Universiteit
Leiden
The Netherlands

Ancient Storage and AI: Automated object detection of prehistoric granaries on archaeological GIS maps: A Deep Learning approach

Penterman, Merel

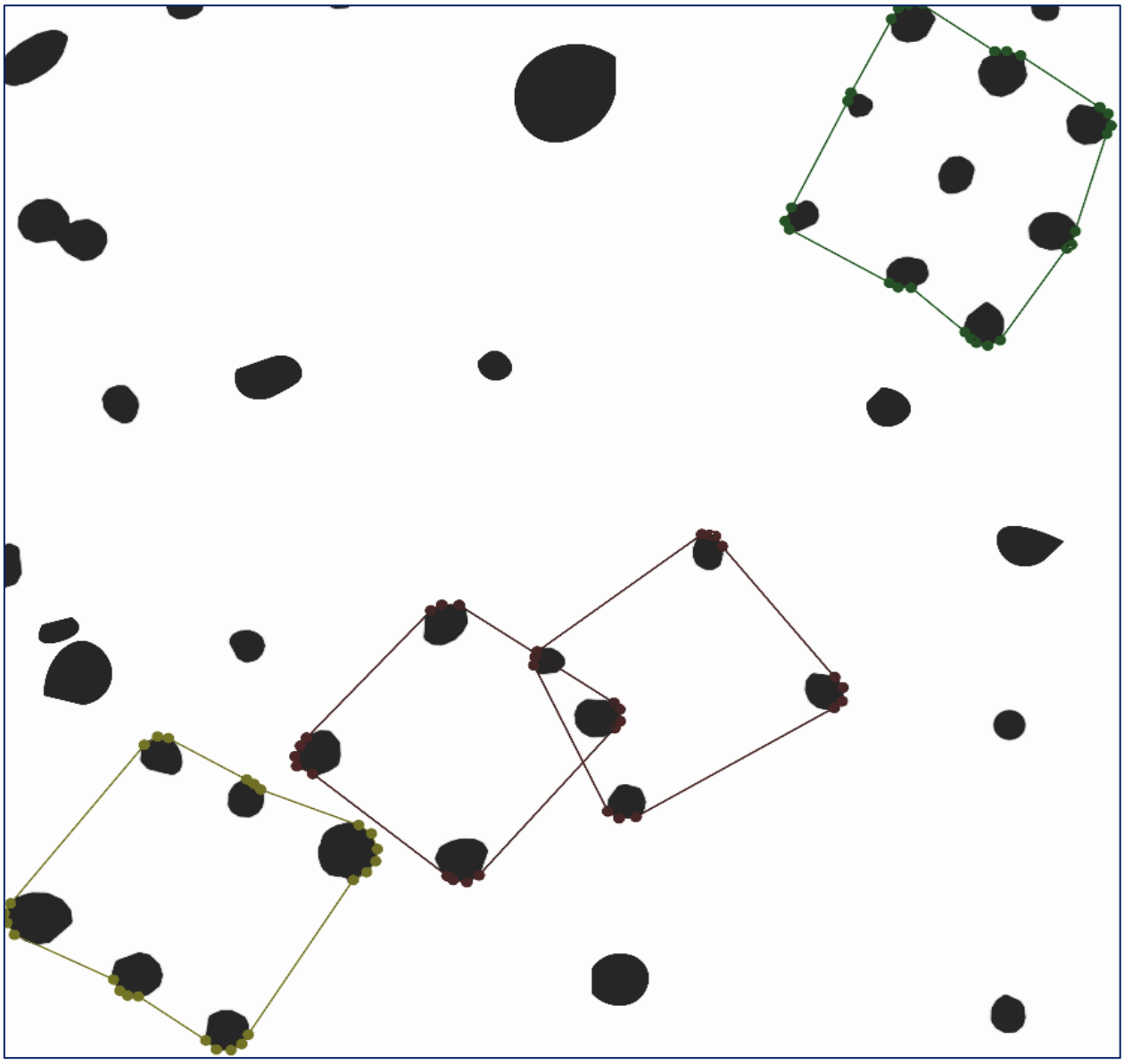
Citation
Penterman, M. (2025). *Ancient Storage and AI: Automated object detection of prehistoric granaries on archaeological GIS maps: A Deep Learning approach*.

Version:	Not Applicable (or Unknown)
License:	License to inclusion and publication of a Bachelor or Master Thesis, 2023
Downloaded from:	https://hdl.handle.net/1887/4195042

Note: To cite this publication please use the final published version (if applicable).

Ancient Storage and AI

Automated object detection of prehistoric granaries on archaeological GIS maps: A Deep Learning approach



RMSc Thesis
M. Penterman

Universiteit Leiden
The Netherlands

Ancient Storage and AI

Automated object detection of prehistoric granaries on archaeological GIS maps: A Deep Learning approach

RMSc Thesis
Merel Penterman – S1616196
Prof.dr K. Lambers
Leiden University, Faculty of Archaeology
13 December 2024

Universiteit Leiden
The Netherlands

Abstract

This thesis explores the application of Deep Learning techniques for automated feature detection within GIS maps in the context of digital archaeology. Specifically, it focuses on leveraging the YOLOv8s algorithm to automate the detection of prehistoric granaries on archaeological excavation maps. Traditional manual analysis methods in archaeological research are often time-consuming and labour-intensive, particularly when dealing with large spatial datasets. Moreover, the overall convoluted nature of archaeological excavations and the diverse range of features they contain present significant challenges for traditional methods. To address these challenges, this research investigates the potential of Deep Learning algorithms to enhance the efficiency and accuracy of automated feature detection on archaeological GIS maps.

The results of this study demonstrate the effectiveness and potential of Deep Learning algorithms to accurately identify prehistoric granaries within archaeological excavation maps. The analysis reveals that the algorithm is able to detect and classify prehistoric granaries with a relative high degree of precision. Despite these promising results, the study underscores the challenges associated with the opacity of DL models, particularly regarding their interpretability and biases. The thesis highlights the importance of addressing issues such as data imbalance, background noise, and the inclusion of contextual information to improve the accuracy and reliability of automated detection models. While the current model demonstrates potential, further research is needed to refine these methodologies, ensuring they contribute meaningfully to archaeological analysis. This work tries to lay some foundation for future advancements in the field, advocating for the development of more comprehensive DL models that can enhance the efficiency and depth of archaeological investigations.

Acknowledgements

I would like to extend my sincere appreciation to Prof.dr K. Lambers for their supervision and guidance throughout the research process. Their insightful feedback and mentorship have been instrumental in shaping the direction and quality of this study.

Furthermore, I am grateful to Archol bv for providing access to their comprehensive datasets, which was a crucial resource for this research. Their collaboration and support have greatly enriched the depth and scope of the study. Correspondingly, a special thanks to T. Hamburg for their guidance and assistance during the course of this project, and M. Steenbakker for their research ideas. Their expertise have been invaluable in navigating the practical complexities of this research.

Finally, I would like to extend my gratitude to all individuals who have contributed to this research in various capacities.

Table of Contents

Tables and Figures

Abbreviations

ABM	Agent-based modelling
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area under the curve
CNN	Convolutional Neural Network
DL	Deep Learning
IoU	Intersection over Union
GIS	Geographic Information Systems
GPS	Global Positioning System
LiDAR	Light Detection and Ranging
mAP	Mean Average Precision
ML	Machine Learning
OBIA	Object Based Image Analysis
PR	Precision-recall (curve)
TS	Total Station
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once (an object detection algorithm)
XAI	Explainable Artificial Intelligence

1. Introduction

1.1 Archaeology in the digital age

In recent years, the field of archaeology has undergone a profound methodological transformation. As can be seen in the examples below, a quick glance at contemporary literature highlights the upcoming importance of terms such as “Big Data”, “quantitative data analysis”, and “digital archaeology”. Facilitated by the advancements in both hardware and software, these digital methodologies have revolutionised archaeologists’ ability to rapidly gather and process vast quantities of data. While traditional methods rely heavily on excavation and manual analysis, and are therefore often limited by time, resources, and their inherent destructive nature; digital archaeology aims to complement these methods through non-invasive data collection and comprehensive analysis. As a result, digital archaeology encompasses a wide range of practices, including, but not limited to: photogrammetric reconstructions of artefacts and sites (e.g. Lercari, 2017; Nikolakopoulos *et al.*, 2017) archaeological virtual reality (e.g. Forte *et al.*, 2011; François *et al.*, 2021), archaeogaming (e.g. Blakely, 2023; Rassalle, 2021; Reinhard, 2018; Winter, 2021), LiDAR technology for documenting sites (e.g. Campana, 2017; Risbøl, 2013; Smith *et al.*, 2014), the use of detection algorithms that identify sites from aerial data (e.g. Papadopoulos *et al.*, 2019; Verschoof-van der Vaart, 2022) network and trade analyses (e.g. Isaksen, 2013; Knappett, 2020), and agent-based modelling (e.g. Campillo *et al.*, 2012; Davies *et al.*, 2019; Kowarik *et al.*, 2012).

1.1.1 Archaeology and ‘Big Data’

Within this digital landscape, the emergence of Big Data has been a driving factor of these new methodologies. A popular concept that is widely employed within many, if not all, scientific disciplines. Originally the term was ‘[c]oined in the 1970’s to refer to datasets that were too weighty to process with existing computing resources’ (VanValkenburgh & Dufton, 2020, p. 1). However, nowadays the concept has evolved to the point where we *can* actually analyse these datasets with increased computational power, making it possible to extract insights from this data through algorithmic and statistical analysis. In general, “Big Data” refers to the systematic collection, management, and analysis of these large and complex datasets. Still, ‘(...) the term lacks any fixed scalar or

dimensional definition' (Ibidem), which makes it difficult to ascertain whether particular datasets are "big" or "small". Some argue that the extent implied by the term can be measured through dimensions like volume, velocity, and variety (Gattiglia, 2015, p. 115). Adding to this complexity, the threshold for what constitutes Big Data varies across disciplines. In archaeology, datasets are often fragmented or incomplete, so the amount of data needed to qualify as "big" is significantly smaller compared to fields (Wesson & Cottier, 2014; Moscati, 2021). Therefore, this relativity is key: archaeological datasets on their own may not fit conventional definitions of Big Data. However, when integrated with large-scale datasets—such as geospatial or environmental data—they contribute to Big Data analysis within an archaeological context. Nevertheless, the question regarding the precise amount of data required has become redundant in contemporary discussions, as the term now predominantly conveys the presence of an associated methodological and theoretical framework.

On a methodological level, as illustrated in the examples above, Big Data archaeology has opened up new ways of tackling research questions. Especially '[t]he recent trends in archaeological practice towards datafication and digitalisation (...), the increased public availability of archaeological and remotely-sensed data, and the developments and decrease in cost of computing power and storage (...)' (Verschoof-van der Vaart, 2022, p. 5) have been the driving factor for the rapid adoption of Big Data methodologies within archaeology. These approaches allow archaeologists to uncover patterns, correlations, and insights that were previously elusive using traditional methods alone. One example is the broad-scaled application of Geographic Information Systems (GIS) to manage and analyse spatial data; enabling researchers to explore the relationships between archaeological sites, environmental factors, and human behaviour. For instance, by using a Global Positioning System (GPS) throughout an excavation, features, artefacts, and other relevant data can be measured with a high spatial accuracy, and later integrated into this GIS system. This integration quickly streamlines the analysis process, allowing for the data to be readily examined, adapted, and understood within the broader archaeological context. All in all, the use of this has improved the accuracy and ease of archaeological site mapping, which is crucial to archaeological research.

Theoretically speaking, as pointed out by VanValkenburgh & Dufton (2020), the Big Data phenomenon has initiated a transformation in the broader outlook on data and science. The framework places greater emphasis on uncovering simple correlations rather than causality, leading to a reorientation in our understanding of what data represents and its potential applications. Thus, although '[t]he centrality of data to archaeological knowledge has always been the case, the burden and expectations placed upon data are subtly shifted in a Big Data paradigm' (Huggett, 2020, p. 9). As a result, this new data paradigm is both embraced and met with heavy resistance. Some see opportunity for groundbreaking innovations, whereas others have concerns regarding the uncertainties this paradigm brings (outlined by Zubrow, 2006; Huggett, 2015). The latter predominantly substantiated by the accused undertheorised nature of the Big Data approach, as digital methodologies are often applied without thorough consideration of their implications, biases, and limitations (Cowley *et al.*, 2021). Still, despite these ongoing theoretical issues and debates (which will be expanded upon in chapter 4), the reality remains that digital archaeology and Big Data are becoming firmly entrenched in contemporary archaeological research, necessitating continued engagement and adaptation, all while acknowledging the opportunities they can offer to the discipline.

1.1.2 Artificial intelligence and automated detection

One prominent methodological framework popularised by the Big Data paradigm is that of Artificial Intelligence (AI). AI's capabilities of processing datasets and recognising patterns have influenced the way in which archaeologists can research the past. For instance, by leveraging the possibilities of AI, researchers have been able to identify archaeological sites on a large-scale, such as Celtic fields systems (Mallick, 2021), burial mounds (Verschoof-van der Vaart, 2022), road systems (Li *et al.*, 2016; Verschoof-van der Vaart & Landauer, 2021), and other structures that have eluded human eyes or traditional archaeological methods. However, similar to the aforementioned Big Data discussion, AI's applications have been met with both support as well as resistance. Even though the technique has been around for some time, '(...) the acceptance and use of these computational approaches were initially limited, since archaeology tends to rely on its own domain-specific toolbox' (de Laet & Lambers, 2009, as cited in Lambers *et al*, 2019, p. 2).

Fundamentally, artificial intelligence can be defined as '[t]he theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, and decision-making' (Oxford Language Dictionary, n.d.). In this thesis, the predominant approach will involve the use of Deep Learning techniques applied to problems in the field of Computer Vision. More specifically, this research will leverage object detection algorithms suitable for archaeological feature recognition. Computer vision based automated detection can be categorised in four distinct approaches: (1) template matching-based object detection, (2) knowledge based object detection, (3) object based image analysis (OBIA), and (4) Machine Learning based object detection (Cheng & Han, 2016, p.12) (figure 1).

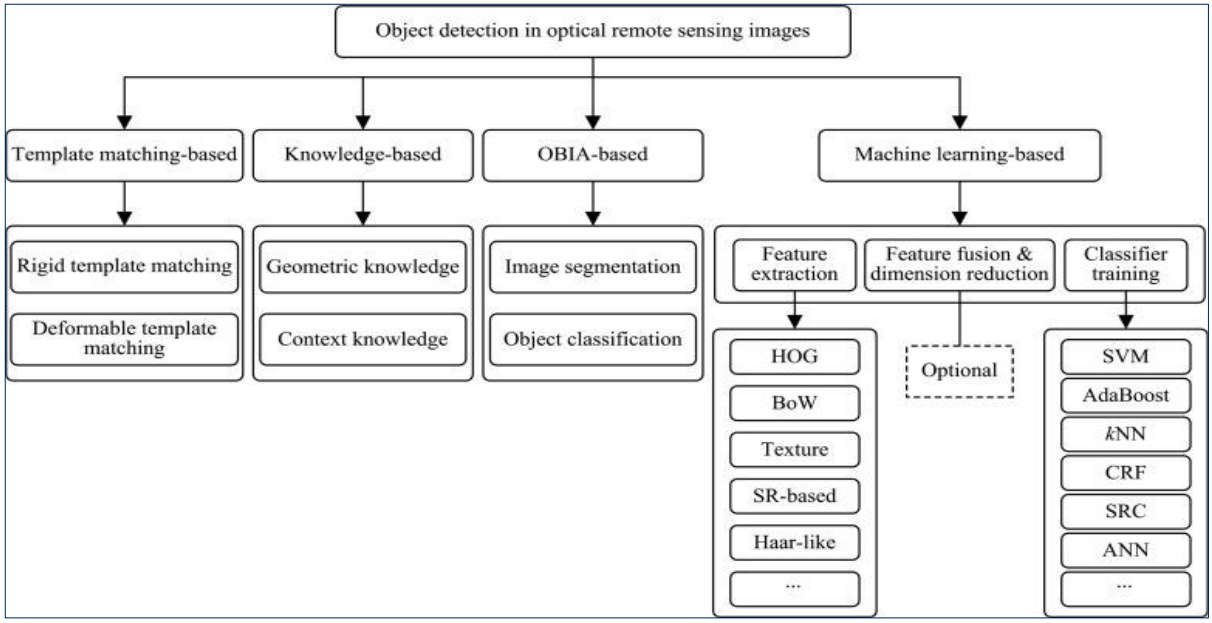


Figure 1: Taxonomy of methods for object detection (Cheng & Han, 2016, p. 12).

