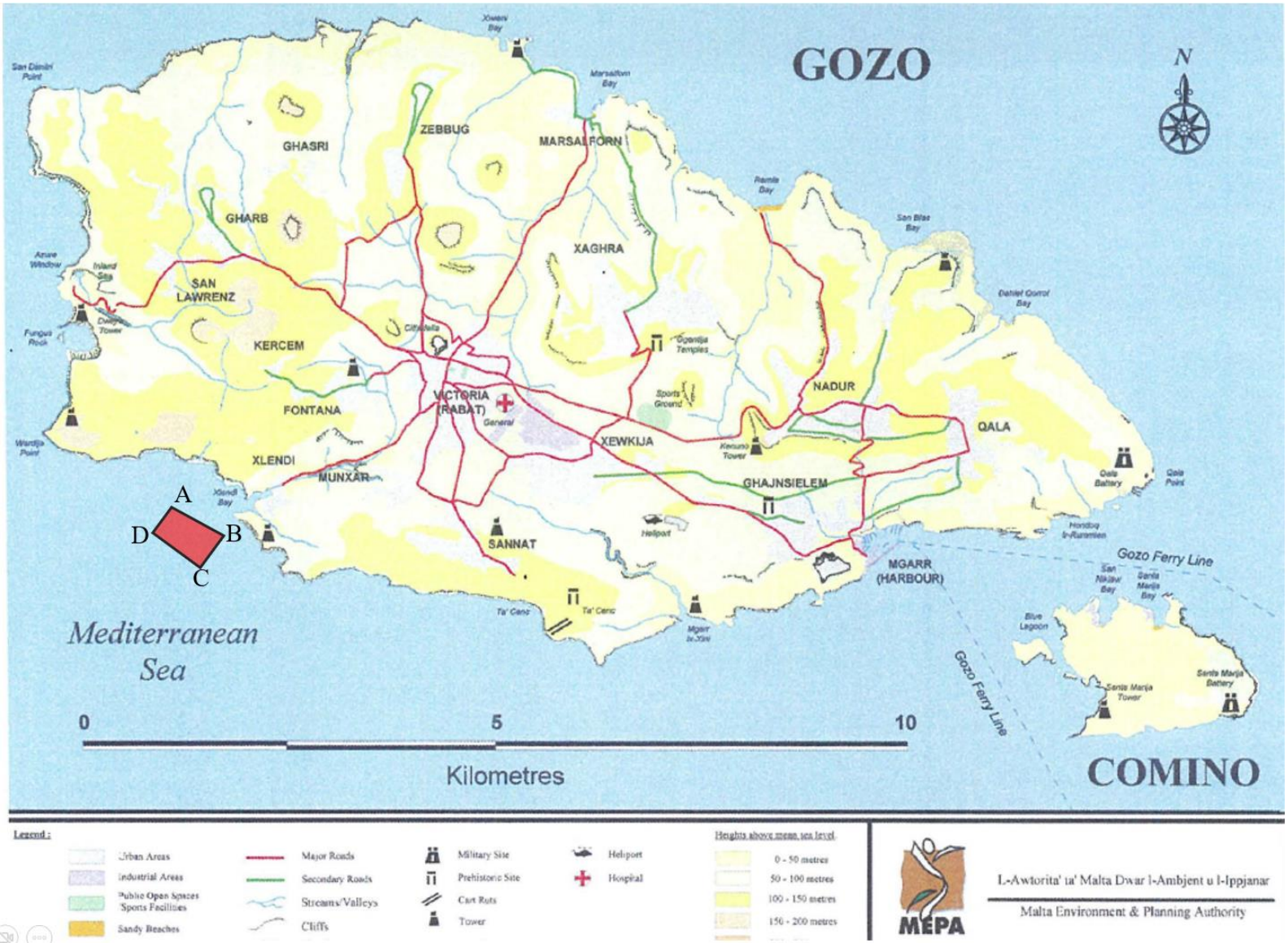


Figure 52. Elevation map of Gozo showing the boundaries of Xlendi Archaeological Park in relation to Gozo (A.B.C.D).

Figure 53. Ras il-Bajda with the Knight's Tower. Ras Mahrax looms in the background (Azzopardi, 2006:170).

1.2 Historical Context

Today Xlendi serves as a small fishing and recreational port, but archaeological evidence suggests a much more active role in the past. Defensive structures such as a semi-circular defensive tower form the 15th century found in the upper part of Xlendi Valley (Vella, 2006: 59), and another coastal defense tower built by the Knights of St. John in the nearby headland in 1650 (De Lucca, 1990: 151) constitute some of the examples that have often been used as evidence to support Xlendi port being one of the main entry points to Rabat from the Middle Ages and onwards. However, Xlendi's importance predates these structures. Archaeological sites in the surrounding landscape—such as Tac-Ċawla and Ta' Marziena temple—demonstrate continuous human activity since the Neolithic period. Substantial Bronze Age and Phoenician remains further highlight the area's historical significance (Azzopardi, 2006: 11-14).

Due to Xlendi's position between Ras il-Bajda and Ras Mahrax, and with Ras il-Wardija as a prominent coastal marker (Figures 52, 53), it is highly likely that, in antiquity, every vessel approaching Malta from Carthage, Pantelleria, Western Sicily, or the Western Mediterranean

would have seen Xlendi as the first available refuge (Gambin 2005: 21).[154] This aligns with the concept of 'cultic topography', linking maritime activity with nearby sanctuary sites. The Ras il-Wardija sanctuary, excavated by the *Missione Archaeologica Italiana* (1964-1967), was an active Punic religious site (4th-1st century BC), reinforcing the idea that Xlendi functioned as a landmark for both navigation and religious practices (Horden and Purcell, 2000: 401-407; Vella, 2002; Gambin, 2005: 21; Azzopardi, 2013: 287).

The archaeological discoveries in Xlendi's seabed further support its importance in antiquity. Xlendi Archaeological Park and the Xlendi Phoenician Shipwreck are two of the most well-documented underwater sites, containing artefacts from periods of high maritime activity (Gambin, 2015; Gambin et al., 2018, 2021). Most of the materials found in this area date to the Phoenician, Punic, and Roman periods, likely resulting from shipwrecks caused by hazardous navigation conditions or alternative depositional processes related to the site's possible cultic relevance (Figure 15).

1.3 Past Interventions

Xlendi Bay has long been a site of interest for divers, including amateur enthusiasts who, while inadvertently causing damage, played a role in bringing attention to its archaeological significance. The discovery of submerged artefacts led to professional interventions and, ultimately, the implementation of protective measures to safeguard this rich underwater cultural heritage. Since the 1960s, multiple investigations have been conducted. These are thoroughly reviewed in Elaine Azzopardi's 2006 assessment of Xlendi's archaeological value (Azzopardi, 2006: 19).

The first recorded discovery of submerged artefacts in Xlendi Bay occurred in 1961, when a diving team from H.M.S. Falcon encountered archaeological materials during training exercises. This prompted further exploration by a team from *Imperial College London* and the Institute of Archaeology (UCL), who recovered and documented artefacts from depths of up to 65 meters (Woods, 1962; Azzopardi, 2006: 20). Their findings were photographed, and a three-dimensional drawing of the reef was produced to aid documentation efforts.

[154] This notion is reinforced by publications that show Malta as a key maritime linchpin for sailors all across the Mediterranean (Gal et al.2023: 3). An example of this can be seen in Figure 14 (p. 33).

Between 1970 and 1993, no major archaeological projects were conducted in Xlendi, but unauthorized artefact retrieval continued throughout this period (Accopardi, 2006). A new phase of investigation began in the early 1990s, when the (Malta) Museums Department with GISMER's logistical support carried out a deep-water survey (Gauci and Grima, 1993; Grima 1993). This survey revealed the first clear evidence of a coherent underwater archaeological deposit, later recognized as Xlendi Archaeological Park, primarily consisting of Punic pottery.

Subsequent exploration surveys employed submersibles to systematically document and analyze the scattered underwater materials (Atauz and McManamon, 2000; Atauz, 2004: 31; Figure 54). More recently, the University of Malta, Heritage Malta, and the Superintendence of Cultural Heritage launched the Tower Wreck Project, dedicated to the in-depth study of the site. Now, in 2025, preparations are underway for the fifth consecutive campaign, with a primary focus enhancing the understanding and preservation of Xlendi's maritime heritage (Gambin and Sauskemat, 2021,2022 and 2023).

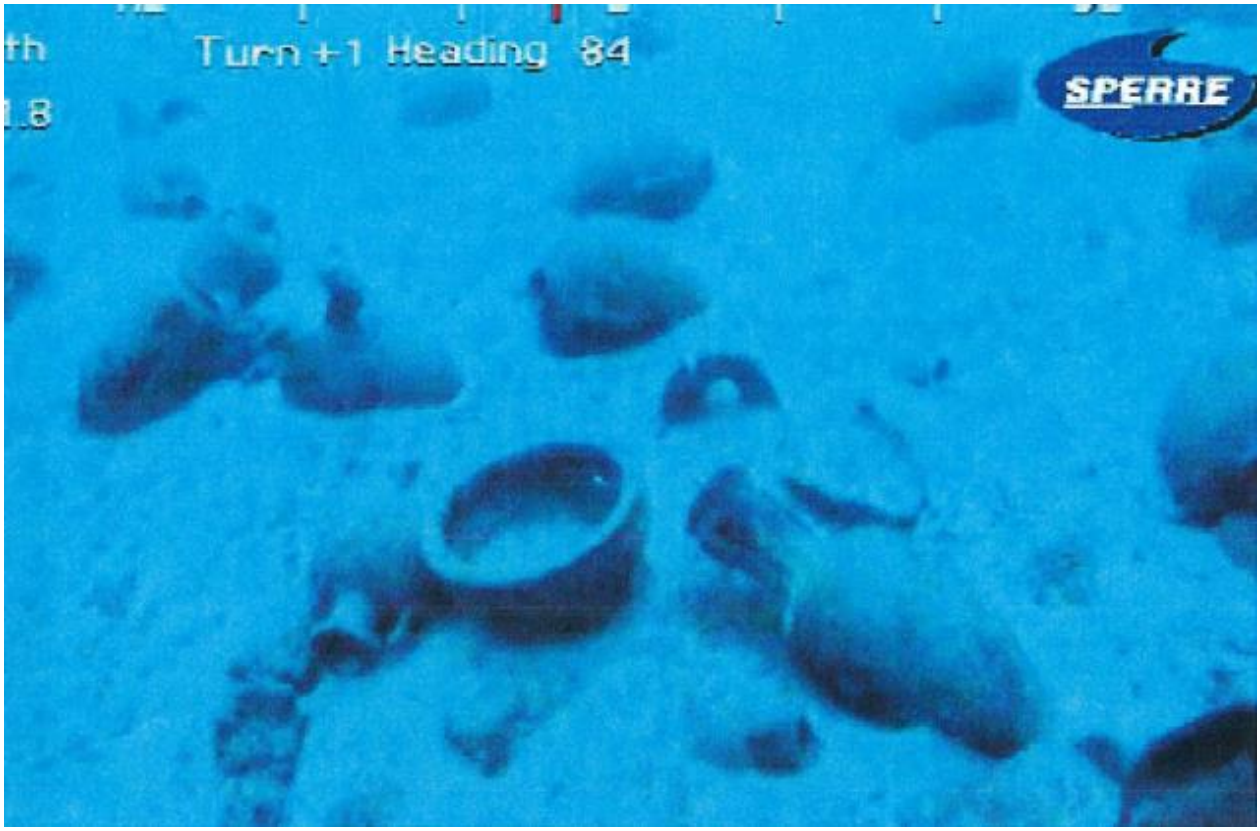


Figure 54. Examples of ROV images taken in Xlendi during the 2001 survey (Atauz, 2004: 31).