As extensively described by Verschoof-van der Vaart (2022) the first three more "traditional" methodologies have shown various restrictions within their implementation in the context of archaeological object detection. The challenges include the overfitting of these algorithms to specific object categories and data sources, difficulties in defining templates for heterogeneous objects affected by various processes, and the expertise required for implementing these algorithms (p. 8). In contrast, Machine Learning based object detection offers distinct advantages over other approaches because it does not rely on explicit predetermined sets of rules or human components that establish said rules. Instead, these methods can autonomously learn patterns and relationships from

data, making them adaptable to contexts and capable of handling complex and heterogenous datasets. By leveraging large datasets and models, Machine Learning-based approaches excel in capturing subtle patterns that may be challenging to specify manually. This flexibility and ability to learn from data make it a valuable tool in tasks like object detection, where rule-based systems may struggle to account for variability and evolving conditions.

1.1.3 Deep Learning and automated detection

Deep Learning (DL) can be considered a subfield of Machine Learning (ML) that is predominantly defined due to the usage of neural network architecture (Argyrou & Agapiou, 2022, p. 13). The concept was first introduced in the 1940s '(….) where the initial goal was to replicate the human brain system to address generic learning issues in a systematic manner' (Jamil *et al.*, 2022, pp. 8 – 9). Although it gained a lot of traction in the subsequent years, the methodology '(…) fell out of favour in [Machine Learning] research in the early 2000s, due to overfitting in training, lack of large training datasets, limited hardware processing capacity, and insignificant performance improvement compared to other machine learning techniques' (Ibidem). Therefore, due to the improvement of computational power and the increased availability of data, the usage of DL techniques in archaeological automated object detection have since become more prevalent.

In general, these neural networks consist of multiple interconnected layers (hence the term "Deep") that progressively learn to represent data at increasingly abstract levels. 'Although the fields of Machine Learning and Deep learning are very intermingled and poorly defined, the main differences lies in the fact that in the case of Machine Learning a problem, for instance object detection, has to be divided into different parts' (Verschoof-van der Vaart, 2022, p. 9). For archaeological automated detection, this means that traditional ML methods often require manually designed feature extraction and multiple stages of processing; DL, on the other hand, combines these many parts into a singular algorithmic expression (Ibidem). Ultimately, this means that DL has the capability to extract features and patterns directly from data, making it particularly suited for complex tasks in areas such as image recognition, natural language processing, and much more.

This thesis will be using the YOLOv8s algorithm (Redmon & Farhadi, 2023). The acronym stands for "You Only Look Once", which implies the algorithm's ability '(...) to accomplish the detection task with a single pass of the network, as opposed to previous approaches that used sliding windows followed by a classifier that needed to run hundreds of times per image' (Terven & Cordova-Esparza, 2024, p. 6). Furthermore, '(...) the YOLO framework is considered a "one-stage" detector, which means that the object localisation and classification is united in one process. This is contrary to "two-stage" detectors, e.g., Faster R-CNN, where this is split in separate processes' (Verschoof-van der Vaart & Olivier, 2021, p. 279). The algorithm has been designed to not only match the accuracy of its predecessors, but also to excel in speed and user-friendliness. The latter exemplified by the availability of a relatively easy-to-use python package created by Ultralytics[1]. All in all, the algorithm's ability to "Only Look Once" enables researchers to analyse larger quantities of data with less computational power and time, all while maintaining relatively high accuracy and precision. While it may not entirely replace more prevalent alternatives, the algorithm offers advantages in efficiency and accessibility. Further information can be found in chapter 2.2.

1.2 Archaeological spatial analysis

1.2.1 Archaeological site mapping

One example of an archaeological Big Dataset is created through archaeological site mapping; a fundamental aspect of archaeological practice. 'Ever since archaeologists have studied the past, the investigation of the location and distribution of archaeological remains in their surroundings, i.e., spatial analysis, has been a central endeavour in archaeology' (Verschoof-van der Vaart, 2022, p. 3). This mapping process involves the integration of various technologies, such as GIS, photogrammetry, and LiDAR, to capture detailed spatial data. With these advanced tools, archaeologists can generate highly accurate maps that not only offer a visual representation of the terrain, but also facilitate the precise recording and analysis of archaeological features. These spatial elements not only aid in understanding the genesis and decline of archaeological sites, but also serve

[1] The package can be found here: https://github.com/ultralytics/ultralytics

as records of features that will be lost during a destructive excavation. By documenting these features during the excavation, archaeologists can ensure that vital information about the site's history and cultural significance is preserved. The ability to analyse spatial data allows researchers to uncover patterns and relationships within archaeological landscapes, shedding light on past human activities and interactions with the environment. Furthermore, archaeological site mapping plays a crucial role in heritage conservation and management. By creating maps of sites and their surroundings, archaeologists can identify areas of significance and develop strategies for their protection and preservation.

Central to this thesis are the excavation maps that delineate archaeological features, achieved by the recording of coordinates within a GIS. These coordinates, typically acquired through measurements using GPS or Total Stations (TS), are subsequently used to construct detailed spatial representations of the archaeological site, and function as a permanent digital record of the identified features present within the archaeological site. Furthermore, within the Netherlands the method of measuring are somewhat standardised due to the national BRL SIKB 4000 protocols (SIKB, n.d.). This has resulted in fairly uniform datasets that can be used through open access data repositories curated by the government (DANS KNAW, n.d.). As can be seen in the figure 2, each feature is individually measured within a GPS coordinate system and later extrapolated to 2D vector polygons. This methodology is especially beneficial regarding large-scale excavations, as the data collection is relatively easy-to-implement and it requires limited data processing steps.

Although these maps create an overview of the archaeological site and its corresponding features, there are various difficulties in interpreting them accurately due to the inherent complexity of archaeological landscapes. Archaeological features are often clustered together, representing various activities and occupations spanning different time periods. This intertwining of features can make it challenging for archaeologists to distinguish and accurately attribute them to specific historical contexts. Furthermore, despite archaeologists' best efforts, features are not always recognised with ease due to preservation issues, such as degradation over time, disturbances, and human error in the recording and interpretation process. In other words, ‘(...) the great majority of the items

recorded in archaeological mapping are not visible in their own right but appear as one kind of reflection or another of buried deposits' (Campana, 2009, p. 22). Thus, although map making seems to be a relatively objective practice, it is important to keep in mind that '[r]ather than simply recording, mapping "translates" and "mediates" (...) [.] More specifically, '(...) maps are inherently subject to the subjectivities of their creators' (Flexner, 2009, p. 8). Consequently, the interpretation of archaeological site maps requires careful consideration to unravel the layers of features within them. Therefore, this debate will be further elaborated within chapter 4.

1.2.1 Archaeological structures

During an archaeological excavation, it is often the case that (pre)historic structures are uncovered. In the Netherlands, these structures are predominantly encountered "below ground", as not much standing architecture is present from deeper pasts. It is generally understood that standing structures are somehow more informative than below ground structures; however, as argued by Reynolds (2009), this is a common misconception. They state that '[w]hile the benefit of standing buildings is clear enough – an appreciation of the intended form of a building, its fixtures and fittings and so on – invariably standing structures, even those of relatively recent construction, have often undergone such transformations as to render any appreciation of the earliest activities that occurred in them almost beyond the reach of the archaeologist' (p. 345). Thus, destroyed underground structures can provide better primary archaeological deposits, which can be more suitable for the primary reconstructions of past human behaviour.

Unfortunately, as is the case with practically all archaeological features, the available evidence of these structures is incomplete and fragmentary (figure 3). Especially, the structures' '(...) form is only fragmentarily accessible, because of the impermanence of building materials (particularly timber, clay, and dry-stone walls) and destructive effects (erosion, later building activities, etc.)' (Trebsche, 2009, p. 506). Consequently, the remains of these structures can range from the actual physical remains, such as walls, floors, wooden posts, and hearths, to more subtle traces in the soil, such as postholes, pits, soil discoloration, and soil texture distinctions (figure 4). These features are often associated with each other, forming interconnected elements within the archaeological landscape. For example, clusters of postholes might indicate the former presence of

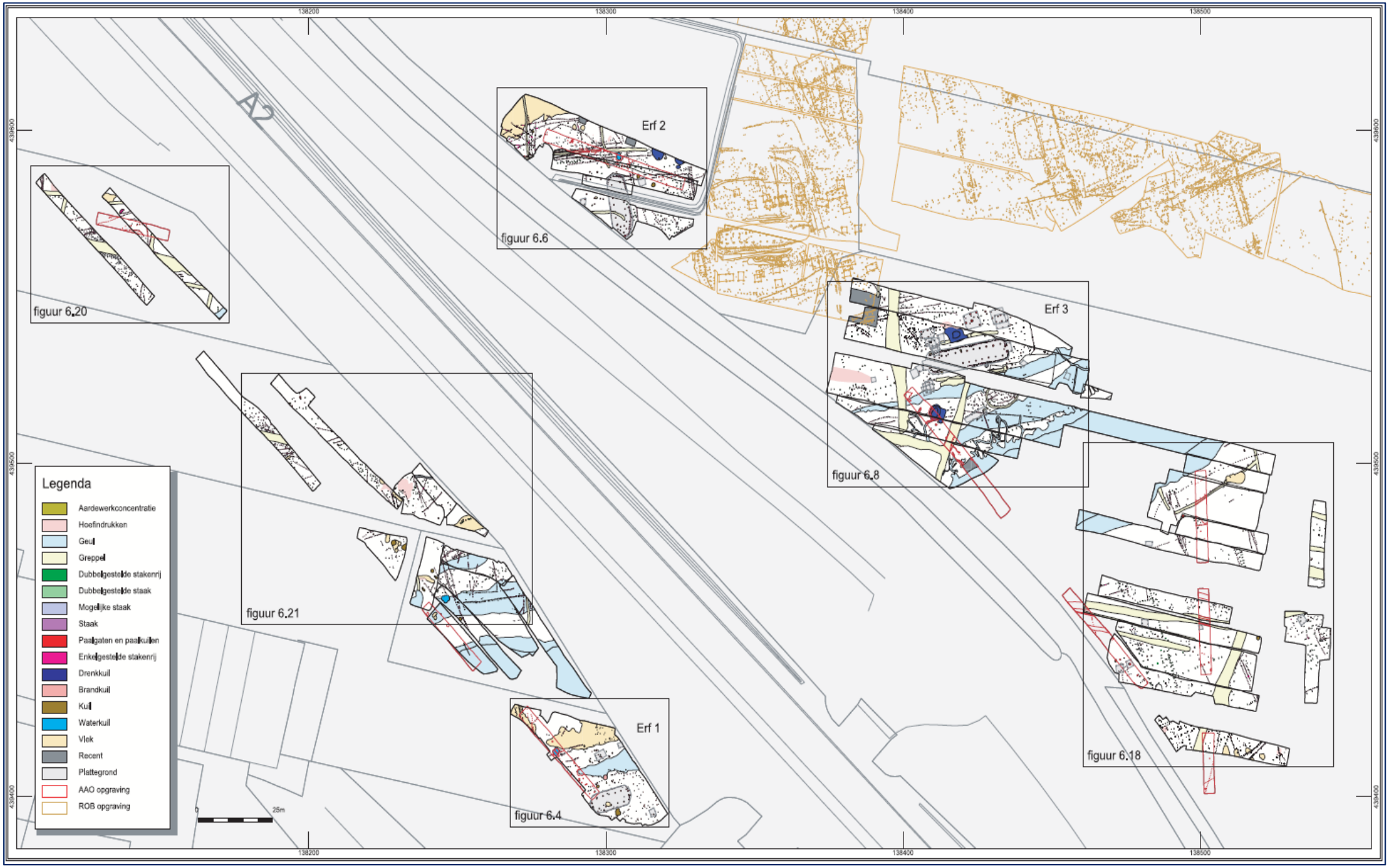


Figure 2: An example of a GIS output of features measured with a GPS system in the large excavation at Zijderveld, the Netherlands (Jongste & Knippenberg, 2005, p. 30).

wooden structures. By analysing the distribution and relationships between these features, archaeologists can reconstruct the layout and organisation of (pre)historic settlements. Over time, as more features are uncovered and examined, a better picture of the overall site emerges, revealing insights into the architectural design, social dynamics, and daily practices of past communities.

However, as can be seen in the figures 3 and 4, there can be a high degree of diversity present with how the structures and their corresponding features manifest themselves in the archaeological record. Therefore, the 2D illustrations and matching interpretations of structures are oftentimes different between archaeological paradigms, periods, and people. This is stated by Trebsche (2009): '(...) the interpretation of an excavation plan also depends on factors additional to the state of preservation: on the density and superposition of features, on the archaeologist's experience, on preconceived ideas of building form, and on already established building types' (p. 508). Although the Netherlands, which is the location for the datasets used within this thesis, has a thorough compendium on its excavated structures (e.g. Arnoldussen & Fokkens, 2009; Lange *et al.*, 2013), there is still much diversity and unknown data that undoubtedly has resulted and will result in a high degree of interpretive variability. This also connects to the issue of typologies, and with that the often associated culture historical paradigm, which tends to generalise, reinforce subjectivity, and perpetuate interpretive biases (Deeben & Theunissen, 2013, p. 7). Still, despite the inherent complexities and interpretive challenges associated with structures, their identification and investigation is important for understanding past human behaviour. Thus, while achieving an overall consensus in interpretations and typologies may be elusive, the process of studying granaries still yields valuable insights into past human behaviour, societal organisation, and architectural practices.

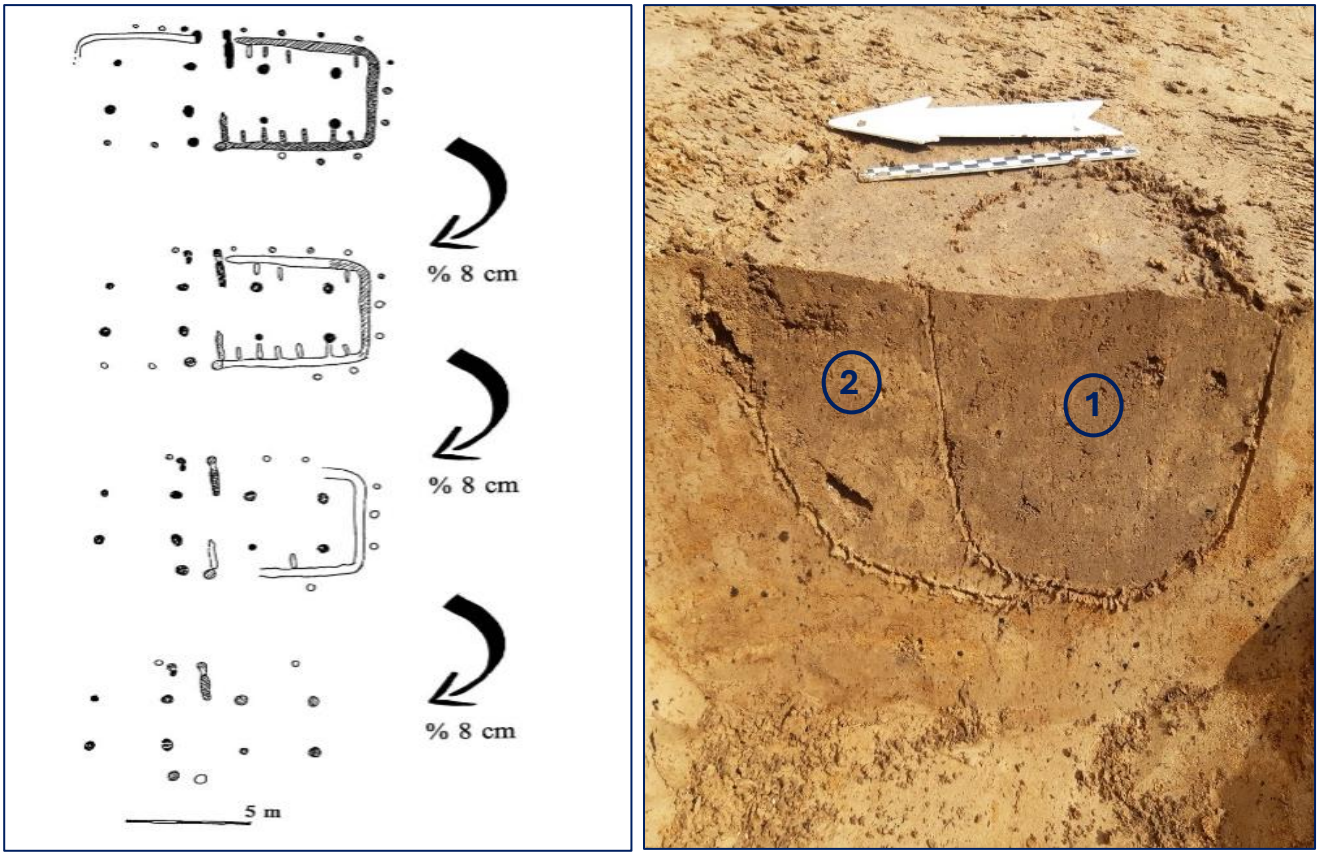


Figure 3: **Left**: state of preservation and legibility of archaeological features; the effects of erosion and ploughing on the Iron Age byre-cum-dwelling-house at Grøntoft (Denmark) (Trebsche, 2009, p. 509). **Right**: cross-section of a medieval post hole from Veldhoven, where (1) is the post-pipe and (2) the post pit (Archol bv, 2023).

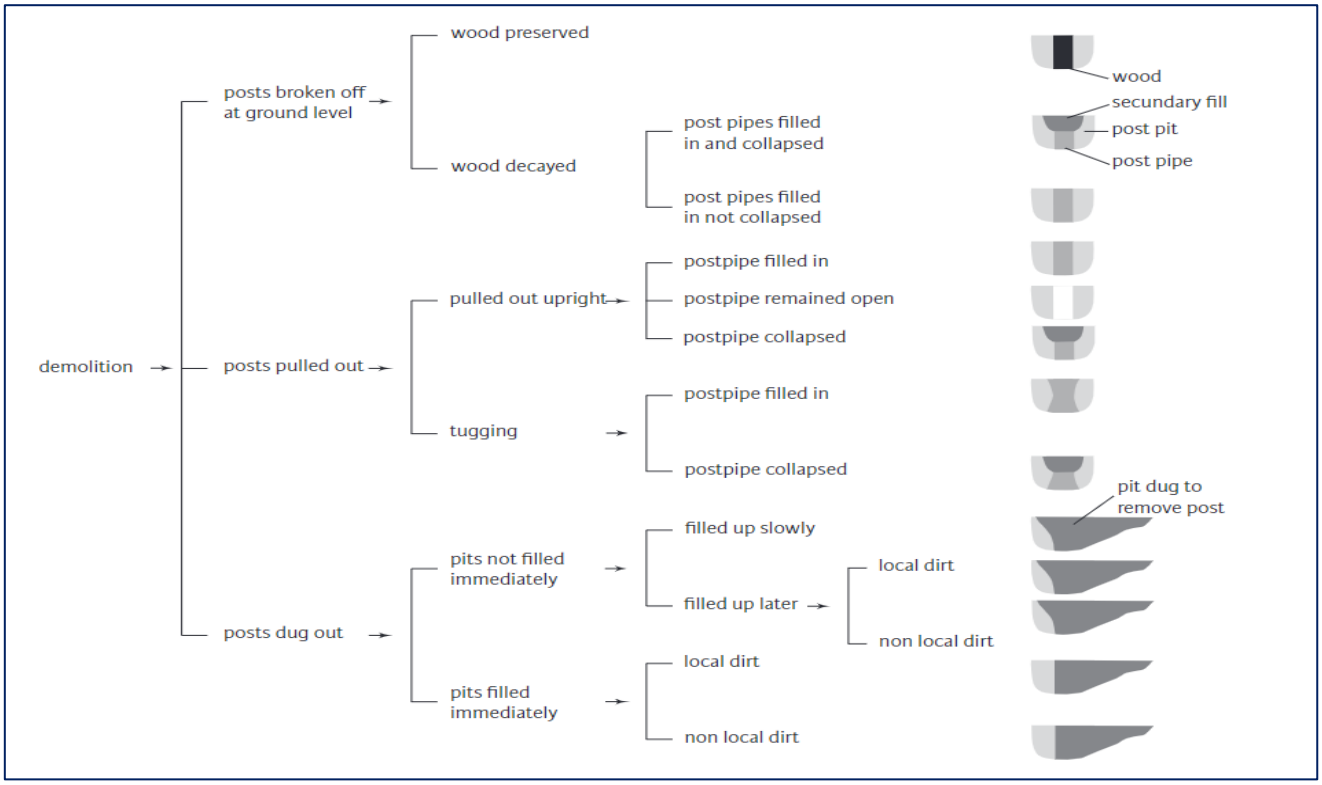


Figure 4: A diagram showing the different variables that can determine the appearance of a post hole during and after the demolition of a building (Theuws, 2013, p. 319).

1.2.2 Prehistoric granaries

In the context of this research, the exploration of structures diverges from the more traditional archaeological inquiries, as it predominantly centres on a methodological development. Here, the primary focus lies in devising and applying a methodology aimed at identifying and analysing archaeological structures within 2D excavation maps. Instead of delving into the details of individual structures from historical or functional perspectives, the emphasis is placed on evaluating the adaptability of a DL approach in detecting and delineating these features. Consequently, this thesis will engage with the existing literature on granaries in the Netherlands on a basic level, with less emphasis on interrogating archaeological contexts and inquiries. Still, in order to make substantiated methodological decisions it is important to understand the context of the data. Therefore, some general information will be provided about the relevant archaeological and historical background.

This research will be looking at granaries dated to the Bronze and Iron Age in the Netherlands (figure 5). This precise period has been selected due to the relatively

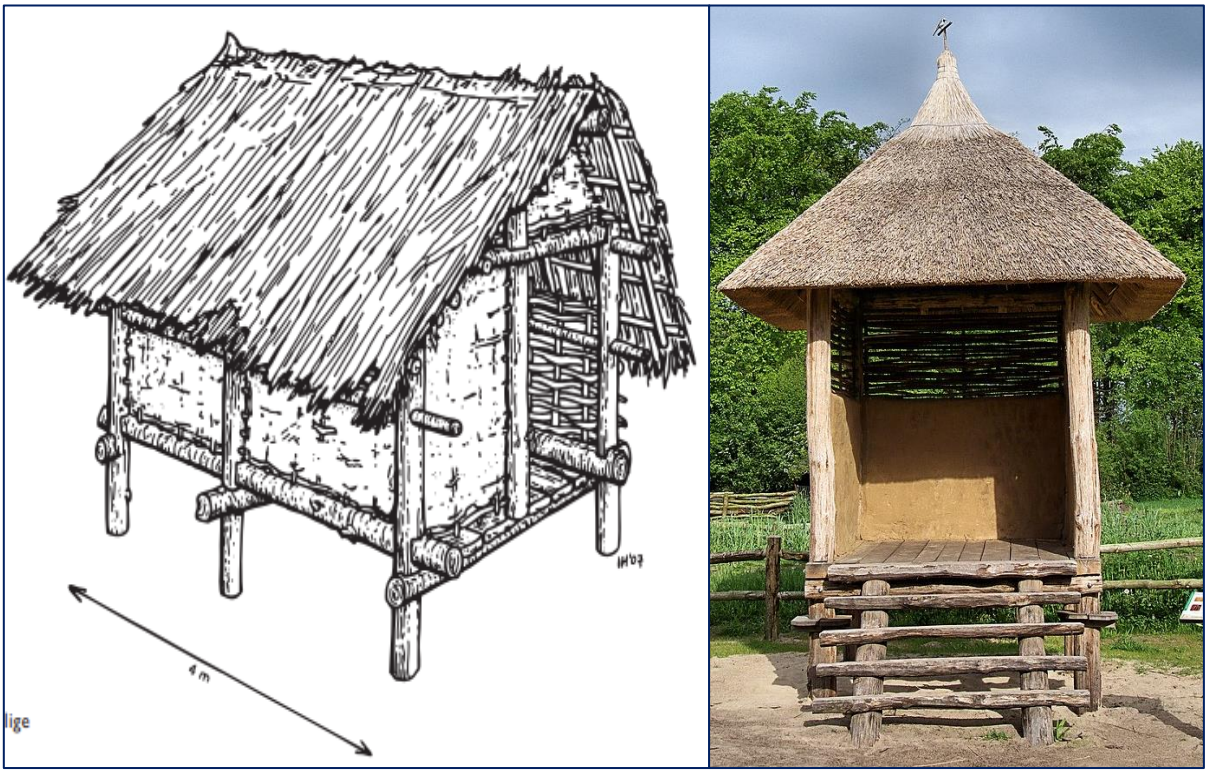


Figure 5: **Left:** Reconstruction drawing of a six-post granary from the early Iron Age (Hermsen & Haveman, 2009, p. 45). **Right**: Reconstruction of a four-post Iron Age granary near the Wekeromse Zand, Gelderland (Wikimedia commons, 2012)

standardised nature of the structures in question (Arnoldussen, 2008; Maes, 2009; Hermsen & Haveman, 2009) and extensive data availability. Furthermore, the focus on the Netherlands is attributed to its uniform data collection and formatting practices, as well as the abundance of excavation reports and general research conducted throughout the region.

Granaries are '(...) relatively small square or rectangular wooden buildings that were used for food storage during the [Bronze and Iron] age in northwestern Europe. They have a raised platform, which makes it difficult for pests and moisture to reach the harvest' (Maes, 2009, pp. 79-80). While there undoubtedly were alternative methods for storing these food products, identifying them archaeologically presents challenges. In contrast, granaries leave distinct evidence of their existence through easily recognisable square or rectangular shaped post-hole imprints. These structures typically have either four, five, six, eight, or nine posts, which are commonly spaced 2 meters apart (figure 6). Granaries are frequently found in close proximity to prehistoric house plans, often in significant numbers. Therefore, these structures are generally interpreted as being part of the larger homestead. However, as the floor plans are too small for domestic purposes, the function has to be interpreted as something different that is suitable for early farming communities. As stated by Maes (2009) '(...) the thickness of the posts in relation to the surface area gives the impression that these had to bear a heavy burden' (p. 81). Correspondingly, ethnographical, classical, and archaeobotanical sources generally categorise these structures as granaries (Malrain *et al.*, 2002; Hermsen & Haveman, 2009; Maes, 2009; van der Meer, 2014).

Furthermore, according to an extensive study of approximately 1500 granaries by Maes (2009) four-post granaries are generally most common in the Bronze and Iron Age; as they

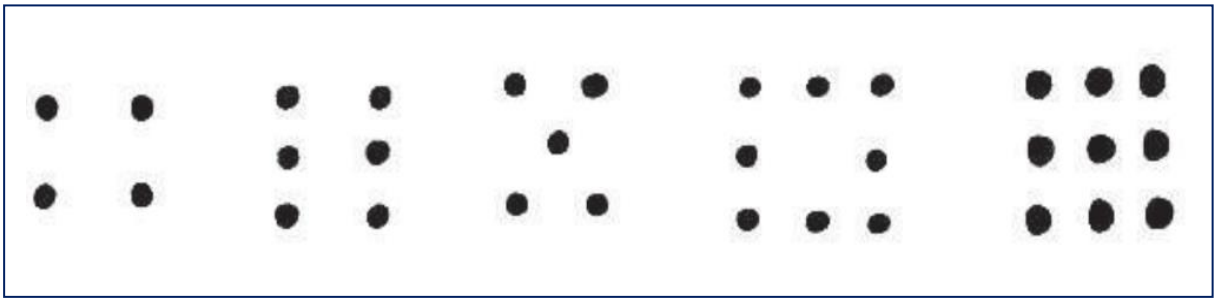


Figure 6: Sketch of the most common types of granaries. From left to right: four-post, six-post, five-post, eight-post, and nine-post granary (Maes, 2009, p. 82)

comprised more than half of the total number of the identifiable granaries (59.88%). Six-post make up the second largest share, namely 31%. The nine-post comprise 6.86% of the total. The five- and eight-pole represent only a small percentage, 0.62% and 1.64% respectively (p. 82). However, it is important to note that the amount of posts does not necessarily indicate a difference in function, instead it could relate to an increase in size and/or capacity. "However, although this increase in capacity is often used as an indicator of agricultural and economic affluence, Hermsen and Haveman (2009) rightly argue that conclusions should be drawn with caution, as 'it is impossible to make one-to-one connections between the number and size of granaries found during archaeological observations and the size of grain harvests or the economic importance of arable farming in the past' (p. 46, [translated by author]). Alternatively, this variation in the number of posts could be attributed to a range of social, cultural, or other factors pertinent to human behaviour. Therefore, no clear reason has been ascertained as to why granaries show different post-hole patterns throughout different places and prehistoric periods. This debate will be further elaborated in chapter 2.1.

1.3 Research topic

Given the information discussed in the introduction above, this thesis aims to advance the current body of research by applying Deep Learning techniques to automate feature detection within GIS maps. While Deep Learning has been widely used for remote sensing data (e.g. Bundzel *et al.*, 2020; Orengo *et al.*, 2020; *Somrak et al.*, 2020; Sorouch *et al.*, 2020) and, to a lesser extent, for analysing topographic or historic maps (e.g. Garcia-Molsosa *et al.*, 2021; Pereira *et al.*, 2024), its application to GIS maps is unusual. Although GIS data is frequently utilised in academic research, the methodological focus has predominantly been on predictive modelling rather than on automated feature detection or similar tasks addressed in this thesis research. Consequently, no other relevant publications directly addressing this context were identified during the course of this study. Therefore, this thesis seeks to fill this gap by exploring the application of Deep Learning techniques specifically within the domain of archaeological GIS data.

More specifically, the focus of this study is on leveraging the YOLOv8s algorithm to automate the detection of archaeological prehistoric granaries on excavation maps. While manual analysis remains common in archaeological research, it is oftentimes

relatively time-consuming and labour-intensive, particularly when dealing with large spatial datasets. Moreover, the convoluted nature of the archaeological record and the diverse range of features it contains present significant challenges for traditional methods. These challenges underscore the demand for digital tools capable of streamlining the process of identifying archaeological features within GIS maps.