2 The Assemblage. Previous work.

The first detailed documentation of some of the archaeological materials from Xlendi Bay was carried out by Ayşe Atauz in 2004 following the INA 2001 survey (Figure 55). This publication identified and catalogued seven distinct amphora types, primarily of Punic and Late Roman origin, dating from the 4th century BC to the 3rd century AD (Figure 55, Table 10, Atauz, 2004: 375–378).

Of particular relevance to this study is the work of Elaine Azzopardi, conducted as part of her master's dissertation, *The Xlendi Bay Shipwrecks: An Archaeological Study* (Azzopardi, 2006). Unlike previous studies, Azzopardi examined the entirety of Xlendi Bay's archaeological materials, drawing upon data from the three major expeditions conducted before 1975 (whose recovered artefacts are housed in the Gozo Archaeological Museum) and the photographic documentation from the 2001 INA survey (Figure 54).

Through this research, Azzopardi identified and classified the available artefacts, contextualized them within their historical background, and compiled a comprehensive catalogue that included drawings and photographs of representative objects (Azzopardi, 2006, Table 11). However, because Xlendi Archaeological Park had not yet been formally defined as a site, her work served more as an archaeological assessment of Xlendi Bay as a whole, rather than a focused study of the wreck's assemblage.

Both Atauz (2004: 381) and Azzopardi (2006: 25) acknowledged the challenges posed by the fragmentary and dispersed nature of the material record, which made it difficult to determine the true significance of the assemblage. Nonetheless, their efforts remain invaluable. Given that our assessment of the assemblage relied heavily on low-resolution images, Azzopardi's analysis of physical artefacts has provided greater accuracy in refining certain typological classifications. Furthermore, these studies have established a baseline typology for the region, offering a foundation for more specialized research projects and potential interpretations of specific artefact distributions.

Today, with the extensive visual dataset gathered over three seasons of the Tower Wreck Project (Figure 56), combined with the foundational research of Atauz and Azzopardi, the development of a refined typology for the site is now a more achievable and methodologically sound undertaking.

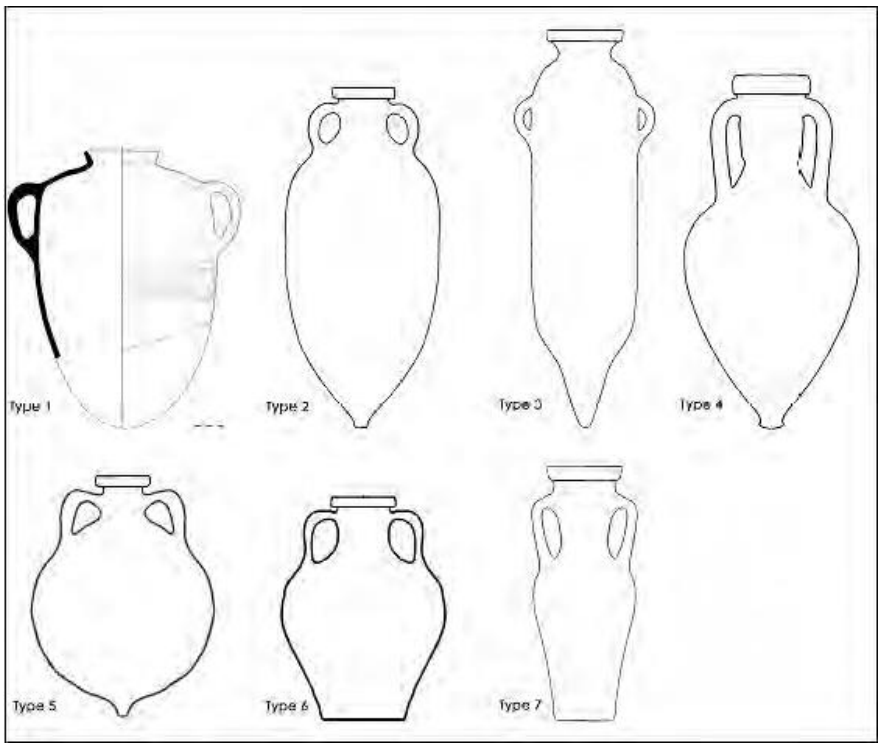


Figure 55. Main typologies identified by Atauz in 2004. From left to right: Ramon 3.2.1.2, Malta 1, Maña C, Graeco-Italic, Dressel 20, an unidentified double handle flat-based amphora, and Sagona urn form III-IV: 4a-b. (Atauz, 2004: 377).

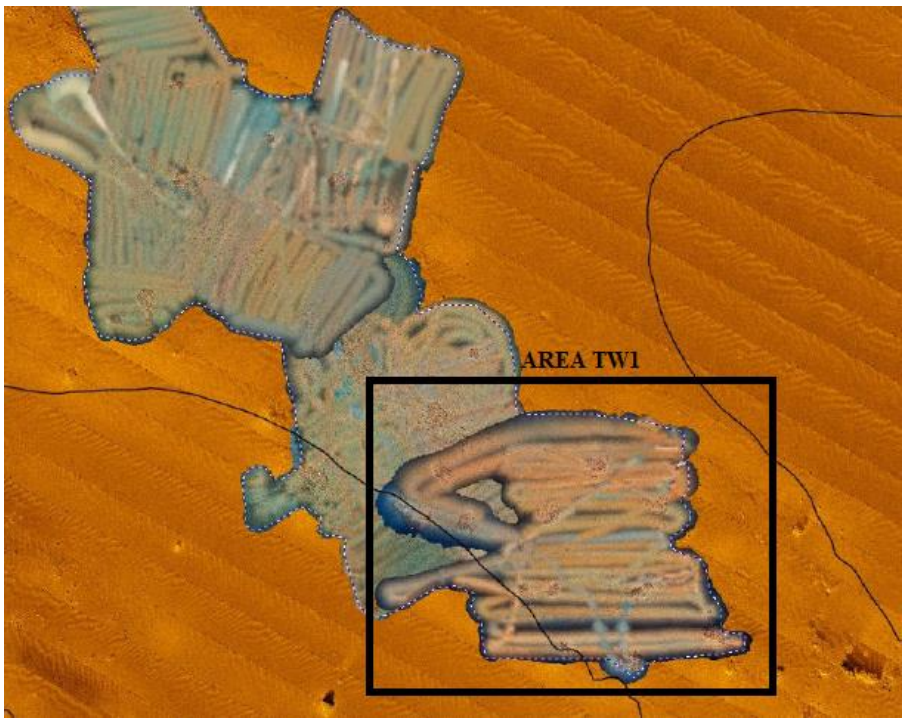


Figure 56. Total surveyed surface of the park against geophysical data from SSS (as of 2023). Area TW1 corresponds to the location of our dataset. Note that even the surveyed surface is still a fraction of the total surface of the site (Department of Classics and Archaeology, University of Malta).

Catalogue No.	Type	Approximate Date	Period	Assemblage	Quantity
Xlendi 1	Tarxien Cemetery	2500-1500 BC	3rd – 2nd millennium BC	1	1
Xlendi 2	Ramon 2.1.1.2	600 BC	End 7th – beginning 6th century	2	1
Xlendi 3	Ramon 2.2.1.2	420-350 BC	End 5th – 1st half 4th century	3	at least 3
Xlendi 4	Urn 3-4: 3	410- 300 BC	4th century	4	1, possibly more in deep water
Xlendi 5	Urn 3-4: 4	410-300 BC	4th century	4	3
Xlendi 6	Ramon 3.2.1.2	290-146 BC	3rd – mid 2nd century	4	3, many more in deep water
Xlendi 7	Cintas 176 jug	3rd c BC	3rd century	4	1
Xlendi 10	Ramon 7.3.1.1	210-150 BC	mid 2nd century	5	1
Xlendi 11	Ramon 7.4.2.1	190- 150 BC	Mid 2nd century	5	1
Xlendi 13	Flat–bottomed amphora	140-100 BC	End 2nd century	5	1
Xlendi 14	Graeco-Italic	Approximately 150 BC	Mid 2nd century	5	2, possibly more in deep water
Xlendi 15	Malta type 1	150-100 BC	2nd ½ of 2nd century	5	10, possibly more in deep water
Xlendi 16	Dressel 1A	Approximately 140-100	Mid 2nd – beginning 1st century	5	4, possibly more in deep water
Xlendi 17	Lamboglia 2	Approximately 140-0 BC	Mid 2nd – end 1st century	5	2
Xlendi 19	Dressel 2-4	50 – 1 BC	1st century	6	7, more in deep water
Xlendi 22	K 109	250 AD	Mid 3rd century	7	1
Xlendi 23	K 107	250 AD	Mid 3rd century	7	1
Xlendi 24	Tripolitana III	250 AD	3rd century	7	1
Xlendi 25	Keay V	200-250 AD	3rd century	7	1
Xlendi 26	Africana II variant	250 AD	3rd century	7	1
Xlendi 27	Keay XXVI	300 – 500 AD	4th – 6th century	8	3
Xlendi 28	Keay XXXVB	450 AD	5th century	8	1
Xlendi 29	Keay XXXII	300-400 AD	4th – 5th century	8	1
Xlendi 30	Keay VIII A	450 – 500 AD	7th century	9	3
Xlendi 31	Islamic jug	10th – 13th c AD	Islamic	10	1

Table 10. Azzopardi's typological chart for Xlendi Bay (Azzopardi, 2006: 106).

Appendix III

Catalogue of Finds

In this section, the various ceramic artefacts present in the dataset will be examined in detail. The primary objectives are to establish a foundational understanding of the site and to develop a typological chart that can serve as a reference for training typological models. The analysis will focus on the physical attributes, measurements, distribution, dating, and correlation with existing typologies. Each form will be accompanied by a drawing and an example from the dataset.

FORM 1

Type:

This amphora type corresponds to Ramon's T-3.2.1.2 and Sagona's amphora IV-V:1, making it one of the most prevalent forms found at the site. It is characterized by an ovoid shape with a high, carinated shoulder that aligns with the vessel's maximum diameter. The rim is short and collar-like, while the body tapers from the midpoint towards the base. The handles, narrow in proportion, are positioned just below the shoulder.

T. Correlation:

- Ramon T-3.2.1.2 (Ramón Torres, 1995: 183).
- Sagona amphora IV-V:1 (Sagona, 2002:92).
- Xlendi 6 (Azzopardi, 2006:48).

Measurements:

- Total height: 55-65cm.
- Maximum diameter (shoulder): 32-35cm and a
- Mouth diameter: 11-13cm.