Furthermore, this thesis will investigate the broader implications of automated feature detection for archaeological practice. By automating a detection process, archaeologists can potentially save significant time and resources, allowing for more efficient data analysis and interpretation. Additionally, automated feature detection may lead to improved interpretive outcomes by facilitating the identification of subtle patterns and relationships within archaeological datasets. However, it is vital to critically assess the impact of this automation; considering factors such as overall data quality, algorithmic biases, and other relevant research constraints.

All in all, this thesis aims to contribute to the ongoing evolution of archaeological practice by helping to develop and understand digital methodologies that could potentially enhance the efficiency and accuracy of data analysis and interpretation. Moreover, by addressing the challenges and opportunities associated with automated feature detection on GIS archaeological excavation maps, this study seeks to foster a more nuanced understanding of the role of technology in everyday archaeological research and interpretation.

1.3.1 Research questions

As not much research has been done in this particular research context, the overall aim of this thesis is to address the general applicability of this methodology within this context. In other words, the thesis will predominantly address the methodological implementation of such a Deep Learning model, and the further implications of this implementation on the manner and quality of archaeological interpretation and research. With this in mind, the research question and two sub-questions will be formulated as follows:

"How can the YOLOv8 algorithm be effectively employed to automate the detection and analysis of Bronze and Iron age granaries within archaeological excavation maps, and to

what extent can this approach potentially enhance the efficiency and accuracy of archaeological site documentation and analysis?"

- How can the Deep Learning YOLOv8 algorithm be adapted and optimised to effectively recognise diverse architectural features representative of Bronze and Iron age granaries on archaeological GIS excavation maps?
- What are the limitations, biases, and challenges associated with implementing the Deep Learning YOLOv8 algorithm for automated feature detection within archaeological contexts, and how can these challenges be addressed to ensure the reliability and accuracy of the automated identification process?

Ideally, these research questions will guide the investigation and analysis conducted throughout this thesis, providing a framework for evaluating the effectiveness and implications of Deep Learning algorithms in the context of archaeological feature detection and analysis. Where the first sub-question mostly relates to the applicability of the model, and the second sub-question to the encountered challenges and limitations. The conclusion of this thesis will then function as a preliminary recommendation for further applications of this methodology in future research. All in all, through these questions, this research aims to give a small contribution to the application of advanced digital methodologies in an oftentimes analogue discipline.

1.3.2 Outline of the next chapters

A total of six subsequent chapters will be used to delve into various facets of this research, providing an exploration of the application Deep Learning in archaeology. Each chapter addresses specific components such as the characteristics of prehistoric granaries, the basic principles of the YOLO algorithm, and the foundational theoretical framework of this thesis:

- Background information: This chapter establishes a context to everything relevant within this thesis. It will explain the basic principles of the YOLOv8 architecture and delve deeper into the literature regarding Bronze and Iron age granaries in Dutch archaeology.
- Theoretical framework: This chapter talks about the theoretical foundations of digital archaeology, and the potential implications and limitations of Deep

Learning in archaeological research. Additionally, the relevant information associated with site mapping will be examined, focusing on data accuracy, critical mapping theory, archaeological visibility, and biases. Ultimately, this chapter will underscore the model's theoretical possibilities and limitations.

- Methodology: The methodology chapter outlines the approach taken in this research, covering data collection and preparation, the YOLOv8 algorithm development, training procedure, and testing techniques.
- Results: This chapter represents the overall performance of the trained models. It will give evaluation metrics, graph representations, and compare the different results across all detected classes.
- Discussion: The discussion chapter further interprets the results presented in the previous chapter, addressing the accuracy and reliability of the algorithm, possible factors influencing detection performance, and implications for archaeological interpretation. Furthermore, it looks at several methodological considerations, strengths and limitations of the approach.
- Conclusion: In the final chapter, the study concludes by summarising the key findings and insights generated throughout the research. It will answer the research questions posited here, give advice for future research, and will reflect on the significance of the findings in the context of digital archaeology.

2. Background information

2.1 Granaries in archaeology

2.1.1 Formation of prehistoric granaries

Granaries can be found all over the world. They provide a useful insight into the agricultural practices, social organisation, and economic strategies of (pre)historic societies. These storage facilities represent some of the earliest visible forms of food and resource management in the prehistoric period. The oldest interpreted granaries can be found the in the Jordan Valley and have been dated to 11,300 – 11,175 cal. BP (Kuijt & Finlayson, 2009). The main function of these buildings has however not changed, as they were '(...) [d]esigned with suspended floors for air circulation and protection from rodents (...)' and are often '(...) located between residential structures that contain plant-processing installations' (Idem, p. 10966). Here, the structures were circular, and predominantly made of mudbricks and stone. Furthermore '(...) the roof was probably flat and covered with a protective coating of mud to shed rain water' (Ibidem). The main indicator that resulted in the structures being interpreted as granaries was the excessive amount of barley and oat remains present, and evidence for food cultivation in its near vicinity. Some debate still remains whether these structures can be interpreted as house plans instead (Weiss *et al.*, 2006). However, a few millennia later, sufficient evidence of the existence of granaries can be found in many corners of the world, such as western Anatolia (Maltas, *et al.*, 2021), China (Liu *et al.*, 2017), northern Africa (Morales *et al.*, 2014) and the Mediterranean (Peña-Chocarro *et al.*, 2015). This evidence indicates a significant evolution in storage techniques and the central role of granaries in these communities.

The first granaries in Western Europe, including the Netherlands, date back to the Middle Bronze Age. The phenomenon appears somewhat simultaneously in various regions across northwestern Europe; however their exact origin is uncertain. Previously, only underground pits seemed to be of importance, but it is likely that a large part of the harvest was also stored in attics indoors (Maes, 2009, pp. 84-85). Unfortunately, the introduction, characteristics, and use of prehistoric granaries in the Netherlands remain relatively understudied. While several archaeological reports do touch on this subject,

most are descriptive rather than analytical. This means that there is no concise dataset detailing the number of granaries discovered, their specific periods, or their locations within the Netherlands. Furthermore, the lack of detailed analytical studies has resulted in a limited insight into the technological advancements, architectural variations, and cultural implications of granary use and development within the region. Nonetheless, some research has been conducted in Great Britain, where granaries from corresponding periods exhibit relatively similar characteristics. Given the scarcity of resources specifically focused on the Netherlands, this research will draw on the British literature to gain some understanding of granaries in prehistoric times. Still, it is important to emphasise that this thesis acknowledges the limitations of this approach, as the direct application of British findings to the Dutch context may not fully account for regional differences. Therefore, while British research provides a valuable framework, there is a pressing need for region-specific studies to accurately interpret the significance of granaries in the prehistoric Netherlands.

An elaborate paper by Gent (1983) gives an overview of the density and time frame to which these structures within northwestern Europe (figure 7 and 8), however it is important to note that this paper is quite outdated, and more data should be added to give a better outline. Despite its age, Gent's work indicates that granaries became most prevalent from the Middle Bronze Age onward and continued to thrive in these regions through all phases of the Iron Age. This observation aligns well with the archaeological datasets used in this study, which show that most excavated granaries are commonly dated to the Iron Age. Furthermore, all the granaries in the dataset can generally be associated with a farmstead in their vicinity, indicating that these structures were used in agricultural contexts. Unfortunately, the time-sequence of activity and use of these granaries and their corresponding farmsteads is often difficult to ascertain due to challenges in accurately dating these structures. Additionally, no clear patterns in orientation, distance, or spatial relationships between granaries and farmsteads can be established due to the lack of comprehensive research in this area (Maes, 2009). This means that any clear connections or inferences about the specific roles or chronologies of the granaries

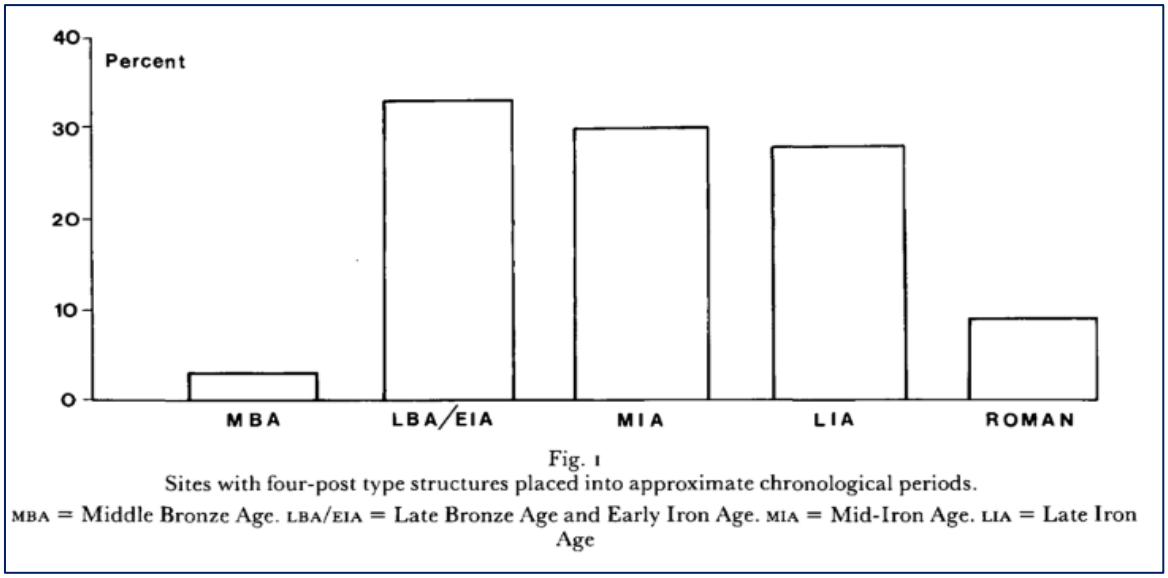


Fig. 1
Sites with four-post type structures placed into approximate chronological periods.
MBA = Middle Bronze Age. LBA/EIA = Late Bronze Age and Early Iron Age. MIA = Mid-Iron Age. LIA = Late Iron Age

Figure 7: Approximate periodisations of four-post granaries in northwestern Europe in 1983 (Gent, 1983, p. 245).

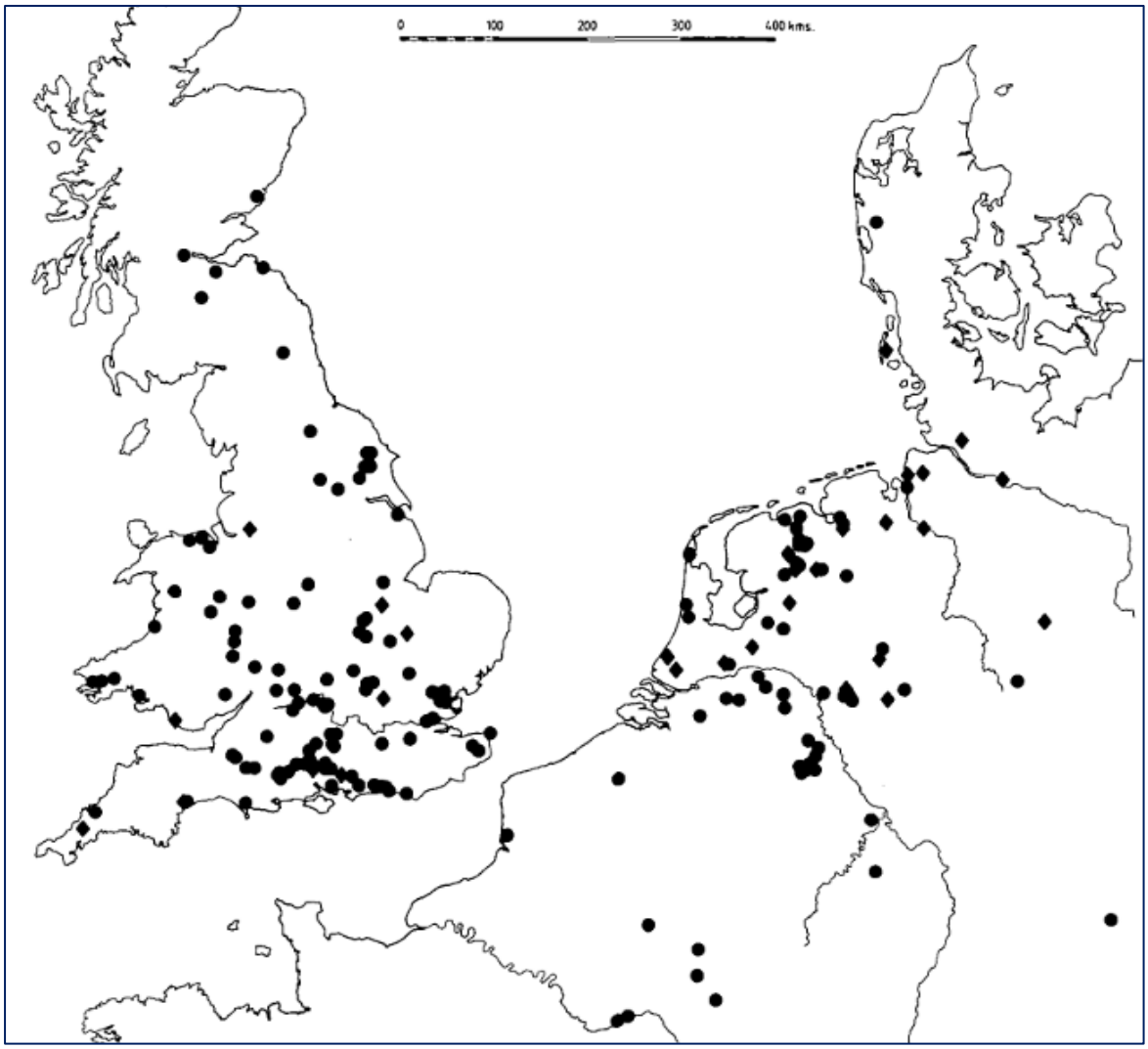


Figure 8: Distribution of four-post granaries and similar structures in northwest Europe. Pre-Roman: circles. Roman Iron Age: diamonds (Gent, 1983, p. 246).

and their farmsteads are difficult to draw. However, it is evident that granaries were utilised extensively within prehistoric agricultural contexts, serving as storage and safeguard for grain supplies.

This does not mean that no other storage options were considered in prehistoric times. Underground pits, (organic) storage vessels, silos, basketry, cashes, raised platforms, and many other examples were still being used contemporaneously to granaries (Jiménez-Jáimez & Suárez-Padilla, 2019) (figure 9).

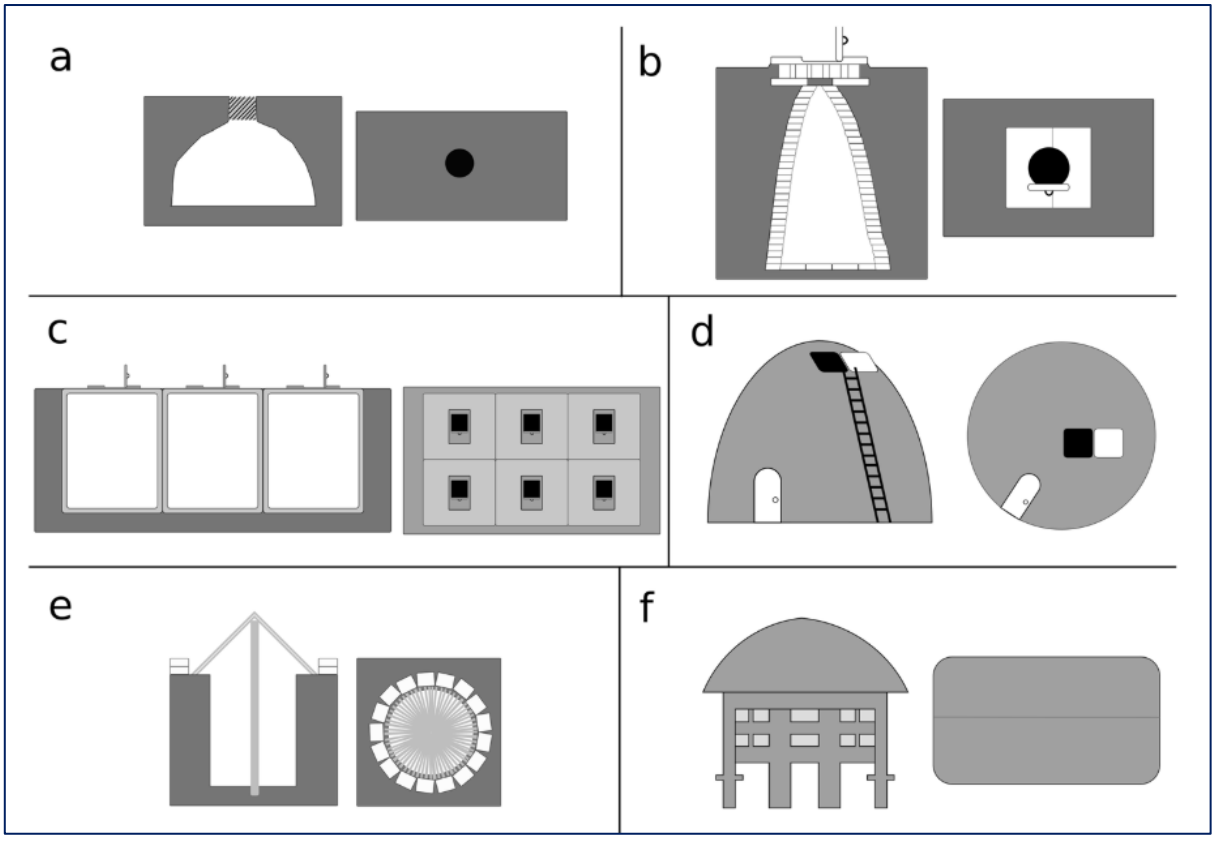


Figure 9: Schematic ideal representation of the types of storage container discussed in the text, both section (left) and view from above (right). (**A**) Simple sealed pit. (**B**) Elaborate sealed pit. (**C**) Underground silo complex. (**D**) Aboveground silo. (**E**) Unsealed pit. (**F**) Granary (Jiménez-Jáimez & Suárez-Padilla, 2019, figure 4).

These examples are also present within Dutch archaeology, and are often seen in close context with Bronze and Iron age farmsteads. Especially the underground silos are prevalent in Dutch contexts (Bakels, 2009; Maes, 2009). However, the archaeological visibility of these particular storage methods is often more limited, particularly when it comes to organic remains, which are susceptible to degradation compared to more durable, deep-entrenched structures. With underground pits it is also often difficult to ascertain whether these were in fact used as storage containers, or had perhaps another

elusive function (see Jiménez-Jáimez & Suárez-Padilla (2019) for a contemporary overview of this debate). Still, this preservation bias might have led to a skewed understanding of the prevalence of different storage techniques. It is plausible that granaries were not as common as other storage methods, but due to the more frequent discovery and preservation of granaries, they are often perceived as more prominent in the Dutch archaeological record than they may have actually been. Especially regarding the fact that granaries can be deemed as quite vulnerable for outside influences, whereas underground storage is more secure against theft and many other intrusive activities (Jiménez-Jáimez & Suárez-Padilla, 2019, p. 807). Furthermore, the construction of granaries, as opposed to other methods, represents a relatively large expenditure of energy, and requires more maintenance. In other words, '(...) long-term preservation in granaries, requires frequent shovelling for improved ventilation, periodic inspection, and continuous monitoring to prevent infestation or weather-related damage' (Idem, p. 808). On the other hand, granaries are durable, easy to monitor, and relatively simple to use (Ollich *et al.*, 2012, p. 215-216). Granaries have been used for millennia, and due to their visibility, they offer interesting opportunities for understanding ancient agricultural practices and social structures.

2.1.1 Architecture and definition of prehistoric granaries

Granaries in the Netherlands are identified by the number of visible posts during archaeological excavations. As previously mentioned, granaries typically consist of 4, 5, 6, 8, or 9 posts spaced approximately 2 meters apart (figure 6). Typically, granaries require consistent spacing between postholes, uniform depth for each posthole, and a final reconstruction that is relatively straight and aligned (Maes, 2009, p. 82). With that these small structures also often exhibit a square or rectangular shape. Consequently, structures with 10 or more posts are often categorised as small sheds. However, attributing a different function or definition to a building based solely on the presence of an additional post-hole can be somewhat arbitrary. Such distinctions are influenced by modern classifications rather than inherent functional differences in the structures themselves. Therefore, it is crucial to recognise that the distinction between a large granary and a shed is quite tenuous (Brijker *et al.*, 2012, p. 21). This is also illustrated in the available literature that regards the granaries of the Netherlands, were six-posters are

sometimes referred to as granaries, sheds, or in a more general term, as outhouses. Where in some instances, the structures can sufficiently be identified or categorised as granaries due to detailed archaeobotanical research, many other excavated structures are not studied in such detail. Therefore, function and significance is often inferred on the basis of a somewhat arbitrary contemporary categorisation that may not fully capture the diverse uses and histories of these prehistoric structures. To be pedantic, what is the difference between a granary and a shed? Functionally speaking, both structures serve as repositories for goods or tools, albeit tailored to different purposes. A shed typically houses implements, equipment, or materials used for tasks ranging from farming to household work. Conversely, a granary specialises in storing harvested grains, protecting them from pests and weather. Despite these distinct roles, both are designed with durability and protection in mind, often featuring elevated floors, ventilation systems, and sturdy construction. While raised floors may be more prominent in granaries—serving to protect stored grain from rodents and moisture—such features are not necessarily exclusive to granaries. Sheds could also feature raised floors depending on the local environment or construction practices, similarly, protecting valuable tools from moisture, dirt, or any other environmental threats. Thus, while their contents may vary, their fundamental purpose of safekeeping and preservation renders them similar in functional intent.

Therefore, in archaeological research, the distinction between the two can become more of a semantic exercise than a practical one. The term "granary" implies a specific function related to grain storage, whereas "shed" is more general and can encompass a wider variety of uses. Yet, in practical terms, the architectural features and principles guiding their construction often overlap significantly. The function of grain storage is usually not proven definitively, even though the structures might resemble traditional granaries. This means that while the functional purpose typically associated with a shed is often not excluded, archaeologists still routinely categorise it as a granary. This overlap can create confusion or unnecessary segmentation in scholarly discussions when the focus should be on understanding how such structures were used and adapted over time. Firmly adhering to terminology can therefore be limiting and arbitrary within archaeological interpretation. Still, within this thesis, this terminology will be adopted as a convention

for clarity and consistency in describing and categorising archaeological finds and features. All in all, the above discussion aims to highlight that the term “granary” and “shed” are often used loosely and interchangeably, without careful consideration of their specific archaeological and historical context.

Setting the semantic discussion aside, the small structures possess characteristics that can be used as evidence of their role in storage. This is clearly outlined in the paper by Maes (2009), who identifies several relevant characteristics. First, the small square and rectangular structures are frequently identified in the obvious vicinity of Bronze and Iron age houses, indicating they formed part of the farmstead. Additionally, it can be assumed that they were not houses themselves, as their ground plans are too small for that particular purpose. Consequently, the substantial thickness of the posts in relation to their surface area suggests they were designed to bear a heavy load (p. 82).

Determining the usable volume of a granary is unfortunately challenging, if not impossible, due to the absence of clear indications of the structure's precise height, resulting in only rough estimates. The lack of reliable evidence regarding roof construction—whether it featured a sloping roof or internal divisions—compounds this difficulty (Idem, p. 85). Additionally, specific information about construction materials and appearance is lacking. Reconstructions of prehistoric granaries often draw on contemporary ethnographic examples and involve a significant degree of conjecture. These granaries were likely shaped by the need to support heavy loads, the availability of local materials, and the requirement for weather resistance. Thus, it was crucial for the supporting posts and platform to be sturdy enough to support considerable weight. Walls might have been constructed from clay or wooden planks, with roofs made from thatch, straw, or wood. Although concrete evidence for a platform is lacking due to the preservation of only ground plans, ethnographic examples of elevated granaries suggest that such platforms were common (Ibidem). This hypothesis, while difficult to accept for granaries with four or six posts, appears more plausible for those with five or nine posts. In these cases, the close placement of the posts indicates limited interior space and suggests they were designed to balance a heavy load (Villes, 1985, p. 429). Furthermore, ground accumulation between the posts could also indicate the presence of an elevated platform (Gent, 1983, p. 247). All in all, due to the lack of clear evidence, it is difficult to

ascertain the appearance of prehistoric granaries. Most reconstructions rely on conjecture and comparisons with contemporary ethnographic examples.

2.1.2 Function of prehistoric granaries

Granaries are a method of storage commonly associated with the prevalence of agricultural practices in its vicinity. 'It is considered one of the mechanisms for coping with the risks and uncertainties that characterised prehistoric subsistence, e.g. seasonal and annual climatic fluctuations, natural hazards, pests, and all sources of variability that will affect food availability' (Halstead & O'Shea, 1989). Storage is, in addition, a way of managing surplus for later use, and therefore a key element in social and economic complexity (Peña-Chocarro *et al*., 2015, p. 379). However, it is important to note that the existence of food storage in earlier hunter-gatherer societies is still a subject of debate and has not been conclusively dismissed (e.g. Rowley-Conwy & Zvelebil, 1989; Stopp, 2002; Cunningham, 2011). Nevertheless, seasonal peaks in food production, aimed at lasting for extended periods, led to the creation of large-scale storage facilities designed for its durability and longevity. In general, Jiménez-Jáimez and Suárez-Padilla (2019) have identified three main concerns regarding inter-annual grain storage in prehistoric societies. Based on ethnographic, archaeological, and experimental evidence, they assert that storage facilities are primarily used to address three critical objectives:

a) Maintaining grain dormancy and preventing premature germination: Grain dormancy is essential to ensure that seeds do not sprout before they are sown in the next planting season. This is achieved by controlling environmental factors such as temperature, light, and, most importantly, moisture. High temperatures and exposure to light can trigger germination, while excessive moisture can disrupt dormancy.
b) Slowing down microbial decay: Microbial decay is a significant threat to stored grain, leading to spoilage and loss of food resources. To combat this, storage methods need to minimise the conditions that promote microbial growth, such as warmth and humidity. This involves creating airtight or well-ventilated storage spaces, using materials that absorb excess moisture. Slowing down decay ensures that the grain remains edible and nutritious over extended periods.

c) Protecting grain from pests and animals: Pests such as insects, rodents, and birds pose a consistent threat to stored grain. Therefore, effective storage solutions incorporate physical barriers to prevent access by these pests.

Granaries are designed to do just that. They provide a controlled environment that regulates temperature, humidity, and access, effectively shielding stored grain from pests such as insects, rodents, and birds. Prehistoric Europe was inhabited by pests like *Sitophilus zeamais* (maize weevil) and *Sitophilus granarius* (wheat weevil), as evidenced by archaeological research (Antolín & Shäfer, 2024). This protection is essential for maintaining food security and preventing losses due to contamination or predation. Additionally, the architectural design of granaries included features like raised floors or elevated platforms to further deter pests and promote ventilation, contributing to the longevity of stored grains.

Stored within these granaries were various types of grains such as emmer wheat and naked barley, which was the staple diet of the communities relying on these structures (Out, 2009; Bakels, 2009; Maes, 2009; Kirleis *et al.*, 2012). Other crops such as '(...) einkorn wheat, free-threshing wheat, pea, flax/linseed, and poppy were in the end also cultivated in the region (...)' (Bakels, 2014, p. 95). However, these crops have yet to be convincingly linked to granary structures. In some cases, this may be due to the rarity of the crops, while in others, evidence is simply lacking. The direct evidence for the storage of food products other than grain is also deficient. However, it cannot be ruled out that fruits, vegetables, and even beverages in barrels were also stored within the granaries (Maes, 2009, p. 82). Lastly, even though the granary is often associated with the storage of subsistence products, it is possible that there were instances of storing tools and other non-food related items.

In addition to their critical roles in storage and protection, granaries may have served broader socio-economic functions within prehistoric societies. Beyond their immediate practicality, some scholars argue that these structures potentially contributed or were the result of the development of social stratification and economic complexity (e.g. Childe, 1954; Bogaard *et al.*, 2009; Bogaard, 2017; Hastorf & Fowhall, 2017). According to this debate, the accumulation of surplus food facilitated by granaries (and other storage facilities) could have enabled specialisation of labour, trade relationships, and

the emergence of elites who controlled access to these stored resources. Still, this "surplus theory", although influential, faces challenges in terms of definitive causal proof. For instance, although the concept of surplus can be seen as a '(...) good proxy for studying wealth, and thus of wealth inequalities. The correlation is not perfect. In fact, wealth may exist without any storage and '[c]onversely, the possibility of storage without payments and without any kind of wealth should not be totally ruled out' (Darmangeat, 2020, p. 67). Furthermore, attributing surplus as the primary catalyst for social inequality oversimplifies the complexities of early human societies. Factors such as political organisation, environmental variability, and cultural practices likely played roles alongside storage facilities in shaping social structures. Moreover, the archaeological evidence linking granaries directly to social stratification is often indirect and subject to diverse interpretations. Lastly, determining whether granaries were intended for communal use or reserved for individual households or elite groups is challenging, adding another layer of complexity to understanding their societal role. In essence, while granaries were pivotal in facilitating food storage and management, their broader socio-economic impact is an intriguing area warranting further research.

For instance, as outlined by Hermsen and Haveman (2009), it is often assumed that the presence of large or numerous granaries on a farmstead indicates grain overproduction and surplus stocks. The argument follows that those who owned these stocks likely played a role in distributing grain harvests throughout the surrounding area (p. 46). However, there are several arguments that can be presented to counter that idea. Firstly, the surface area of granaries alone does not reliably indicate their storage capacity, as volume is also influenced by the height of the structures. Moreover, accurately assessing the total number of granaries associated with a farmstead during archaeological surveys can be difficult. Preservation conditions and the extent of excavation areas affect how easily granary layouts are identified. When multiple farmsteads were situated together, an additional challenge arises in determining which and how many granaries belonged to each individual farm. Another challenging aspect of interpreting granaries is determining their duration of use and, consequently, their frequency of replacement (Idem, pp. 47-48). In other words, there is too much ambiguity, data bias, and variability in interpreting

granaries as markers for ancient societies' socio-economic dynamics. Still, they remain an interesting proxy for understanding economic affluence.

Another factor that is often associated within the functional framework of prehistoric granaries is that of a ritual, cultural, or religious significance. This idea, mainly based upon ethnographic and anthropological studies, suggests that granaries may have served ceremonial purposes or held symbolic importance within ancient societies. Although this is archaeologically difficult to prove, there have been contemporary studies that showcase this cultural significance for agricultural communities. Hermsen and Haveman (2009) give a relevant example to this phenomenon, by referring to the village of Songo on the Dogon Plateau in Mali. Here families live in compounds consisting of two or three houses, depending on the number of wives the household head has. Adjacent to these houses are storage sheds that closely resemble granaries reconstructed for prehistoric contexts in the Netherlands. The architecture of these sheds reflects the gender of the owner and their status. In this instance, the men own the largest sheds, primarily used for millet storage, equipped with three doors. On the other hand, women manage smaller sheds with only one door, divided into multiple smaller compartments for storing personal items like clothing, small food products, and jewellery. The doors of the sheds are sometimes adorned with intricate wood carvings, signifying the status of their owners. As married sons remain on their father's compound, new houses and grain sheds are added over the years, gradually filling the compound (p.48) (figure 10).