Date:

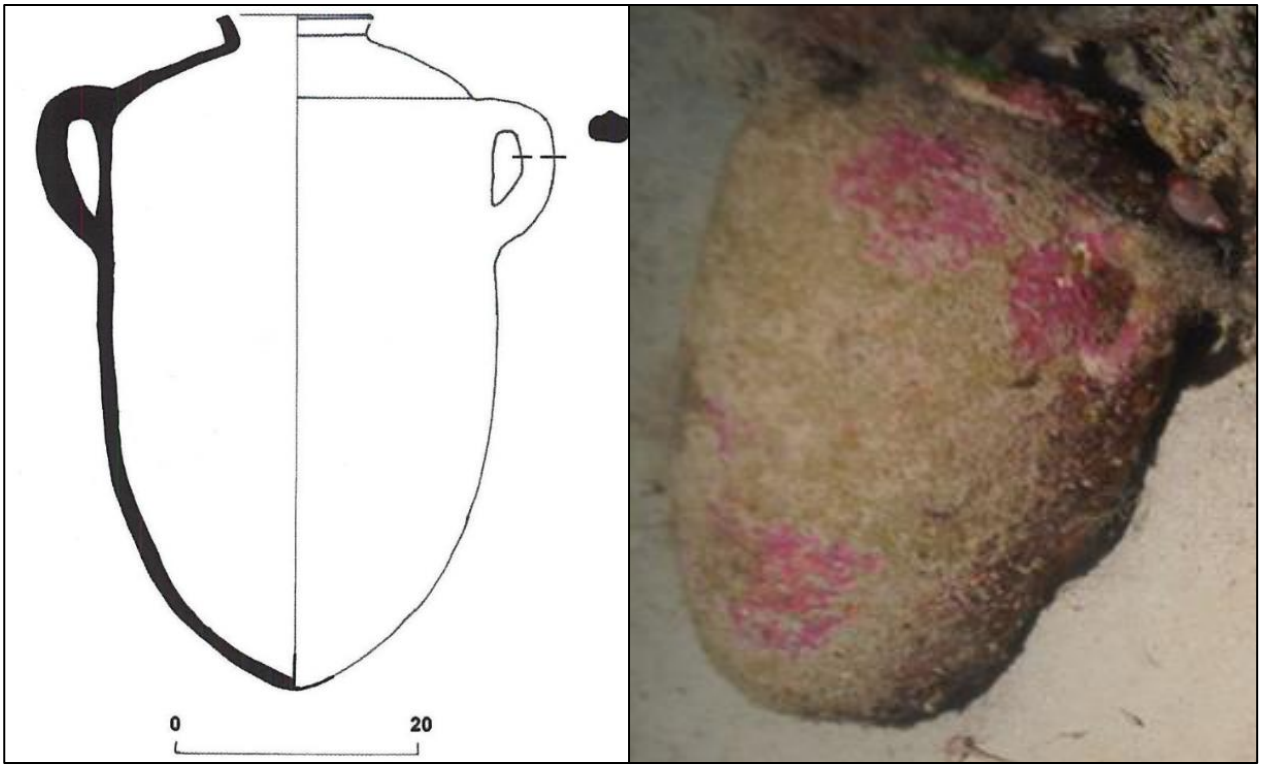
Traditionally dated to the 3rd century BC.

Place of production and distribution:

While this amphora type is most commonly associated with Western Sicily and Carthage, its numerous typological variations and fabric differences suggest a broader production and distribution network (Ramón Torres, 1995: 519; Sagona, 2002: 667–8). Notably, its resemblance

to Sagona's amphora form V:1, a prevalent type in Malta's funerary context, supports the hypothesis that the Maltese archipelago may have been an additional center of production (Brunella, 2004).

Drawing and example from the dataset:



FORM 1

FORM 2

Type:

This amphora type is incredibly abundant at the site and appears in several variations of the general form known as Ramon 2.2.1.2. Its shape is typically ovoid and symmetrical, tapering towards both the mouth and the base. Some examples exhibit a narrower, more cylindrical body. The neck is often very short or entirely absent, with a small, folded-over rim. The large, arched handles, with a half-circle or oval cross-section, are positioned around the midpoint of the upper section of the vase. While sharing similarities with Form 1, this type lacks the high carinated shoulder characteristic of the Ramon T-3.2.1.2. Instead, its maximum diameter is typically found around the central part of the vase rather than near the shoulder.

Type Concordance:

- Ramon T-2.2.1.2 (Ramón Torres, 1995: 179).
- Sagona amphora IV:1 (Sagona, 2002:90).
- Xlendi 2 (Azzopardi, 2006:40).

Measurements:

- Total height: 60-64cm.
- Maximum diameter (lower half): 30-34cm.
- Mouth diameter: 12-14cm.

Date:

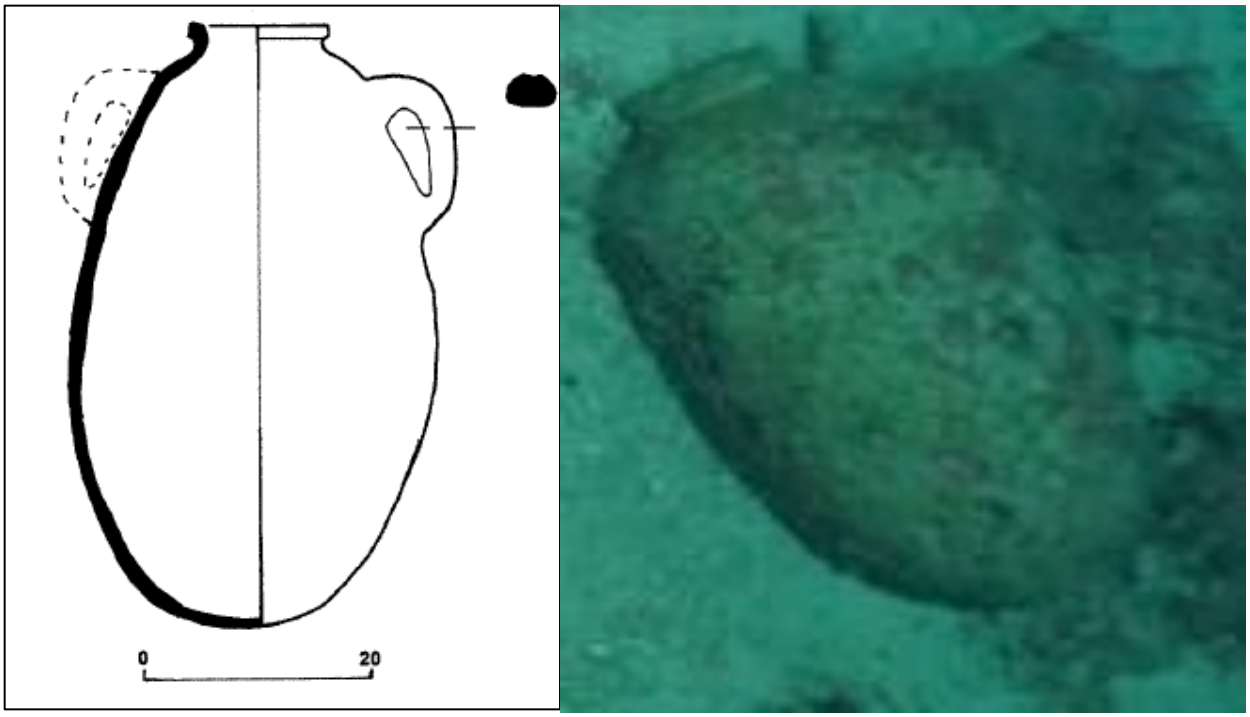
This form originates in the early 5th century BC, with its peak usage occurring in the first half of the 4th century BC.

Place of production and distribution:

Produced in Tunisia, western Sicily, and Malta, this amphora type represents one of the earliest Punic forms to be exported in significant quantities from the beginning of the 5th century BC. It also serves as a precursor to many amphorae of the 3rd century BC and later (Ramón Torres, 1995: 179; Azzopardi, 2006: 40). Variations within this form, observable in Xlendi Archaeological Park as evidence of sustained commercial activity, suggest the existence of multiple workshops (Ramón Torres, 1995: 179). In Malta, these amphorae frequently appear in terrestrial contexts, where they are identified as Sagona Amphora IV:1. This suggests a local production influenced by broader

Punic ceramic traditions (Sagona, 2002: 90). Differences in dating proposed by various scholars further support this interpretation.

Drawing and example from the dataset:



FORM 2

FORM 2b

Type:

Amphorae of this form are relatively common at the site, appearing in various sizes. This variation, combined with the challenges of identifying precise shapes from (relatively) low-resolution images, makes distinguishing between Ramon's T-2.2.1.4 and T-2.1.1.2 difficult. This is a significant issue, as these shapes belong to different periods.

Ramon T-2.2.1.4 was originally defined based on two examples discovered in Malta, which the author linked to the previously discussed Form 2 (T-2.2.1.2). He described it as a slightly more bellied version of Form 2 and speculated that it could be specific to the Maltese archipelago (Ramón Torres, 1995: 180). In the Maltese context, Claudia Sagona later identified a similar form as Sagona Amphora Type III-IV:1, basing her definition on a much larger dataset (Sagona, 2002: 89).

In terms of morphology, both authors describe this amphora as characteristically egg-shaped, with its maximum width positioned lower on the body. It often features a distinctly rounded, somewhat paunchy lower half that tapers into a leaf-shaped base. Additional defining features include a short neck, a compact rolled rim, and high-positioned, rounded handles with a circular cross-section.

Type Concordance:

- Ramon T-2.1.1.2, Ramon T-2.2.1.4 (Ramón Torres, 1995: 180).
- Sagona amphora III-IV:1 (Sagona, 2002:89).

Measurements:

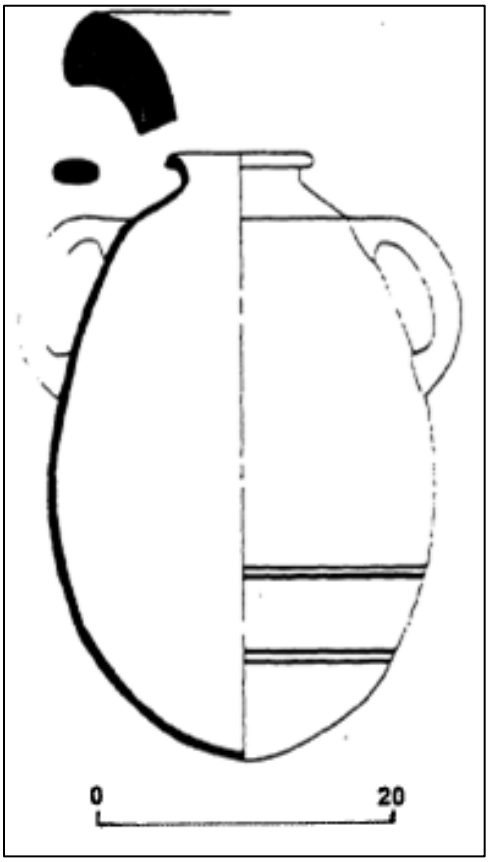
- Total height: 58-66cm.
- Maximum diameter (body): 31-40cm,
- Mouth diameter: 14-16cm.

Date:

4th century BC.

Place of production and distribution: Malta.

Drawing and example from the dataset:



FORM 2b

FORM 3

Type:

Quite common at the site. Clearly of Punic origin, it closely resembles the egg-shaped Form 1 but features a noticeably longer and more cylindrical body. This elongation follows a broader trend observed in Punic amphorae typologies from the 4th century BC onward (Ramón Torres, 1995: 204).

Form 3 corresponds to Ramon's T-7.1.1.1 or the similar T-7.1.2.1, both of which the author identified as commonly found alongside T-2.2.1.2 (Form 2 in this catalog) in shipwrecks such as the 'Pecio del Sec' (Ramón Torres, 1995: 61).

On Xlendi Archaeological Park, Form 3 is characterized by a cylindrical, elongated body and a very short neck, which acts as a transitional surface between the body and the short, vertical rim. The handles, with an oval cross-section, form three-quarters of a circle with a slight elbow and are positioned just below the shoulder.

Type Concordance:

- Ramon T-7.1.1.1, Ramon T-7.1.2.1 (Ramón Torres, 1995: 204).

Measurements:

- Total height: 60-70cm.
- Maximum diameter (mid-body): 25-27cm.
- Mouth diameter: 11-13cm.

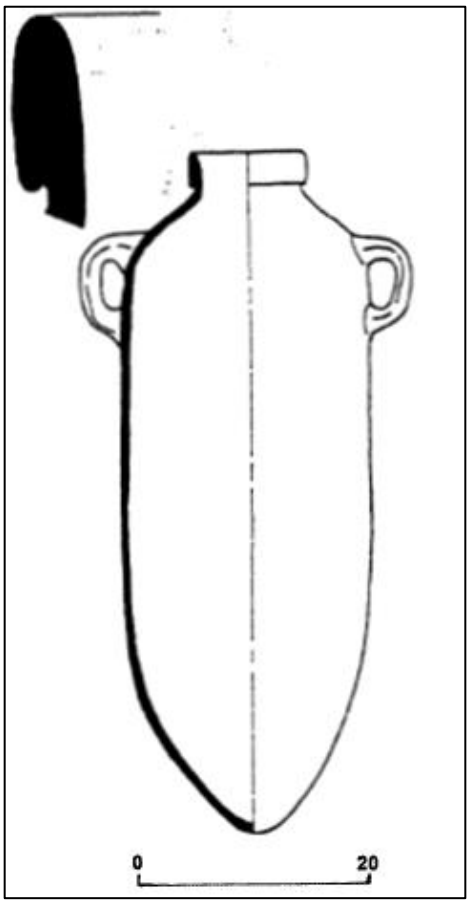
Date:

3rd century BC.

Place of production and distribution:

The production of T-7.1.1.1 and its variants appears to have been concentrated in Tunisia and western Sicily, based on the frequency of finds (Ramón Torres, 1995: 205). In terms of distribution, this type of cylindrical Punic amphora is widespread across the western Mediterranean, often appearing alongside more globular forms.

Drawing and example from the dataset:



FORM 3

FORM 4

Type:

Graeco-Italic amphorae are among the most common and recognizable forms in the assemblage. The term "Graeco-Italic" refers to amphorae found throughout Greek and Roman contexts, dating from the 4th to the 2nd centuries BC (Vandermersch, 1994).

The examples visible at the site align with the most general description of this form. They feature a pear-shaped body, widest at the top and tapering downward, terminating in a short, solid spike. A sharply carinated shoulder transitions into a long cylindrical neck, which ends in a triangular rim. The ovoid handles extend from just below the rim to the high point of the shoulders.

Type Concordance:

- Xlendi 14 (Azzopardi, 2006:60).

Measurements:

- Total height: 80-83cm.
- Maximum diameter (body, below the shoulder): 40-41cm.
- Mouth diameter: 20-21cm.

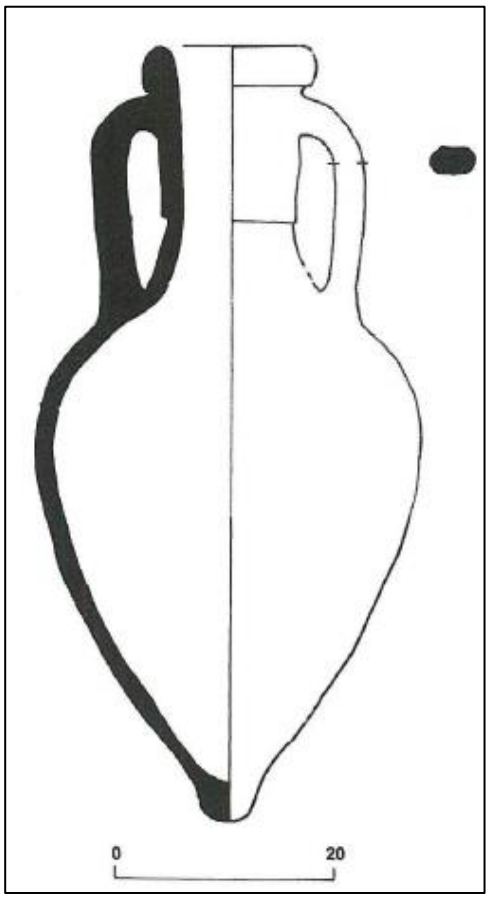
Date:

Graeco-Italic amphorae are generally dated between the 4th and 2nd centuries BC, with variations in body, neck, and rim morphology serving as chronological markers. A notable trend is the increasing elongation of the body in later models (Will, 1982: 357, 1987: 178). Based on these characteristics, Form 4 likely belongs to the later phase of this typology.

Place of production and distribution:

These amphorae were primarily produced in Tyrrhenian and Adriatic Italy, including Magna Graecia. However, their widespread distribution across the western Mediterranean has led to speculation about additional production centers in France, the Iberian Peninsula, and North Africa (Williams, 1986: 85).

Drawing and example from the dataset:



FORM 4

FORM 5

Type:

This form is easily recognizable in the dataset images and is widely present across the site. In the Maltese context, it was first classified as Malta Type 1 (Brunella, 2004: 88). However, the author notes its resemblance to a broader morphological group of amphorae with ovoid bodies and short cylindrical necks, particularly the Brindisi types such as Dressel 26, Tripolitanian 1, and Apania VII. Malta type 1 are described as a break from the local Maltese Punic tradition adopting typological features more akin to Republican-period amphorae (Anastasi, 2019:42).

Form 5 features an ovoid body, characteristic of many Punic amphorae like the Ramon Series 2 and 3, but on a generally larger scale. It has a knobbed base, rounded shoulders, and a cylindrical neck that ends in a thick-collared rim. The handles are circular in section and extend from just beneath the rim to the upper part of the shoulder. They are the distinctive element that separates them from Punic types.

Type Concordance:

- Malta type 1 (Brunella, 2004:88)
- Xlendi 15 (Azzopardi, 2006:61).

Measurements:

- Total height: 90-100cm
- Maximum diameter (body): 30-40cm
- Mouth diameter: 14-18cm

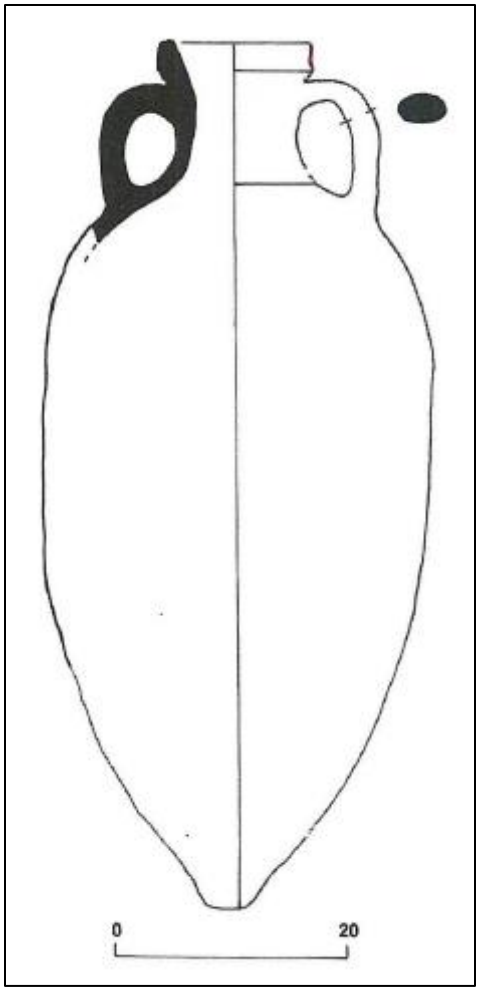
Date:

Often dated to the 2nd century BC based on contextual ceramic markers and its relation to the Brindisi types.

Place of production and distribution:

Initially believed to be of Maltese origin due to its similarities with local forms and its concentration in the archipelago, mineralogical studies have confirmed this assumption (Brunella and Capelli, 2000; Azzopardi, 2006: 63, Brunella, 2004.). However, it cannot be said that Malta type 1 is exclusively found in Malta. Similar versions of Malta 1 have also been found in the Adriatic coast of Italy (Anastasi, 2019: 24-42) and Tunisia (I. Ben Jerbania, 2017: 180-88).

Drawing and example from the dataset:



FORM 5

FORM 6

Type:

This is the only form at the site with such an elongated shape, making it easily identifiable. The amphorae found at Xlendi Archaeological Park closely resemble the Maña C or Ramon T-7.4.2.1 types, characterized by a large, "mushroom-shaped" rim, a long and narrow cylindrical body, and an equally elongated peg base. The handles, though often difficult to discern, are short and positioned just below the shoulder on the upper part of the straight body.

Type Concordance:

- Dressel 18 (Dressel, 1899).
- Maña C (Maña, 1951:75).
- Ramon T-7.4.1.2 (Ramón Torres, 1995: 209).
- Xlendi 11 (Azzopardi, 2006:56).

Measurements:

- Total height: 100-110cm.
- Maximum diameter (shoulder): 25-28cm.
- Mouth diameter (outer rim): 22-25cm.

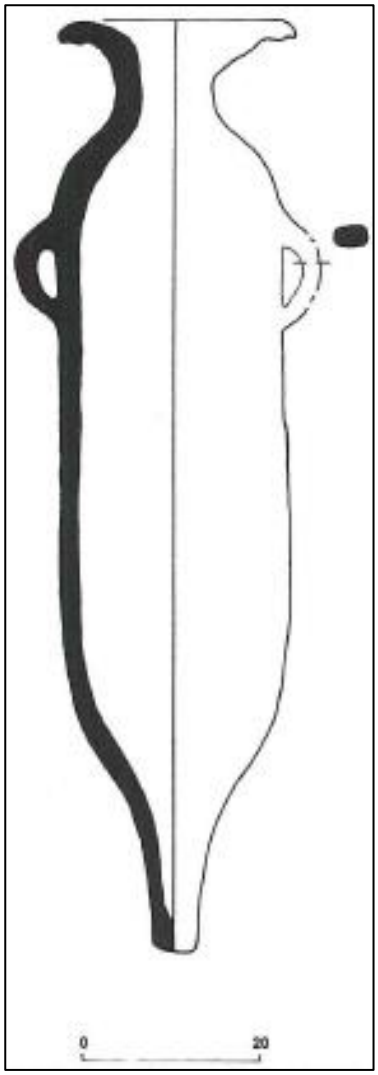
Date:

First half of the 2nd century BC.

Place of production and distribution:

This type of amphora was produced in well-known ceramic centers in North Africa, particularly in the Carthaginian area of Dermech-Ben Attar (Ramon Torres, 1995: 209). Examples of this form are found along the northern coast of Africa and throughout the western and central coasts of the Mediterranean (Bechtold 2012: 6; Bechtold, 2018)

Drawing and example from the dataset:



FORM 6

FORM 7

Type:

Dressel 1 amphorae are characterized by their ovoid body, long cylindrical neck, and straight handles that extend from just below the collared rim to the top of the carinated shoulder. Only a few examples of this type have been identified on the site. Their recognition in underwater photography is particularly challenging due to their close resemblance to the more common Form 5, from which they evolved into one of the first truly distinctive Roman amphora productions.

Dressel 1 amphorae are typically subclassified into three types: Dressel 1A, 1B, and 1C. While they share similar body shapes, these variations are dated to specific timeframes from the second quarter of the 2nd century BC to the 1st century BC and can be distinguished by subtle differences, such as rim shapes (Peacock, 1977; Tchernia, 1986). Given the condition of the amphorae in the underwater environment and the quality of available images, it was not possible to determine the exact subtype for the examples from Xlendi Archaeological Park. However, based on Azzopardi's extensive study of materials from Xlendi Bay, they can tentatively be classified as Dressel 1A (Azzopardi, 2006: 65).