This anthropological example illustrates that the quantity and size of storage sheds near houses do not always directly correlate with harvest size, the significance of arable farming in the local economy, or the physical lifespan of the building structures. Cultural traditions associated with changes in family composition, marriages, and other factors could have equally influenced their construction and use. Unfortunately, archaeological methods do not provide this information directly, which means that much of the information regarding cultural significance of these structures is lost. Therefore, it is important to not assume a certain function for these structures based solely on their size or quantity, as their roles within ancient societies were possibly multifaceted and influenced by various cultural dynamics and practices.

Figure 10: Example of several Dogon granaries in close proximity to the village (Wikimedia commons, 2010).

However, there are instances in the Dutch archaeological context were certain cultural or ritual practices can be associated with granaries. These are most clearly evidenced in the form of large quantities of burnt pottery and the presence of substantial parts of one or multiple vessels in post-holes of houses and granaries. These depositions are commonly interpreted as a ritualised deposition of abandonment. Especially when the deposited ceramic vessel is intact, decorated, unaccompanied by typical refuse artifacts, and carefully placed, it is often described as "ritualised." While proving this conclusively is always challenging, in these cases, '(...) there is clearly a process of conscious selection. Therefore, it is plausible that there is a deeper significance to the deposition of these objects, hinting at a ritual of abandonment' (Hermsen & Haveman, 2009, p. 82). Although this concept is most commonly associated with the main buildings or farmsteads in these periods, there have been sufficient instances of supposed ritual depositions within granary context as well (Benallou, 2021). Brück (1999) even mentions that for British Bronze Age settlements, there is often a distinction between the material deposited in the main building, and secondary storage structures such as granaries. In the main buildings, discoveries are often associated with consumption, production, and status, whereas in storage structures, they are often linked to food preparation (p 150). The concept of ritual abandonment was introduced by van den Broeke (2002), who connected these deposits to the abandonment phases of post-holes, based on

stratigraphy and approximate artifact dating. Where direct causational evidence is lacking due to the fragmentary nature of the archaeological record, the interpretation relies heavily on contextual clues. In particular, the intactness and careful positioning of objects within post-holes suggest a symbolic act rather than mere disposal, hinting at the possibility that these depositions were part of a broader ritualistic practice marking the transition or closure of a structure's use. However, without more definitive evidence, these interpretations remain speculative, underlining the challenges in reconstructing the full cultural significance of such practices in prehistoric contexts.

In conclusion, granaries were essential components of prehistoric societies, serving as both practical storage solutions and, potentially, as indicators of socio-economic and cultural dynamics. While their primary function was to ensure the preservation of surplus food and manage the risks associated with agricultural production, their broader impact on social organisation remains a subject of ongoing debate. The possibility that granaries also played roles in social stratification, communal versus individual ownership, and even ritual practices adds layers of complexity to their interpretation. As the archaeological record often lacks direct evidence, caution is necessary when drawing definitive conclusions about the socio-economic implications of these structures. Future research that combines archaeological, ethnographic, and experimental approaches are therefore crucial in deepening our understanding of the multifaceted roles granaries played in prehistoric societies.

2.1.3 Prehistoric granaries in practical archaeology

While this overview provides a general understanding of granaries in prehistory, it is interesting to consider how these structures are actually identified in the practical context of archaeological excavations. Identifying a granary often involves a combination of factors, including the layout, number, and nature of postholes, as well as the historical and environmental context in which the structure is found. However, as mentioned before, granaries are notoriously understudied in archaeological research, which means that much of the identification process in the field relies on educated guesses. This lack of detailed study makes it particularly difficult to definitively distinguish granaries from other types of structures, further emphasising the need for more targeted research in this area.

In Dutch excavations, granaries are typically identified by examining specific structural features (outlined by Hermsen & Haveman, 2009; Maes, 2009). Their most noticeable distinctions are the layout and characteristics of the postholes. Granaries tend to have postholes spaced approximately 2 meters apart, with the structure often supported by four, five, six, eight, or nine posts. The relatively small number of posts and the tight spacing between them make these structures easier to differentiate from larger buildings, such as main houses or side buildings, which are usually much more expansive. The posts themselves are often quite deep and of even depth, suggesting that they were designed to support considerable weight. Consequently, this means that when fewer posts are excavated—either due to disturbances, overlapping features, or other practical limitations—the most significant evidence for identifying these structures is often compromised. In such cases, the interpretation of the building's function as a granary becomes more difficult and speculative, as the structural features that would typically indicate storage purposes may be lacking or unclear.

Contextually speaking, granaries are commonly found near prehistoric farmhouses, located within agricultural contexts. The proximity to residential areas points to a functional connection with the storage of harvested crops. In some cases, botanical research has helped confirm the agricultural role of these buildings. For example, traces of burnt grain found in postholes can provide evidence of the structure's function. However, this type of evidence is far from conclusive, and in many cases, botanical data is either sparse or indirect.

Given the limitations of current research, the identification process remains tentative and open to debate. The absence of clear, definitive evidence means that archaeologists must rely on broad assumptions and comparisons with other sites, making granaries a subject of ongoing uncertainty in the field. This highlights the importance of further research into the identification and study of these structures. More focused investigations into granaries, including for instance experimental archaeology and comparative studies are essential to improve the accuracy of these interpretations. Until more data is gathered, the task of distinguishing granaries from other types of buildings will remain a challenging and often speculative aspect of practical archaeology.

This limitation also highlights an inherent problem with the dataset used to train the Deep Learning model. The data itself is largely derived from educated guesses made by archaeologists, given the lack of definitive evidence for many of these structures. Furthermore, while archaeologists can incorporate a wide range of contextual information—such as environmental, historical, and social factors—into their analyses, DL models do not have the capacity to interpret these contexts. This inability to account for broader contextual data is a fundamental challenge for the model, limiting its ability to make accurate and reliable identifications of granaries. However, implementing such a model might also reveal and use patterns or relationships within the data that archaeologists have not yet considered. DL models can analyse vast amounts of data and will factor characteristics in their selection process that might be elusive through regular analysis. This capability offers the potential to generate new insights into the identification and understanding of granaries, even if the models themselves lack the ability to contextualise their findings. This theoretical concept will be further outlined in chapter 4.1 regarding the human and algorithmic "black-box" and the "garbage in – garbage out" concept.

2.2 YOLOv8 architecture

As briefly described in the introduction, this thesis will make use of the YOLOv8 Deep Learning object detector. YOLO, which stands for "You Only Look Once," is a family of one-stage object detection algorithms that have been originally designed for real-time processing. YOLOv8 represents one of the latest iterations in this series, building on the strengths of its predecessors while introducing several enhancements to improve performance and accuracy. Although this algorithm is relatively new, its methodology is already widely discussed in the academic literature. However, in the context of archaeology, the amount of literature still remains somewhat limited. A brief review of the available literature reveals that most papers focus on remote sensing imagery and the localisation of archaeological sites (e.g. Olivier & Verschoof-van der Vaart, 2021; Canedo *et al.*, 2023). Furthermore, several of these papers are evaluating the usefulness of the YOLO object detector compared to more traditional two-stage detectors like R-CNN (e.g. Marçal *et al*, 2024; Vokhmintcev *et al.*, 2024). This indicates that the algorithm is still

being explored and tested for its practical applications within the archaeological domain. Therefore, much more research is still needed in order to fully understand its potential and optimise its use in this field.

2.2.1 Basic principles of YOLO

In short, YOLOv8 continues the tradition of YOLO models by employing a single neural network to process an entire image in one go, predicting bounding boxes and class probabilities directly. This approach contrasts with two-stage detectors like Faster R-CNN, which first generate region proposals and subsequently classify them (Olivier & Verschoof-van der Vaart, 2021, p. 279). According to Jiang *et al*. (2021) there are several core principles of YOLO that can be summarised as follows:

1) Single-stage detection: Unlike two-stage detectors like R-CNN, YOLO performs detection in a single stage, directly predicting bounding boxes and class probabilities from the entire image in one pass, which greatly enhances speed and efficiency.
2) Real-time processing: The algorithm's architecture is designed for real-time applications, allowing it to process images and videos quickly. Its streamlined network structure, which avoids complex pipelines, ensures fast computation.
3) Regression-based detection: YOLO frames object detection as a single regression problem, predicting bounding boxes and class probabilities simultaneously, which simplifies the detection process.
4) Non-maximum suppression: The algorithm predicts multiple bounding boxes per grid cell but uses non-maximum suppression to select the bounding box with the highest Intersection Over Union (IOU) with the ground truth, improving the precision of the final detections.
5) Network simplification: Improvements in YOLO v2 and subsequent versions focus on enhancing accuracy and recall without significantly deepening or broadening the network, which helps maintain its speed advantage. Furthermore, the architecture is relatively small, which makes it feasible to implement on hardware with limited computational resources.

In short, where CNNs require the input image to iterate through several convolutional, pooling, and fully connected layers, the YOLO methodology processes the entire image in one forward pass through its network. Instead of generating region proposals and subsequently classifying them, YOLO treats object detection as a single regression problem. It directly predicts the bounding boxes and class probabilities, optimising for both localisation and classification simultaneously.

However, in a general sense it can be said that CNN based methods are better suited where high accuracy is paramount and computational resources are abundant. Therefore, although the loss of accuracy within this research is, of course, unfortunate, it can be argued that the benefits of using YOLO in archaeology could outweigh this drawback. YOLO's capabilities might allow for the rapid localisation of archaeological features on maps, possibly providing an aiding tool for archaeologists. This speed facilitates the analysis of large datasets, making it possible to quickly generate comprehensive maps that can then be refined through further investigation. Furthermore, a broad initial classification can be iteratively refined, by leveraging expert knowledge to enhance accuracy.

2.2.2 Evaluation metrics

In order to assess the workings of a DL model, standardised evaluation metrics have been developed. These metrics make it possible to compare the results of other developed models and methods and essentially see how well the model performs. These metrics will also be used within chapter 5, and are therefore outlined below.

For this thesis the model will be evaluated on the metrics of recall (eq. 2), precision (eq. 3), and mAP (eq. 4). Each of these evaluation techniques will measure a different component of the overall model's performance. Importantly, these metrics are commonly based upon the amount of True Negatives (TN), False Negatives (FN), True Positives (TP), and False Positives (FP) present within the algorithmic output. These predictions are based upon either the difference or similarity between the model prediction and the actual supervised input. However, as this thesis is not dealing with a standard binary classification (true/false), but with a bounding box input and bounding box prediction these categories are calculated by the means of Intersection over Union

(IoU; eq. 1, 1a & 1b). In short, '[w]hether a prediction falls in one of these categories is determined by the amount of overlap between the generated bounding box and the ground truth bounding box' (Olivier & Verschoof-van der Vaart, 2021, p. 282) (figure 11).

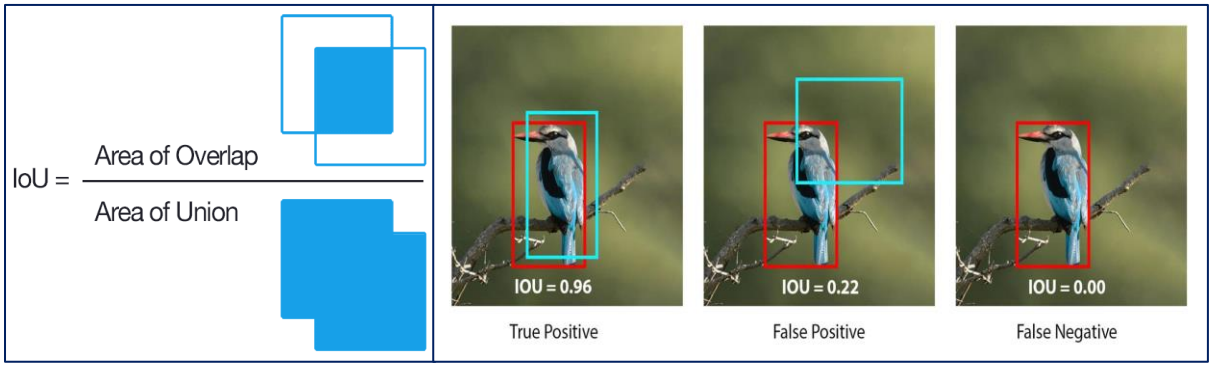


Figure 11: Schematic overview of how the Intersection over Union (IoU) metric is measured (Wikimedia commons, 2019).

This means that the IoU can be used as a threshold of what is considered a correct detection. This threshold is able to control the trade-off between precision and recall in the DL object detection model. For instance, within this thesis it might be beneficial to choose a lower IoU threshold if the primary goal is to ensure that potential structures are not missed, accepting that some false positives will occur. Conversely, a higher IoU threshold might be used if the focus is on precise localisation, ensuring that identified features are accurately bounded. However, in a standardised format, '[t]he threshold for a detection being a TP is normally set to an overlap of 0.5. If the overlap is less, the detection is considered a FP. The (average) IoU can not only be used as a measure for loss during training, but also gives an indication of the quality of the bounding boxes' (Olivier & Verschoof-van der Vaart, 2021, p. 282).

$$Intersection\ over\ Union = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad (1)$$

$$Area\ of\ Overlap = Area(b_{pred} \cap b_{true}) \quad (a)$$

$$Area\ of\ Union = Area(b_{pred}) + Area(b_{true}) - Area\ of\ Overlap \quad (b)$$

Furthermore, the metrics of recall and precision are also commonly used to evaluate the performance of a DL model. Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify positive instances from the entire pool of actual positives. It answers the question: out of all the actual positive instances, how many did the model correctly identify? Thus, a high recall indicates that the model

is good at finding all positive examples. In the context of this thesis, high recall means that the model will be able to identify most of the actual granaries present in the dataset, and with that, minimising the risk of missing these archaeological structures.

On the other hand, precision focuses on the accuracy of positive predictions. A high precision indicates that when the model predicts something as positive, it is likely to be correct. For example, in the context of this thesis, a high precision means that when the model identifies an area as containing a granary, it is usually correct and not identifying unrelated features as granaries.

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \quad (2)$$

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \quad (3)$$

The last metric within this thesis is that of the Mean Average Precision (mAP) (eq. 5). This metric evaluates the overall accuracy of an object detection model by considering both the precision and recall across all the detection classes. The Average Precision (AP) is the key component of mAP. This measures how well a model detects and localises objects across various classes in a dataset. It is calculated by generating a precision-recall curve for each class. The area underneath this curve (AUC) represents AP, where higher values indicate better model performance. Mean Average Precision (mAP) then averages the AP values across all classes in the dataset.

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \quad (4)$$

The metric of mAP is commonly used in research papers to benchmark and compare the performance of models. Different variants of mAP, such as mAP 0.5, 0.75, and 0.95, can be specified to evaluate performance at different thresholds of IoU. Therefore, this metric can commonly be understood as an overall performance indicator of the DL model. Furthermore, this metric is commonly preferred over the F1-score in multiclass object detection, because mAP assesses precision and recall across all classes, accounting for varying levels of confidence in object detections. F1-score, on the other hand, is more suited for binary classification tasks and may not provide an assessment of the model's performance across multiple classes.