Type Concordance:

- Dressel 1 (Dressel, 1899).
- Xlendi 16 (Azzopardi, 2006:65).

Measurements:

- Maximum height: 97-117cm.
- Maximum diameter (high body): 28-29cm.
- Mouth diameter: 17-18cm.

Date:

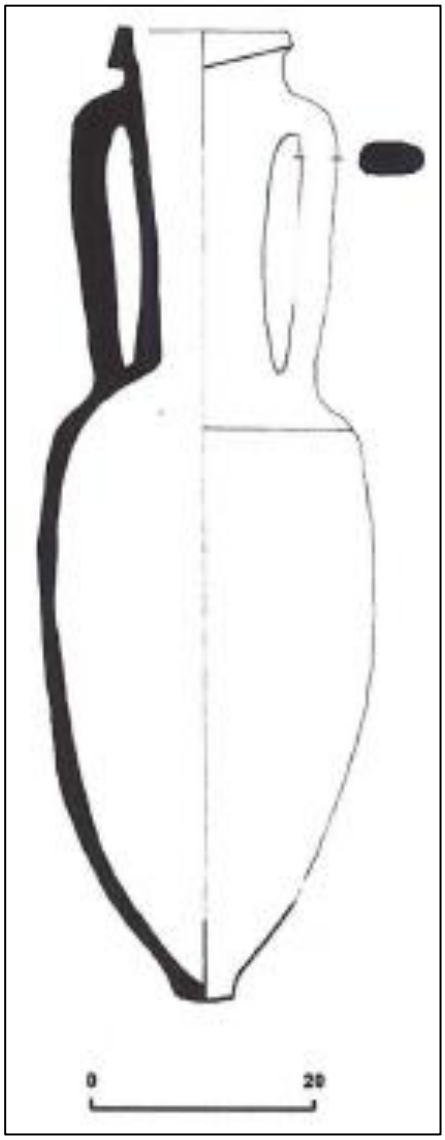
130-50 BC.

Place of production and distribution:

Dressel 1 amphorae were primarily produced along the Tyrrhenian coast of Italy, from Etruria to Campania. While examples are found in the Eastern Mediterranean, they were widely distributed across the Western Mediterranean, particularly in Italy, southern France, Spain, and even as far as Britain and Germany.

Drawing and example from the dataset:

.



FORM 7

FORM 8

Type:

An uncommon ceramic form found at the site is the Sagona urn type III-IV: 4a-b. This form is easily recognizable as one of the few ceramic items that are not amphorae. The urns are elegantly crafted, featuring tall, slightly outwardly flaring necks and rims. Their widest point is at the shoulders, from which curved double cordon handles extend, attaching at the mid-neck and shoulder. The form exhibits a distinct S-shaped profile, transitioning from a flat, slightly flaring base to a convex lower body, then expanding at the shoulders before narrowing again into the neck.

While these general characteristics remain consistent across all examples, their size varies. Claudia Sagona, who first defined them in the Maltese context (2002:103), notes this variability. Although we cannot discern the specific decorative patterns she describes, the difference in size between individual pieces is evident when one looks at the examples from the site. Form 8b within this classification may represent an example of this variation.

Type Concordance:

- Sagona urn III-IV: 4a-b (Sagona, 2002: 103).
- Atauz Type 7 (Atauz, 2004: 380).
- Xlendi 5 (Azzopardi, 2006: 47).

Measurements:

- Maximum height: Traditionally, they keep to heights from 19 to 35 cm. Sagona, however, mentions how some urns from marine contexts off the coast of Gozo (like Xlendi) have been found to be quite larger, standing up to 65cm. This is the case for the ones we present at Xlendi Archaeological Park.
- Maximum diameter (upper body): 18-22cm
- Mouth diameter: 12-16cm

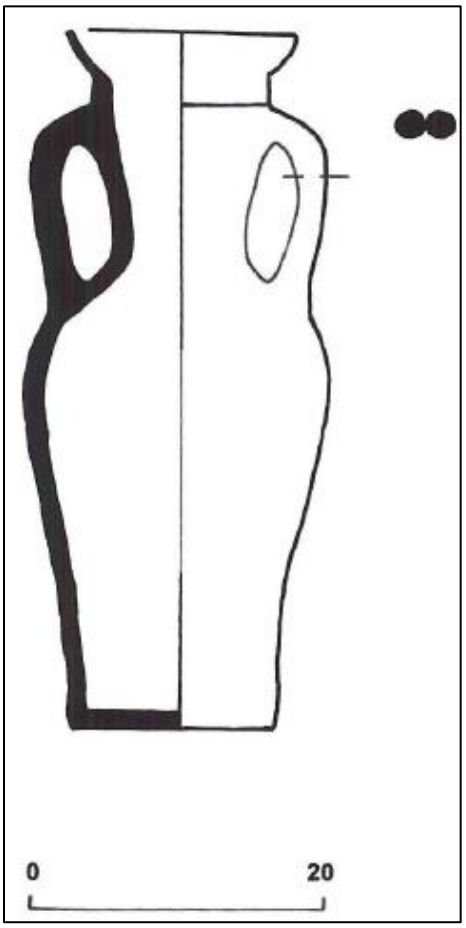
Date:

The dating of this form remains uncertain due to the variety of sources discussing it (Vidal Gonzalez, 1996: 87; Bruno, 2004: 86). In the Punic context of Malta, these urns are generally dated between 410–300 BC (Sagona, 2002: 24).

Place of production and distribution:

This form is widely distributed throughout the Mediterranean, with examples found in Carthage, Sicily, and Ibiza. While it is widely accepted that these urns were locally manufactured, there is no definitive evidence to confirm that Malta was the sole production center. Minor regional variations suggest the existence of multiple production sites (Sagona, 2002: 104).

Drawing and example from the dataset:



FORM 8

FORM 8b

Type:

This form is a variation of Form 8, with only one complete example identified in the studied assemblage. Although smaller in overall size, it retains key characteristics, including an outwardly flaring rim, a convex profile above the flaring base, and looping double cordon handles.

Type Concordance:

- Sagona urn III-IV: 4a-b (Sagona, 2002: 103).
- Xlendi 5 (Azzopardi, 2006: 47).

Measurements:

- Maximum height: Typically ranges from 19 to 35 cm. However, Sagona notes that some urns from marine contexts off the coast of Gozo, such as those from Xlendi, can be significantly larger, reaching up to 65 cm. This is also the case for the examples found at Xlendi Archaeological Park.
- Maximum diameter (upper body): n/a
- Mouth diameter:n/a

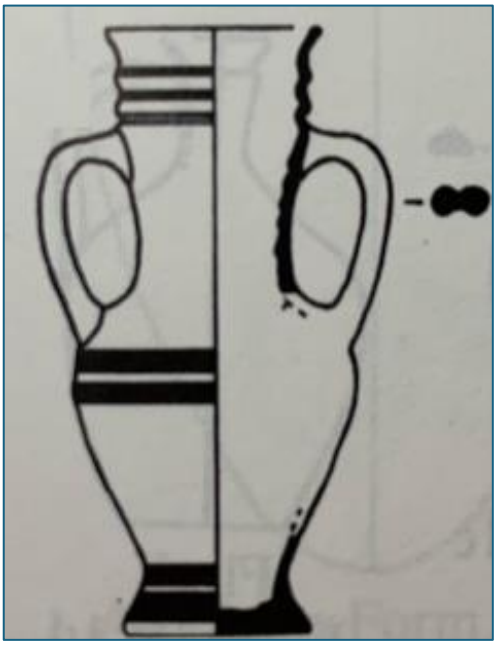
Date:

410-300 BC (Sagona, 2002: 24).

Place of production and distribution:

(see Form 9).

Drawing and example from the dataset:



FORM 8b

FORM 9

Type:

This vessel, identified as an urn of the Sagona type III-IV: 3, is relatively uncommon at the site and can easily be mistaken for examples of Form 11 in photographs. Like Form 11, it features a flat bottom and double handles that connect the neck to the high shoulder. However, this type is distinguished by its slenderer body and longer neck.

Type Concordance:

- Sagona urn III-IV: 3 (Sagona, 2002: 102).
- Xlendi 4 (Azzopardi, 2006:45).

Measurements:

- Maximum height: 30-46cm.
- Maximum diameter (high body): 18-26cm
- Mouth diameter:10-14cm

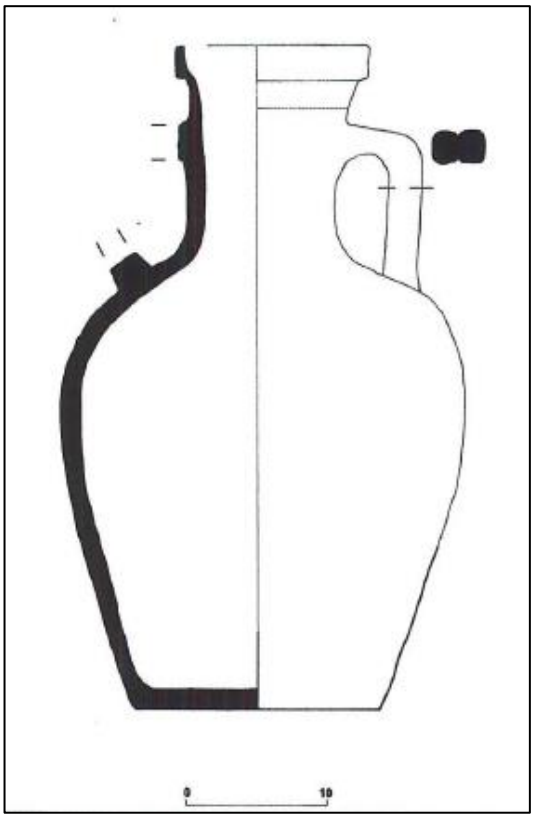
Date:

410-300 BC

Place of production and distribution:

Although similar forms have been found in Tunisia and other North African sites, these vessels are believed to be of local production (Sagona, 2002:102).

Drawing and example from the dataset:



FORM 9

FORM 10

Type:

This form was identified by Elaine Azzopardi during her study of the ceramic assemblages of Xlendi Bay (Azzopardi, 2006: 53). Cylindrical in shape, this amphora is widest at both the upper and lower sections of the body, narrowing noticeably at the center. The handles are small and positioned high above the shoulders, close to the short neck. The base tapers sharply into a short, pointed foot.

Type Concordance:

- Xlendi 9 (Azzopardi, 2006:53).

Measurements:

- Maximum height: 50-65cm.
- Maximum diameter (high body): 25-30cm
- Mouth diameter:16-24cm

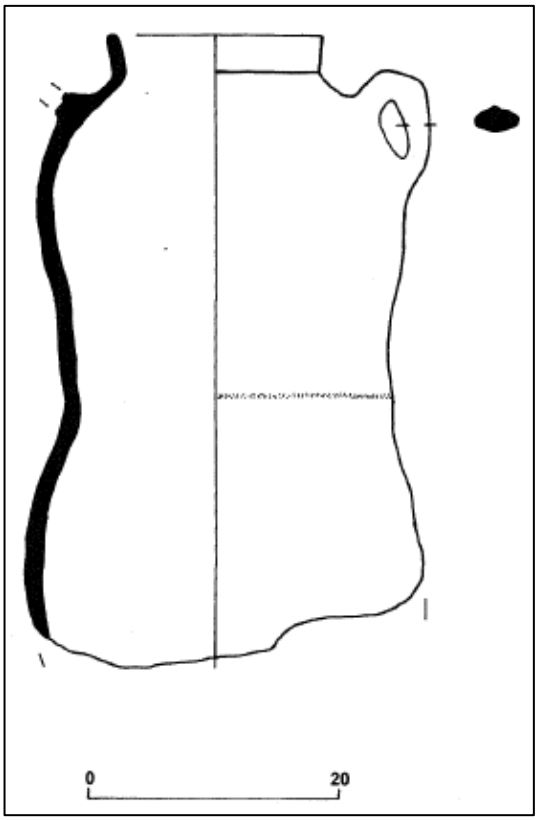
Date:

The general shape suggests a date between the 5th and 4th centuries BC (Azzopardi, 2006: 54).