3. Dataset

The dataset comprised of excavation datasets collected and created by professional archaeologists during excavations and trial trenching projects. The datasets were selected based upon an approximate similar temporal context of the Iron and Bronze age in the Netherlands. The datasets of granaries were sourced from excavation sites across the Netherlands. The dataset consists of 19 different excavations, with a total of 477 corresponding labelled granaries; the precise locations of which can be seen in figure 12. As can be seen, most site locations originate in Noord-Brabant, Utrecht, and Gelderland, with few exceptions.

All the data collected is owned by Archol bv, and due to embargo constraints and confidentiality requirements, the images in this thesis have been anonymised to protect sensitive information. This means that specific details about the excavation sites, as well as examples of the raw data, cannot be fully disclosed within this thesis. This decision reflects Archol bv's policy to maintain control over unpublished or proprietary data while ensuring the confidentiality of the archaeological information. As a result, while the data serves as the foundation for the research conducted in this thesis, its anonymisation ensures adherence to their policy.

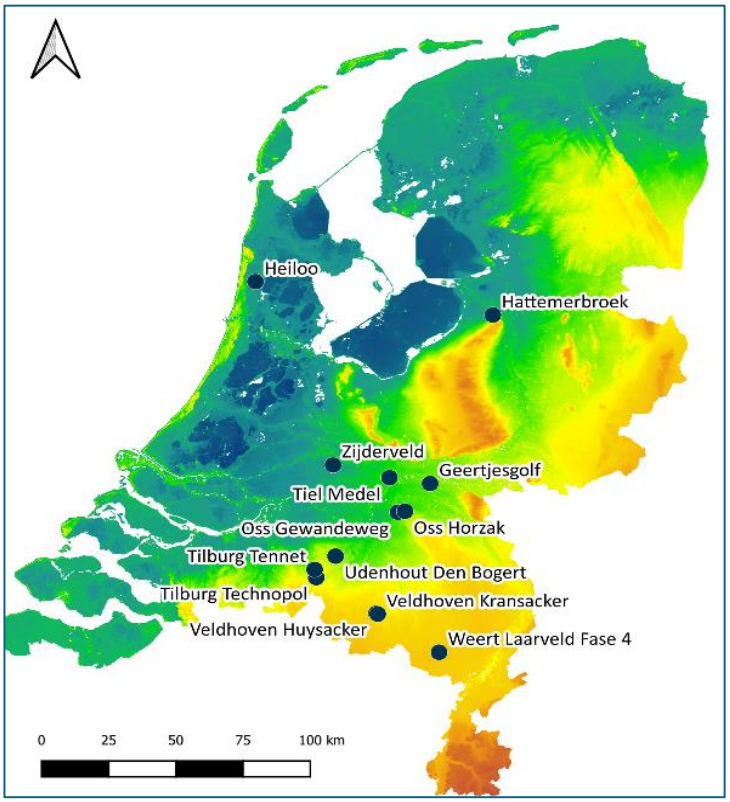


Figure 12: Location of all the data that has been used within this thesis.

This thesis is allowed to show the specific excavations or trial trenching projects that the data was sourced from, as can be seen in table 1. This table showcases the amount of granaries and their corresponding type that were identified in each project. This already displays the discrepancy present between the 4-post granaries as opposed to other categories, as these are most prevalent in the archaeological record and are possibly recognised easier to their standardised format.

	4-post	6-post	9-post
Heiloo Zuiderloo	11	0	1
Tiel	16	2	3
Veldhoven Huysackers	64	22	4
Udenhout den Bogerd f3	29	5	2
Tennet Tilburg-Zuid	26	5	2
Udenhout den Bogerd f4	5	1	0
Veldhoven Huysackers f2	10	4	3
Veldhoven Huysackers f3	27	7	2
Hattemerbroek	6	5	2
Udenhout den Bogerd f5	9	0	4
Oss Gewandeweg	5	1	4
Tilburg Technopol 7	30	0	0
Zijderveld_A2	29	1	2
Zijderveld DO-I	13	2	0
Tiel Medel	80	0	0
Oss Horzak	5	1	2
Weert Laarveld fase 4	7	1	0
Veldhoven Kransacker	3	2	1
Geertjesgolf vpl 1 + 3	7	0	4
Total	382	59	36
	80,1%	**12,4%**	**7,5%**

Table 1: Toponyms of the 19 excavation datasets used within this thesis, along with the number of granaries and their corresponding number of post-holes (Archol bv, n.d.).

The data utilised in this thesis was collected using GPS technology. The datasets were provided in shapefile format, a widely used geospatial data format that facilitates integration with GIS software. Each shapefile was linked to accompanying Microsoft Access databases containing metadata, including details on the shape, NAP, colour, texture, depth, and associated artifacts of individual archaeological features (figure 13). Other available metadata linked to the Microsoft Access system are photographs of the features, digital drawings, and possible botanical samples that were collected from it.

spoor	type	contour	diepte	datering	structuur	opmerking
977	KL	SCH		NT		
978	PK	VVG		PREH		
979	GR	VVG	0	NTC		onderkant greppel, niet zichtbaar in coupe. Met mogelijk spitsporen
980	PK	VVG	10	NTC		waarsch nieuwe tijd, palenrij
981	PK	VVG	13	NTC		met hout
982	SS	VVG	6	NT	32	onderkant greppel (spitsspoor)
983	SS	VVG	5	NT	32	onderkant greppel (spitsspoor)
984	SS	VVG	2	NT	32	onderkant greppel (spitsspoor)
985	SS	VVG	3	NT	32	onderkant greppel (spitsspoor)
986	SS	VVG	6	NT	32	onderkant greppel (spitsspoor)
987	PK	VVG	12	PREH		botspikkels in bovengrond,vaag spoor
988	CR	VVG	36	PREH		
989	CR	VVG	11	PREH		
990	PK	VVG	12	NTC		met hout en piepschuim
991	PK	VVG	10	NTC		met hout
992	CR	VVG	23	PREH		met verbrandingsresten (v3) en crematiedepositielaag (v4). V3 en v4 zijn aanwezig in vlak 2.
993	PK	VVG	11	NTC		met hout
994	NV	VVG				
995	CR	VVG	21	PREH		op bodemgrafkuil kleine concentratie crematieresesten; in twee putten gevonden
996	CR	SCH	14	PREH		
997	CR	SCH	18	PREH		type b
998	CR	SCH	21	PREH		graftype B
999	REC	SCH		RECENT		
1000	CR	VG	17	PREH		Met glazen kraal (v911); enkele crematiespikkels
1001	PK	VVG	16	PREH	8	

Figure 13: Example of one of the Microsoft Access databases, displaying feature number, interpretation, contour, depth, approximate dating, structure number, and comment (Archol bv, n.d.).

As can be observed, certain data points, such as depth, are occasionally missing, which is largely a result of the nature of archaeological investigation. For example, during trial trenching, not all features are examined in detail, as the purpose of such investigations is to conduct a broad survey rather than a full-scale excavation. As a result, some features are intentionally left in situ, meaning their full extent, including depth measurements, cannot be recorded. Additionally, archaeological features that are clearly recent are similarly excluded from detailed documentation, as they are often not considered relevant to the focus of the investigation. This is not problematic for the purposes of this thesis, as the primary metadata utilised includes the feature number,

interpretation, and structure number. These data points are sufficient to accurately identify granaries and their corresponding post-holes within the shapefile.

By linking these Access databases to the shapefiles in QGIS software using the feature number as the primary key, feature maps can be plotted in the Amersfoort/RD New coordinate system (EPSG:28992). This practice is common within commercial archaeology, where such maps are often referred to as AllFeatureMaps or, in Dutch, AlleSporenKaart (ASK). These ASK maps serve as a foundation for preprocessing the data for the deep learning (DL) model, a process that will be detailed in chapter 5 (figure 14).

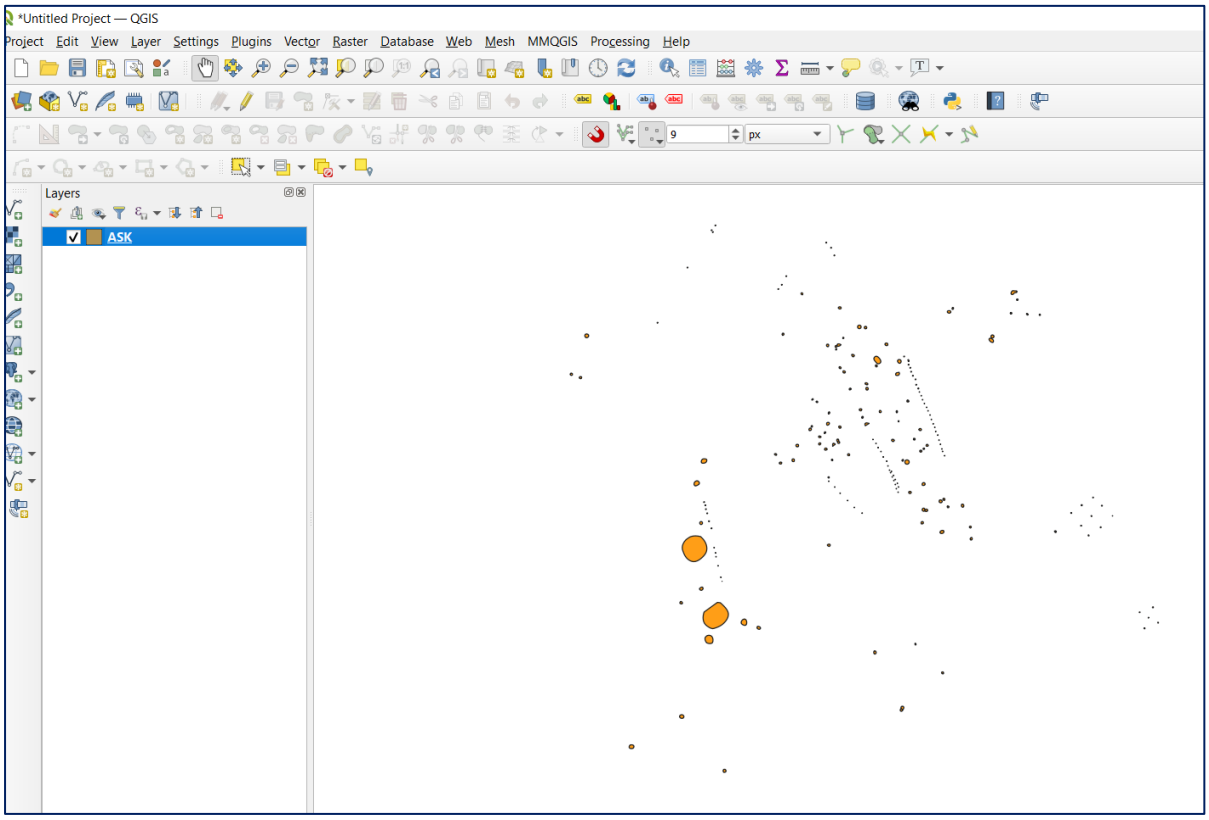


Figure 14: Example of an ASK from an excavation in the QGIS environment (Archol bv, n.d.).

The used granaries were identified by professional archaeologists at either three distinct stages: during the field excavation, during the subsequent data processing phase, and during the reporting phase. In the reporting stage, archaeologists synthesise their findings, drawing on relevant literature to contextualise and confirm the identification of granaries. Consequently, these structures are explicitly labelled as granary features within the metadata, simplifying their classification and making them readily accessible for this study. This can be seen in figure 14 where all the associated features are given a structure number.

Most of the granary examples used in this thesis have been published within their corresponding archaeological reports. This means that there is a somewhat reasonable assumption that the granaries are correctly identified based upon the characteristics described in chapter 1 and 2. As will discussed in chapter 4, the concept of the human "black box" is particularly relevant here, as there remains no universally established or clearly defined set of criteria for accurately identifying granaries within archaeological research. In some cases, the data or report reflected some interpretative doubt by the archaeologist due to missing post holes, atypical feature shape or depth, or other inconsistencies with typical granary representations. In these instances the examples were carefully considered, and oftentimes removed due to the ambiguity and potential for misinterpretation.

Although every effort has been made to use data with the highest accuracy, there is an inherent possibility of misinterpretations, leading to both false positives and false negatives. In particular, the dataset might contain mislabelled granaries or overlook unlabelled ones due to the complexity of the archaeological record. Yet, despite these limitations, the YOLOv8 model is designed to handle such variations. Its ability to generalise from diverse examples enables it to detect patterns even when some data points are inconsistent. This adaptability ensures that, even with a small margin of error, the DL model should still deliver reliable results in practical applications.

4. Theoretical framework

4.1 Digital archaeology and theory

Digital archaeology is often criticised for lacking a strong theoretical foundation.. Digital archaeology has been '(...) accused of being technocratic, apolitical and indifferent to social and cultural concerns and of relating poorly with theoretical orientations currently found in archaeology' (Dallas, 2015, pp. 177-178). Several reasons are oftentimes given to substantiate this accusation. First of all, the rapid pace at which the technological advancements are introduced seem to outstrip the development of the necessary theoretical frameworks. Secondly, there is often a focus on the immediate practical application of digital tools at the expense of theoretical development. The immediate benefits of using AI, such as increased efficiency and new insights into archaeological data, can overshadow the need for a thorough theoretical examination. Practitioners may prioritise the tangible results of these technologies over the slower, more abstract process of theory-building. This is stated by Zubrow (2006), who notes that '[t]here is a tendency to use digital technological solutions simply because one has the "toys" available' (p. 22). The appeal of quick wins and the pressure to produce measurable outcomes can divert attention away from the more time-consuming task of developing a theoretical foundation. This is also distinctively regarded as the "law of the hammer" (Moore & Keene, 1983) '(...) in that the appeal of the technology has caused excessive application, or pounding, without regard to purpose, appropriateness, or theory' (Drennan, 2001, p. 668). All in all, the discipline '(...) has been subject to what is called an "anxiety discourse," wherein the identity, nature and academic legitimacy of archaeological computing was questioned and concerns expressed about its theoretical core, the rigour and relevance of its methodologies, the value of its outputs, and the extent to which its contributions were recognised as having any significance to the broader field (Huggett *et al.*, 2018, p. 43)

However, in recent years, there have been many publications which add relevant theoretical underpinnings to the discipline (e.g. Perry *et al*. 2016; Morgan, 2019; Morgan, 2022; Huggett, 2024). Digital archaeology '(...) has become an interdisciplinary perspective in which integration, collaboration, and the introduction and use of

methodological and theoretical digital tools are reshaping the broader discipline' (Huggett, 2024, p. 328). These developments reflect a growing recognition of the importance of engaging with theoretical frameworks, offering new ways of establishing digital methods as a component of the broader archaeological project. Therefore, in line with this sentiment, this thesis will examine both the opportunities and the limitations of applying these tools in archaeological inquiry, while reflecting on the broader theoretical implications for the discipline as a whole.

4.1.1 Digitalisation of archaeological practice

The growing prominence of digital archaeology has fundamentally changed the way the archaeological discipline operates as a whole. As outlined by Morgen (2019) '[t]he digital has become pervasive, tedious, and worryingly invisible in archaeological labour, embedded in the craft of archaeological knowledge production' (p. 325). Correspondingly, some argue that '(...) there is no digital archaeology' (Huvila, 2018, p. 1), but instead archaeological research inherently incorporates digital tools and techniques into its practices. This perspective suggests that digital archaeology is not a standalone subfield, but rather an integral aspect of the broader archaeological process. With that in mind Huggett (2020) developed a model to illustrate the increasing role of digital tools in archaeological practice (figure 15). Traditionally, archaeology has been a hands-on and creative process, however digital technologies are now being integrated to standardise and streamline various tasks. This includes the use of consistent methods for data recording, which are then organised into systems that enhance efficiency. In some cases, tasks can be fully automated, with digital tools performing functions that were once carried out by the archaeologist. While these tools may assist the archaeologist in certain tasks, they can also take on more control as automation increases, shifting the balance of agency between the practitioner and the technology (Idem, p. 419).