Place of production and distribution:

Based on morphological characteristics, Sicily and Tunisia have been proposed as potential production centers. However, in terms of distribution, the examples from Xlendi appear to be unique.

Drawing and example from the dataset:



FORM 10

FORM 11

Type:

This form encompasses a large number of similar vessels that are challenging to classify based solely on photographic data due to their close resemblance to one another. These small, flat-based amphorae feature an oval profile, with their widest point at the upper body. They have cylindrical necks and circular handles that attach to the high shoulder just below the flaring rim.

Type Concordance:

- Xlendi 13 (Azzopardi, 2006: 59) and Xlendi 37D.

Measurements:

- Maximum height: 35-45cm.
- Maximum diameter (high body): 25-32cm
- Mouth diameter: 18-20cm

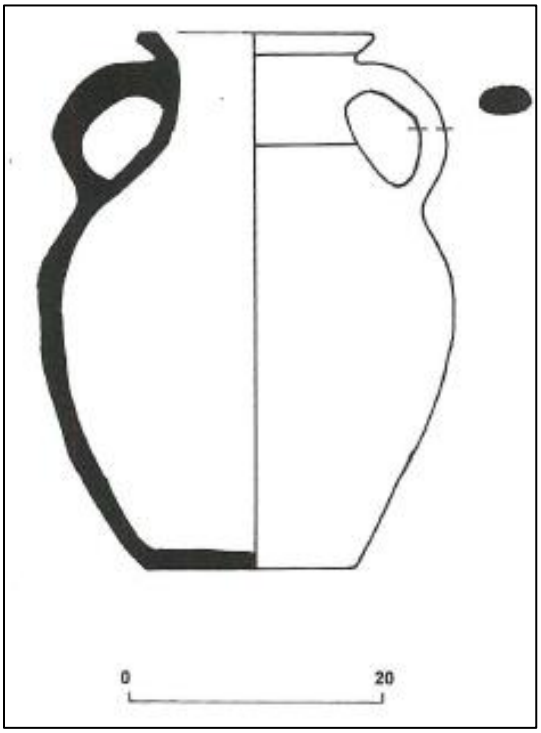
Date:

n/a

Place of production and distribution:

Fabric analysis suggests production in Sicily, North Africa, and Malta (Azzopardi, 2006: 59).

Drawing and example from the dataset:



FORM 11

FORM 11b

Type:

Closely resembling Form 11, these small, flat-based vessels are distinguished by their reduced size, conical neck, smaller handles, and smooth rim that opens into a wider mouth. Azzopardi identified up to four different variations of this same shape showing slight differences in all these features as well as the proportions of the oval shape they all share. She was able to tell them apart from physical examples recovered from Xlendi, but was not able to use those differences to establish parallels in other types or interpret information about their origin and distribution. In our case, these slight differences are not pronounced enough for our detection models to have a chance at identifying them given their relatively low presence at Xlendi Archaeological Park.

Type Concordance:

- Xlendi 37a-d (Azzopardi, 2006: 96).

Measurements:

- Maximum height: 35-45cm
- Maximum diameter (high body): 24-28cm
- Mouth diameter: 10-15cm

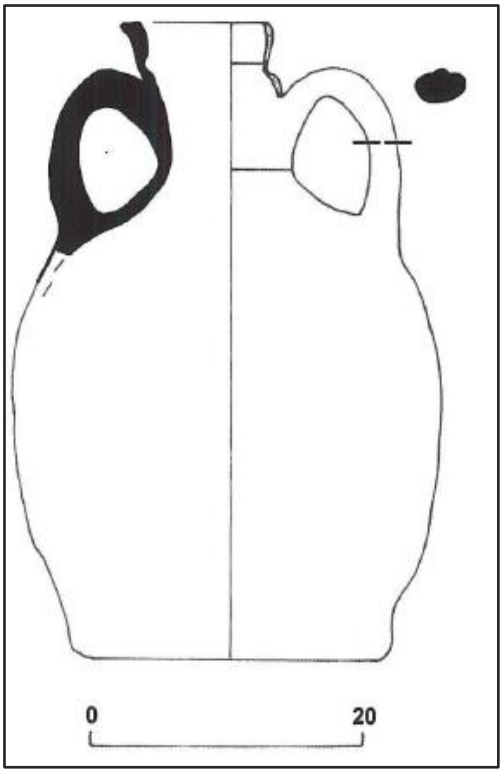
Date:

n/a

Place of production and distribution:

n/a

Drawing and example from the dataset:



FORM 11b

FORM 12.

Type:

Similar to Form 12, this category includes a group of vessels that are difficult to distinguish from one another when analyzing raw photographic data from maritime environments. These vessels have oval-shaped bodies and are generally classified as jars due to their flat bases, flaring necks suited for pouring, and singular wide handles that connect the shoulder to the rim—features also characteristic of jugs. Their design suggests they may have been used as shipboard items.

Type Concordance:

- Xlendi 36A (Azzopardi, 2006:). Xlendi 36B.

Measurements:

- Maximum height: 20-25cm
- Maximum diameter (high body): 12-18cm
- Mouth diameter: 10-14cm

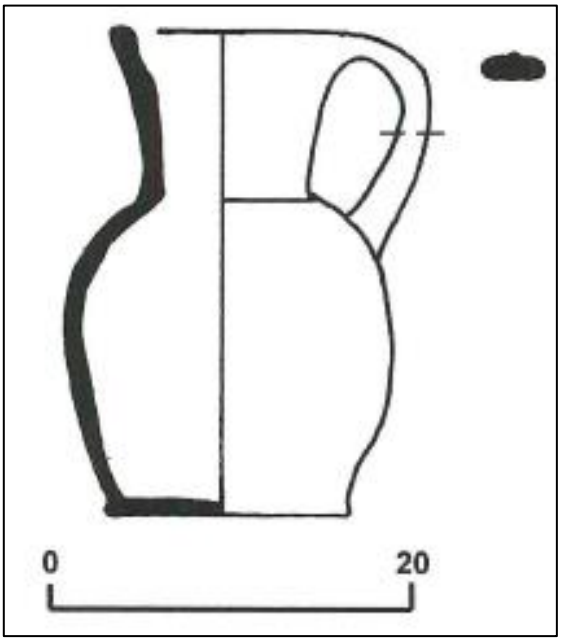
Date:

n/a

Place of production and distribution:

n/a

Drawing and example from the dataset:



FORM 12

FORM 13

Type:

Several examples of this vessel type are visible across the site, though their small size makes their identification a challenging one for the detection models. These flat-based, oval-shaped vessels feature relatively long, conical necks with a smooth rim. Their oval handles connect the shoulder to the neck just below the rim.

Type Concordance:

- Xlendi 35 (Azzopardi, 2006:45).

Measurements:

- Maximum height: 14-20cm
- Maximum diameter (handles): 10-12cm
- Mouth diameter: 5-8cm

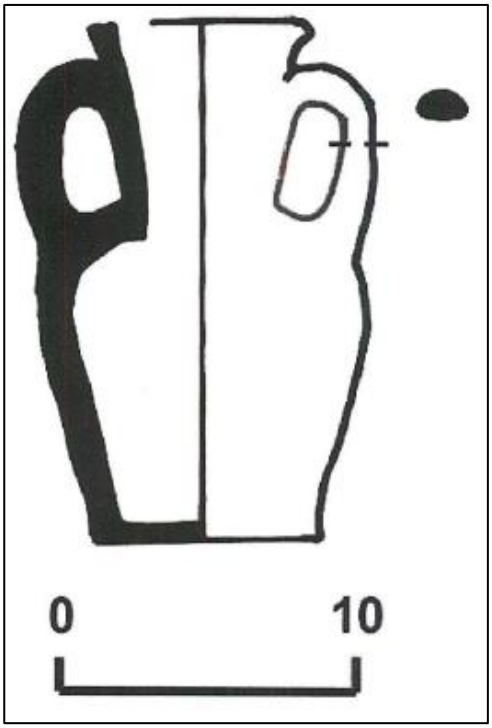
Date:

n/a

Place of production and distribution:

n/a

Drawing and example from the dataset:



FORM 13

FORM 14

Type:

Only two examples of this form have been identified at the site. These are globular amphorae with thick, sharply bent handles and a short neck, classified as Dressel 20—a type primarily used for transporting olive oil. Although Dressel 20 amphorae were mainly distributed in Spain, Atauz identified examples in remotely operated vehicle (ROV) images captured during the Xlendi survey of 2004 (Atauz, 2004: 380).

Type Concordance:

- Dressel 20 (Dressel, 1899).
- Beltrán 5 (Beltrán, 1970: 471).
- Atauz type 5 (Atauz, 2004: 380).
- Peacock and Williams 25.

Measurements:

- Maximum height: 75-100cm
- Maximum diameter (high body): 60-67cm
- Mouth diameter:8-10cm

Date:

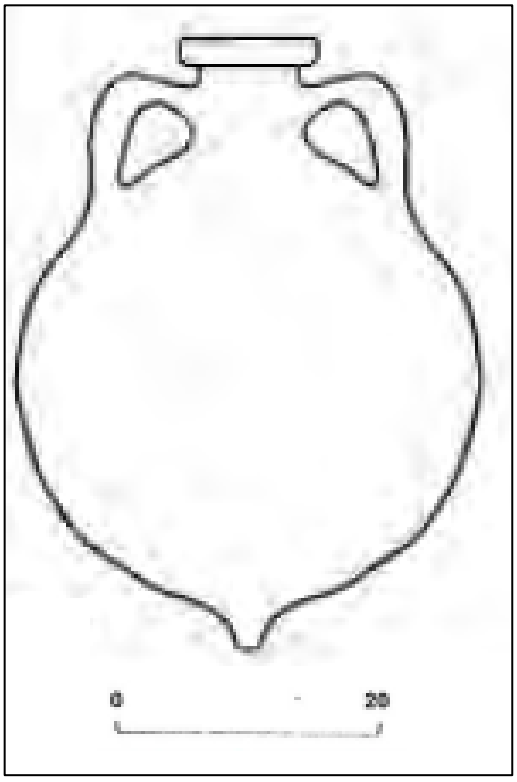
1st to 2nd century AD.

Place of production and distribution:

These amphorae were produced at multiple centers along the Spanish Levantine coast, including Oliva, Almadrava, and Catarroja, the banks of the river Guadalquivir, and the Roman provinces of Baetica (Santonia, 1991).

In terms of distribution, Dressel 20 are one of the most commonly distributed amphorae in the western Roman provinces, including the British Isles (Peacock and Williams, 1986).

Drawing and example from the dataset:



FORM 14

Appendix IV

Model Evaluation. Metrics

Contents

Each model generates both numerical and visual metrics as output. Visual metrics typically plot numerical data, aiding in interpretation and allowing for a quick assessment of a model's validity before conducting further tests.

To thoroughly evaluate a model's performance, numerous metrics must be considered, each reflecting different aspects of the model during training. In this dissertation, we have utilized these metrics to design, develop, and refine our models effectively. With a comprehensive list of these metrics already available for reference in the YOLO model's online repository,[155] this section will focus on those—both numerical and visual—that are most essential or useful for the interpretative process.

1 Detection Metrics

On output, each model generates numerical and visual metrics. The visual ones usually plot the numerical ones to help with interpretation. Before we see those, we must first look at another general set of metrics that are necessary to explain the significance of those produced by the output:

-Intersection over Union (IoU): IoU measures the overlap between a predicted bounding box and the ground truth bounding box (input by a human). It is a fundamental metric for evaluating object localization accuracy and is particularly important when precise object positioning is crucial.[156]

[155] For even more information on the metrics used by YOLO and how to interpret them: https://docs.ultralytics.com/guides/yolo-performance-metrics/.
[156] Explanation of the concept behind Intersection over Union (p.164).

-Precision and Recall: Precision quantifies the proportion of true positives among all positive predictions, assessing the model's ability to minimize false positives. On the other hand, Recall measures the proportion of true positives among all actual positives, evaluating the model's ability to detect all instances of a given class.

When deploying a trained model, selecting an appropriate confidence threshold is essential to balance precision and recall (Paraskevas et al., 2023:5). This threshold can be adjusted based on the specific detection task. From an archaeological perspective, A higher confidence threshold should be used when the goal is to verify archaeologists' findings. This increases precision but may result in more false negatives. A lower threshold is preferable when the objective is to alert archaeologists to potential discoveries. This maximizes recall, detecting more possible findings at the expense of increased false positives.

-Average Precision (AP): AP computes the area under the precision-recall curve, providing a single value that encapsulates the model's precision and recall performance.

-Mean Average Precision (mAP): mAP extends the concept of AP by calculating the average AP values across multiple object classes. This is useful in multi-class object detection scenarios to provide a comprehensive evaluation of the model's performance. Suitable to conduct a broad assessment of model performance.

1.1 Output Numerical Metrics

Each model produces both numerical and visual metrics. Visual metrics typically represent the numerical data graphically, making interpretation easier. However, before analyzing these, it is important to first examine a broader set of fundamental metrics that provide context for understanding the significance of the output results.

-Precision (Box(P)): The accuracy of the detected objects, indicating how many detections were correct. It is important when minimizing false detections is a priority.

-Recall(R): The ability of the model to identify all instances of objects in the images. It’s vital when it is important to detect every instance of an object.

-mAP50: Mean average precision calculated at an intersection over union (IoU) threshold of 0.50. It is a measure of the model's accuracy considering only detections made with a confidence rating superior to 50%.

-mAP50-95: The average of the mean average precision calculated at varying IoU thresholds, ranging from 0.50 to 0.95. It gives a comprehensive view of the model's performance across different levels of detection difficulty.

-Speed: The speed of inference can be as critical as accuracy, especially in real-time object detection scenarios. This section breaks down the time taken for various stages of the validation process, from preprocessing to post-processing.

1.2 Output Visual Metrics

The evaluation of a model also yields visual outputs that provide an intuitive understanding of the model's performance. Each of these is valuable for a variety of things:

-Plotted general metrics (Figure 57): This composite graph illustrates the evolution of the model's key metrics (Y-axis) throughout the training process,[157] measured in epochs (X-axis).[158]

-Precision-Recall Curve (Figure 58): A crucial visualization for classification tasks, this curve illustrates the trade-off between precision and recall across different confidence thresholds. It is particularly valuable when handling imbalanced classes, as it helps assess the model's ability to balance false positives and false negatives.

-Normalized Confusion Matrix (Figure 59): The confusion matrix offers a detailed breakdown of model performance, displaying the proportional counts of true positives, true negatives, false positives, and false negatives for each class.[159] In practice, it helps the user understand how well a model is predicting different classes by comparing the model's predictions with the ground truth labels introduced during training.

---

[157] For the training to be successful according to the **general plotted metrics**, the precision, average precision, and recall graphs should display an upward trend that stabilizes toward the end, indicating the model's refinement. Conversely, other metrics should exhibit a declining trend. These visualizations provide an immediate indication of whether the model is performing as expected. The training process of a model is outlined in p.24.

[158] Epochs (p.16).

[159] A perfect **confusion matrix** will have all the intense colors on the downwards diagonal line from left to right (Shinde et al., 2018: 836; Fan, 2025).

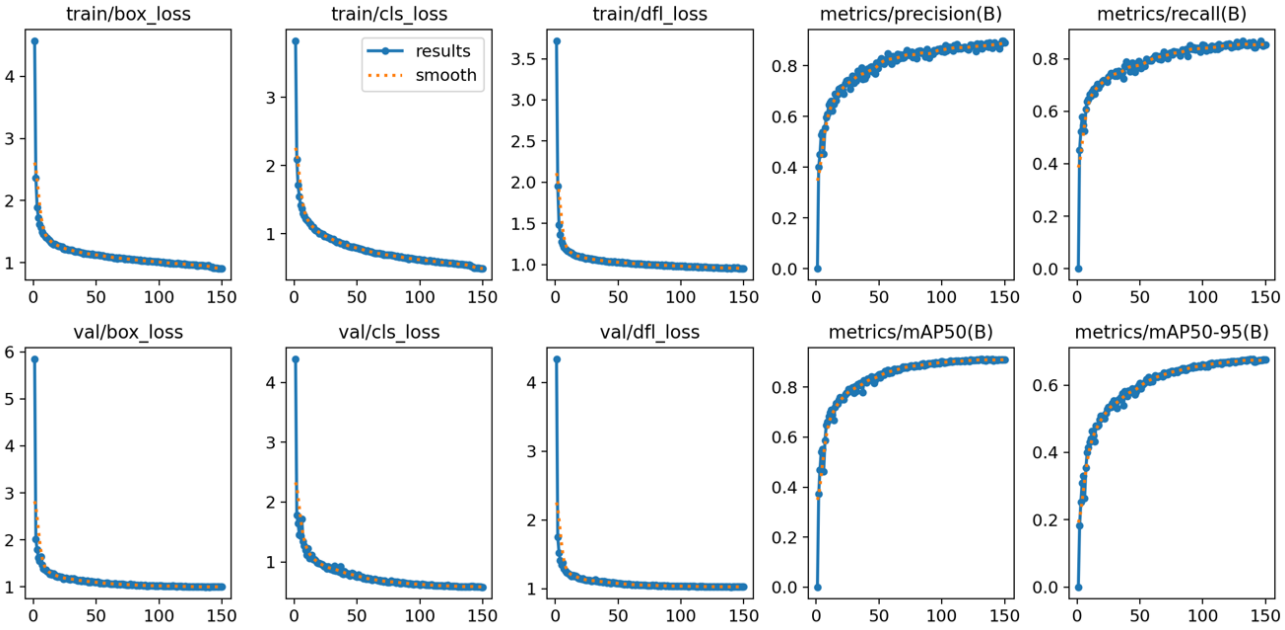


Figure 57. Example of plotted metrics of an S1 model.

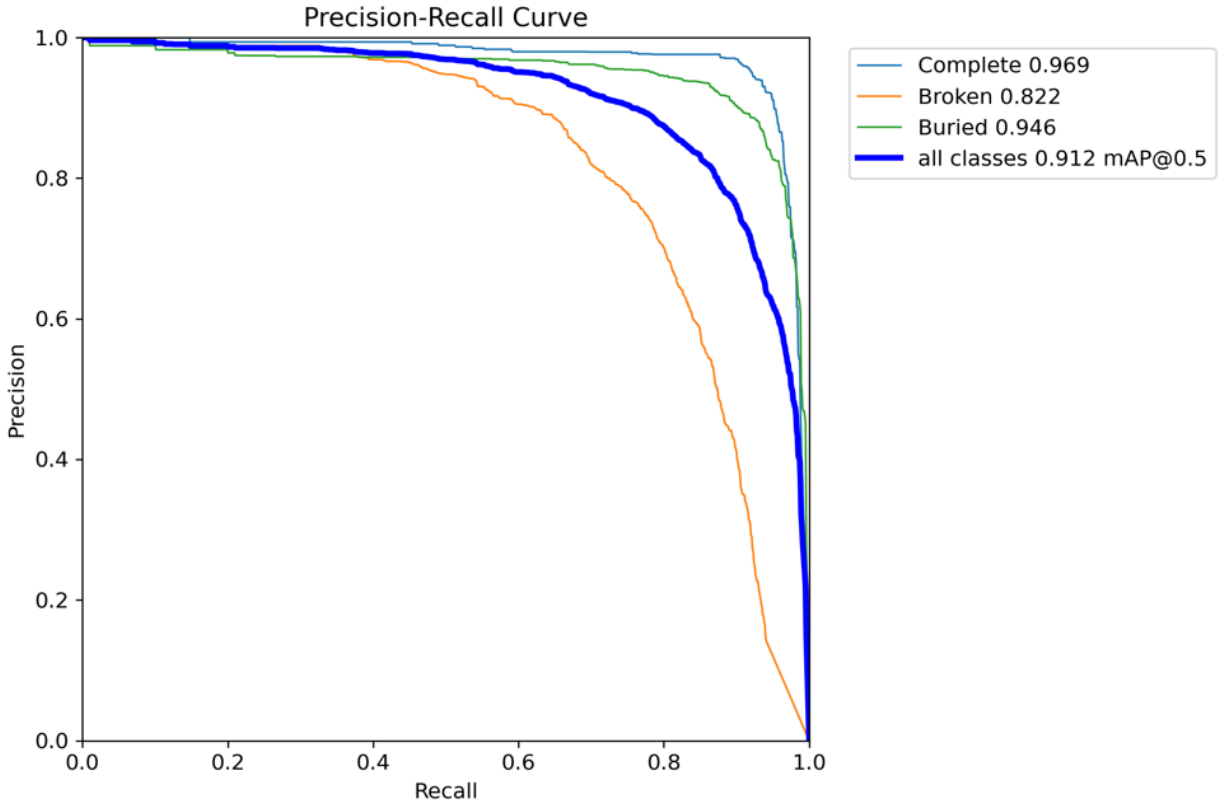


Figure 58. Example of Precision-Recall Curve for an S1 model.

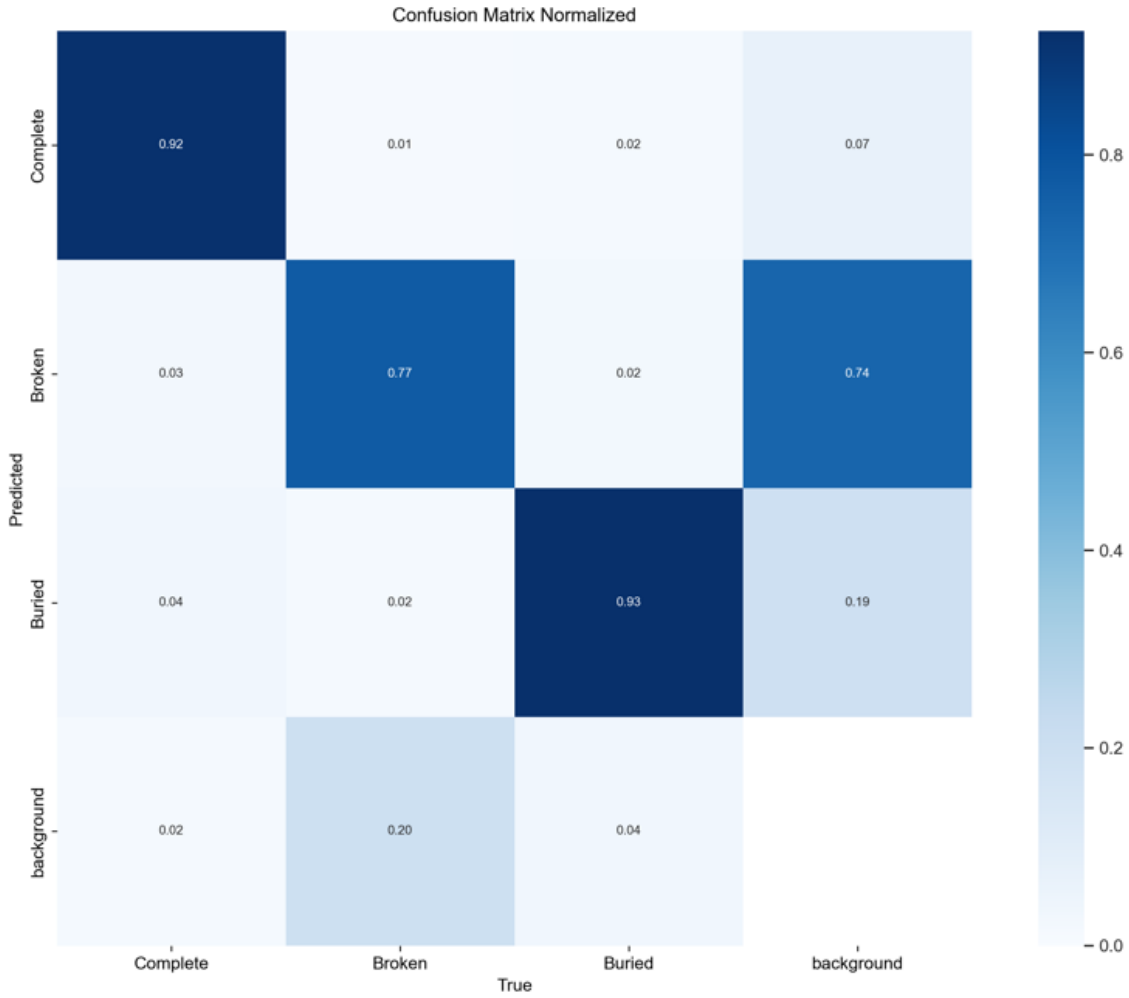


Figure 59. Example of the Normalized Confusion Matrix for an S1 model.

Appendix V

Test Videos

As part of the process of visual evaluation, 12 out of the 72 models designed during this dissertation were evaluated for archaeological validity.[160] As an example, the videos corresponding to the most representative models from each PCI (N1, S1 and T4) are presented here as examples through Google Drive links.

N1 Evaluation video:

https://drive.google.com/file/d/1mCHkRRWpJN0MpkTtoR7DUp7YnAwePble/view?usp=sharing

S1 Evaluation video:

https://drive.google.com/file/d/16RwMTem6olQgPbBLRs_FDcaVXU_7VEvy/view?usp=sharing

T4 Evaluation video:

https://drive.google.com/file/d/1qemd9BktoKQdiK5ntOXWIMARIoKf9dRU/view?usp=sharing

[160] Model validity (p.63).

Appendix VI

Examples of Exportability

Contents

The examples have been chosen for the sake of variety. Some correspond to shipwrecks excavated more than 25 years ago and others to ongoing projects with no publications. The ideas presented here answer to hypothetical scenarios established to illustrate concrete situations for the assessment of utility and exportability of the test models developed at Xlendi Archaeological Park, and therefore the conclusions we can extract from them are experimental and not to be generalized in any way.

1 The Uluburun shipwreck

Excavated from 1984 to 1995 in Uluburun, Turkey, this single shipwreck dated to 1350 BC lay scattered across a slope with rocks and sand pockets at around 45m of depth. It carried a very varied cargo of raw materials like ingots, ceramics, glass, weapons, and other items like tools, jewels, and miscellaneous, thus serving as a prototypical ancient wreck—an assemblage of diverse materials arranged in the general shape of the hull, with additional dispersed elements due to site morphology (Bass, 1986; Pulak, 1988; Pulak, 1998, Figure 60).

Given Uluburun's very distant chronological, material and geographical similarity to the source of our models, we would struggle to directly implement any of them without adaptations. This is especially true for more complex analyses. [161] The varied cargo at Uluburun makes the concept of nature models particularly valuable, but it also means that our models, trained at a site with nothing but ceramic, would not work. While it is true that our N1 models could isolate ceramic materials,

[161] As stated before, the models trained for this dissertation have a testing purpose. All of them could be refined for increased robusticity (which in turn means exportability) and precision.

determining the total number of ceramic pieces using the site's NISP[162] while plotting density of material, N2 models would have niche applications, and N3 models would refine N1's analysis by providing an exact count of ceramic vessels through the MNV,[163] they would all be very disrupted by the presence of other alien materials like ingots and glass and weapons. The same goes for our N4 models, which identify only plastic and ceramic materials. They would not be directly applicable here either.

Like we said, this does not mean that the concept of nature models is useless in Uluburun. Quite the opposite in fact. Theoretically, more developed PAI (like a hypothetical N5, for instance) could be trained to function like N1 and N4 but with the added ability to identify all site's materials—including ingots, wood, and weapons—alongside ceramics. The problem is that if any detection models are to be developed for shipwrecks as **unique** as Uluburun, it is going to be hard to have a similar site to function as a base training site.[164] They would have to be developed and trained at the site of Uluburun itself.

While the case is the same for state models, there is an additional thing to consider: State models are not necessary at a site like Uluburun. This is because state models are often useful to evaluate site formation processes through the reflection these have on the preservation of materials. In the case of Uluburun—any most other single shipwrecks—site formation can be understood without them. However, the potential still exists to develop additional state PAI on sites like Uluburun. For instance, a hypothetical S5 model could assess the condition of a boat's wooden elements, offering insights into their preservation and rate of degradation.

As for typological models, if we were to successfully train valid ones at Xlendi Archaeological Park, the chronological and geographical differences between the sites would make it impossible for us to export them to a site like Uluburun without modifications. Alternatively, we could develop a dedicated set of typological models for Uluburun, though the effort required might not justify the benefits in this case. Nevertheless, such work could prove valuable for future applications to similar sites.

The Uluburun shipwreck is an underwater archaeological site of such exceptional uniqueness that it falls beyond the scope of the models we could develop at Xlendi Archaeological Park.

[162] Number of identified specimens (p.53).
[163] Minimum number of vessels (p.53).
[164] Base training site (p.125).

Currently, no other shipwrecks exist that could serve as training data for these models. As a result, using detection methods at sites like Uluburun—and similarly unique discoveries in the future—would probably only be feasible if detection methods were specifically trained and developed for each individual case. This site is also an example of how different PCI would or would not be of use on a case-to-case basis. In the scenario proposed by this exercise, for instance, state models would not be very valuable.

a)

b)

Figure 60. a) Excavation of Uluburun**.** b) Reconstruction of the Uluburun assemblage (from INA.org).

2 Fourni Islands

The Fourni archipelago is known as the shipwreck capital of the world, with a presence of 53 ancient shipwrecks spanning over 2500 years on an area key to the trade routes between the Black Sea and Alexandria. As a result, the whole area can be considered a huge assemblage that offers a contrast with the other examples by representing a survey area rather than a single wreck (Murray, 2018; Figure 61). This shift from traditional shipwrecks alters the role of detection models and presents new analytical opportunities. Additionally, some of the models developed for testing in this study in Xlendi Archaeological Park could potentially be implemented directly with success because of this reason.

In terms of nature models, because of their simplicity, our models could be applied directly in Fourni despite their test-stage development. All from N1 to N4 models could function as survey tools on diver-held cameras or underwater vehicles with minimal adaptation. Given the survey context, their real-time capabilities would also be practical in this case. For example, nature models could be used to plot density maps, identify areas of interest, and generate statistical data on the site's activity over time, offering benefits similar to those observed at Xlendi Archaeological Park.[165] Further nature PAI could be developed to include other types of materials present in the survey area as well.

All our state models could be applied in Fourni directly too, offering real-time information of pattern deposition based on state of preservation with **at least** all the other benefits that applied for Xlendi Archaeological Park.[166] For instance, plotting the results on a chart would yield valuable information on the site formation and post depositional processes that have affected materials across time at different areas of the site, or help us identify the presence of more shipwrecks around the archipelago.

The typological models designed for our site, however, even if they worked, would not be applicable to Fourni without an adaptation process. This is because of the chronological and typological difference that is to be expected between Xlendi Archaeological Park (3rd century BC to 2nd century AD) and Fourni itself, which presents materials up to the Byzantine period. In this regard, while typological trained models would not be compatible between both sites, Fourni itself

[165] Possible uses of nature models (p.95-98).
[166] Possible uses of state models (p.101-104).

is large and varied enough to become a base training site[167] on its own, one that would cover the training of typological models for the entire area and its chronological and cultural spheres.

As for the models themselves, they would be of great use in a place like Fourni. They would provide at least the benefits mentioned in our case. For instance, they could be used to map the whole survey site and instantly characterize and date shipwrecks based on their typologies, better understand the affluence of trade across the archipelago, generate statistical data on trade relations across time, and much more. [168]

[167] Base training site (p.125).
[168] Possibles uses of typological models (p.109-113).

Figure 61. Assemblage of Fourni Islands (from Greece-is.com).

3 The Slope I Shipwreck

The Slope shipwrecks are two archaeological sites recently discovered at 1500m of depth in the South China Sea and dated between 1368 and 1655 AD. Respectively, their cargoes consisted of thousands of porcelain objects for Slope I and wooden logs for Slope II.[169] Slope I was chosen as an example to highlight how our models could be adapted to an assemblage outside the Mediterranean (Figure 62). With the site sharing a widespread distribution of materials similar to that of Xlendi Archaeological Park, and given its depth of 1,500 meters, the real-time capabilities of YOLO models would be highlighted in what would exclusively be object detection through 2D imagery acquired with submersibles or other underwater vehicles.

In this scenario, while the core concepts behind our nature, state, and typological PAI remain relevant, significant refinements and retraining would be required to accommodate the site's distinct material composition and any possible variations resulting from the change in maritime environment (from the Mediterranean to the South China Sea). For example, porcelain vessels have very different shapes than Mediterranean amphorae and, as we can see in Figure 63, they were stacked in different ways as what we normally encounter on Mediterranean assemblages. In addition, porcelain, because of its manufacturing characteristics and the nature of kaolin as the finest of clays, preserves its polychrome nature even after centuries underwater. Including color as one of PAI differentiating factors is an example of how this difference in material could influence our training capabilities.

Developing tailored models for Slope I would improve detection accuracy, lead to new applications in detection methodology, and be a first step in the generation of exportable models for this particular archaeological context. In other words, like Fourni in the Aegean, Slope I could become a base training site for porcelain in the South China Sea for the 15-16th century period.[170]

[169] There are no scientific publications of meaning on the shipwrecks yet. More information can be found at the news reports submitted by the Chinese Academy of Sciences: https://www.ecns.cn/news/2023-05-22/detail-ihcprrta3249639.shtml).
[170] Base training sites (p.125).

Figure 62. Porcelain assemblage of the Northwest Continental Slope I Shipwreck of the coast of Hainan in the South Chine Sea (from ecns.cn/news).