This standardisation, systemisation, and automation of archaeological practice can be regarded as a significant development. Morgan (2019) even popularised the term "cyborg archaeology" which '(...) draws from feminist posthumanism to transgress bounded constructions of past people as well as our current selves' by '(...) using embodied technologies, we can push interpretation in archaeology beyond traditional,

skeuomorphic reproductions of previous methods to highlight ruptures in thought and practice' (p. 326). The concept posits that contemporary archaeological practice increasingly involves the integration of human expertise and digital technologies, creating a hybrid approach to knowledge production. In this framework, archaeologists are considered "cyborgs" in the sense that their work is inseparable from the tools and technologies they use. An archaeology that '(...) integrates posthuman principles to create a viable interstitial space where things from the past and from the present can commingle in commensurate space' (Morgan, 2022, p. 216).

In other words, digital technologies not only offer new insights, but they also challenge traditional methodologies, highlighting potential flaws within them. By addressing similar research problems through digital tools, archaeologists can test and refine existing methods, providing alternative perspectives and solutions. This approach allows for a more critical examination of past practices, helping to uncover biases, inaccuracies, or limitations that may have previously gone unnoticed.

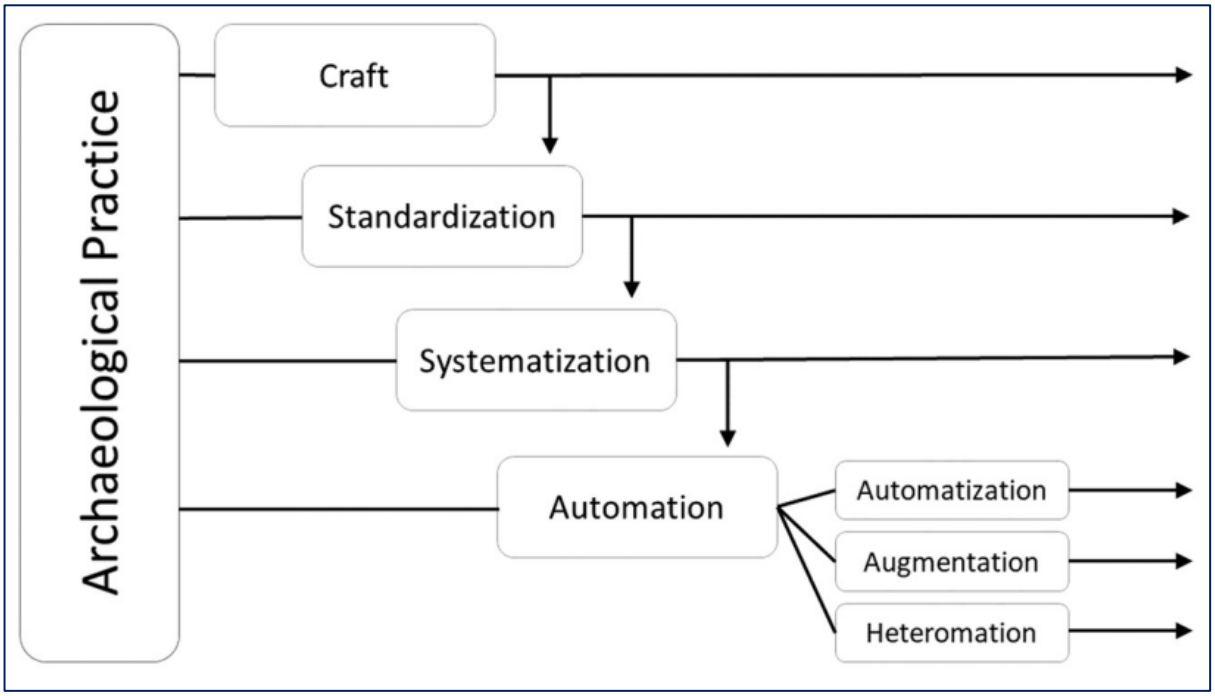


Figure 15: A model of archaeological practice (Huggett, 2020, figure 1).

4.1.2 Artificial intelligence, transparency, and ethics

Still, the digitisation of archaeological practice has resulted in some challenges and concerns. One of the major issues within archaeological Deep Learning research is that of transparency and the so-called "black-box" issue (Huggett, 2017; Huggett, 2021;

Gattiglia 2022; Tenzer *et al.*, 2024; Vadineanu *et al.*, 2024). 'A black box generates outcomes, but knowledge of how they arrive remains hidden. It is seen as a mysterious, inscrutable, powerful entity connected to a "data-driven algorithmic culture"' (Anichini & Gattiglia, 2022, p. 78). AI models are capable of generating predictions or classifications, yet the internal mechanisms by which these outcomes are produced often remain opaque to researchers and users. This becomes even more difficult with DL architectures, as these models '(...) essentially reprogram themselves as they learn from data, rather than the software being written in its entirety by human programmers' (Huggett, 2022, p. 9). As a result, researchers may struggle to explain why a model makes certain predictions or classifications, even though the algorithm itself might be highly effective in identifying patterns or solving problems.

The question one might then consider is whether it is reasonable to trust classifications and evaluations made by methods that operate in ways researchers cannot comprehend. Carabantes (2019) observes that transparency is a design principle frequently demanded of AI systems but less of human beings. Humans, too, are "black boxes" in many respects— '(...) we do not have access to other human beings' minds or brains, and we do not demand it to trust them. We only have access to their behaviour and their explanations, which are not a reliable depiction of the real mental processes' (p. 316). Therefore, it can be said that this challenge is not a new problem caused by AI, but instead has been a longstanding issue in the relationship between humans, technology, and knowledge production. As stated by Gattiglia (2022) 'the lack of explainability is merely an aspect of the ontological revealing-concealing dimension, not a novel concern in its own right, and archaeologists may be more cognizant of their own by paying attention to this dimension of technological mediation' (p. 330). This "human black box" is particularly relevant in this thesis, as the identification process of granaries, which this thesis relies on heavily, also can be found to be arbitrary and based upon subjective interpretations and experiential knowledge of the archaeologists involved. Although this thesis has tried to clarify which characteristics are oftentimes used to identify granaries, these factors are not codified into a transparent, reproducible framework, leaving much of the process reliant on individual expertise and tacit knowledge. This arbitrariness mirrors the opacity of AI systems, raising questions about the parallel challenges of

explainability in both human and machine contexts. Just as AI models operate as "black boxes," archaeologists' decisions in identifying granaries can lack explicit documentation making their reasoning similarly opaque to external observers. Consequently, this thesis grapples with a dual-layered black-box issue: the inherent complexity of AI systems and the implicit subjectivity embedded in human decision-making processes. These concerns highlight the challenges of transparency, accessibility, and trust in AI applications as well as archaeologists, emphasising that explainability is not merely a technical requirement but a foundational element for academic research.

Still, while human cognition and AI share similarities in their opacity, the lack of explainability in AI models, particularly within DL, raises concerns about their reliability and the ethical implications of relying on them without a clearer understanding of how they reach their conclusions. It has been argued that '(...) AI algorithms alienate archaeologists from their object of research and create further divisions between those who have the knowledge for implementing AI models and those who do not, creating inequality in the research. If the technology cannot be explained, archaeologists may also feel alienated from their role as knowledge-bearers. (...) When explainability threatens to alienate archaeologists from their ability of interpretation, there is an opportunity for them to become sceptical towards the use of AI' (Gattiglia, 2022, p. 330). Therefore, explainability has always been high on the list of concerns within archaeological AI research, as it directly impacts the trust and ethical acceptance of these technologies in archaeological practice.

Similarly, this ethical side is something that is often raised within AI research in archaeology. Simply put, '[t]ransparency is necessary for understanding biases in the data and the functioning of the algorithms. If the archaeologists cannot trust and verify that the AI algorithm has made a correct identification, the result cannot be used in research' (Anichini & Gattiglia, 2022, p. 78). When datasets are incomplete, biased, or skewed, the algorithms that rely on them can inadvertently perpetuate these biases in their classifications or predictions. If archaeologists cannot verify the integrity of the AI's decision-making process or understand the potential biases embedded within the data, they risk making conclusions based on flawed or incomplete information. This highlights the importance of ensuring that both the data used to train AI models and the algorithms

themselves are critically examined, ensuring that the resulting insights are as accurate and inclusive as possible. This data problem often referred to as the “garbage in, garbage out” concept, where, simply put, bad data will result in a bad model. As outlined by Kansteiner (2022): ‘[i]f we think that the stories and images we consume influence our memories, identities, and future behaviour, we should be very wary about letting AI craft our future entertainment on the basis of our morally and politically deeply flawed cultural heritage' (p. 124). In other words, if AI systems are trained on biased or incomplete data that reflects morally or politically flawed narratives, they could reinforce and perpetuate these issues, influencing not just archaeological research, but also broader cultural perceptions and future decision-making. This underscores the ethical responsibility of archaeologists and other researchers to ensure that AI tools are used thoughtfully, with an awareness of the potential consequences of relying on flawed data to shape our understanding of the past and the narratives we build about collective heritage.

Building onto the “(human) black box” and “garbage in, garbage out” issues, AI applications in archaeology often work with fragmentary and flawed datasets. The integration of AI into archaeological research has initiated more discussions about the nature of archaeological datasets and expertise. As outlined by Lambers (2024) AI models, particularly in supervised learning contexts, rely on benchmark datasets created by experts to classify archaeological phenomena. However, these benchmarks are often understood as a "gold standard." However, AI research has shown that ‘[w]hat has led experts to certain classifications in our benchmark datasets is often far from clear, transparent, or consistent, and this results in flawed datasets’ (Idem, p. 8). This has resulted in archaeologists reconsidering the reliability and objectivity of their own classifications, as AI models frequently expose inconsistencies or errors in the very data meant to validate them.

Fortunately, new fields are emerging both within and outside of archaeology to address and mitigate these issues. The discipline often calls that the ‘[r]esearchers across the discipline of archaeology should work closely with data scientists and social scientists to design representative sampling strategies and data gathering methods, and to develop protocols for assessing and correcting for bias in datasets’ (Tenzer *et al.*, 2024, p. 4). Furthermore, the introduction of Explainable Artificial Intelligence (XAI) is a step in the

right direction for improving transparency and accountability in (archaeological) AI systems (e.g. Barredo Arrieta *et al.*, 2020; Labba *et al.*, 2023; Li *et al.*, 2023; Matrone *et al.* 2023; Tenzer *et al.*, 2024; Vadineanu *et al.*, 2024). XAI aims to make AI models more interpretable, allowing researchers to understand and explain how decisions are made. Still, although the discipline is starting to understand better how AI systems function internally, the explainability of these systems is still far from straightforward. As a result, while XAI techniques provide tools for uncovering these processes, they often require specialised knowledge and may still offer only partial insight into the inner workings of complex systems. Huggett (2021) even states that '[w]here a digital device appears capable of explaining its reasoning, such as an explainable artificial intelligence providing some visibility of its underlying process, is it in reality creating a new black box through supplying a gloss that is understandable by the user?' (p. 425). An AI system might still be hiding its deeper complexities under a simplified version of its reasoning, thereby not fully addressing the problem of transparency. Therefore, creating a balance between transparency and usability remains a challenge in the integration of AI within archaeology. While XAI can offer more accessible explanations of how decisions are made, it may not always provide a complete or fully understandable picture of the underlying processes. Still, as there are many researchers dedicating their efforts to addressing these challenges, actively refining AI methodologies and to ensure that AI systems in archaeology are both transparent and ethically sound, ultimately contributing to more reliable and inclusive interpretations of the past.

Lastly, it is important to note that this thesis will not directly utilise XAI systems as a central focus, as the primary objective is to assess the feasibility of a DL model to detect structures on GIS excavation data. While transparency is a critical consideration for future applications, this research is grounded in evaluating the capabilities and limitations of AI systems for these purposes, without delving into the explainability or transparency at this stage.

4.1.1 Artificial intelligence, agency, and autonomy

Closely related to the black-box issue is the concept of artificial intelligence and research agency. 'Roles and tasks that were previously thought to be incomputable are beginning to be digitalised, and the presumption that computerisation is best suited to well-defined

and restricted tasks is starting to break down' (Huggett, 2021, p. 417). Although the question of agency has been raised before in relation to digital tools, AI has brought this debate to the forefront once again. As noted by Gattiglia (2022) '[a]rchaeologists do not directly control AI the way they control a total station; neural networks, once programmed, are internally autonomous' (p. 329). Where the concept of "agency" is a term traditionally tethered to human or actor-driven decisions, the introduction of autonomous systems, like AI, complicates its theoretical boundaries. This shift prompts a theoretical re-examination of the roles both researchers and machines play in the generation of knowledge. Huggett (2021) outlines this debate of algorithmic agency extensively, where they summarise their argument as follows:

> Ultimately, whether or not agency in the sense of a capacity for intentional and/or cognitive action can be legitimately associated with digital devices, agency can certainly be attributed to devices by humans, especially given the tendency to anthropomorphize them. (...) On that basis, if a device affects subsequent human actions and decisions then it can be said to have agency, even if that agency masks the human agency involved in the design and creation of that device. (...) Furthermore, human actors can be seen to share agency with devices where the task could not be done without the participation of the nonhuman components in what may be characterized as a symmetric or asymmetric relationship. (Huggett, 2021, p. 422)

Huggett chooses to err on the side of caution. Whether algorithms have inherent agency cannot be concluded, however their influence on human actions and decision-making is undeniable. Huggett's stance emphasises the symbiotic relationship between humans and AI, wherein the device's input is essential for completing tasks that would otherwise be impossible or impractical for humans to perform alone. This perspective reframes agency as a dynamic interaction, where both human and nonhuman elements contribute to the outcome, but it is the human designers and users who retain ultimate responsibility for directing and interpreting the process. Thus, while AI may exhibit forms of autonomy within certain contexts, it remains, at least in Huggett's view, a tool—one that extends human capacity, but does not replace the essential role of human agency in the production of knowledge and action. This idea is often echoed by other researchers in the field, however some go a step further, where Anichini and Gattiglia (2022) even state that '(...) AI algorithms have autonomy and intentionality; they require cognition and create a trace in the world. (...) In the AI age, archaeology's challenge is to recognise

technology as an agent on whom we depend for extracting meaning and, at the same time, as something that partially reflects our hermeneutics' (p. 81). They discuss the increasing role of AI in shaping how we interpret and understand the world. AI is not just a neutral tool but an active participant with its own (designed) intentions and autonomy, influencing how researchers access and interpret information. This idea shifts away from seeing humans as the sole interpreters, inherently acknowledging that AI now can play a significant role in the interpretation process.

The role of this debate in this thesis is to simply outline the effect that AI technology has on the archaeological discipline. If one where to create DL models that can accurately identify structures on archaeological excavation maps, then, one could argue, the research agency of this identification procedure is be shared between the human researchers who design and train the model and the AI system itself, which performs the identification task. In this case, the AI's ability to recognise patterns and structures could be seen as a form of computational agency, where the system contributes a significant level of autonomy in executing the task. However, this agency is ultimately shaped and constrained by the human input in the form of data, programming, and decision-making during the development of the model. The AI does not independently choose what to recognise but rather follows the directions embedded within its design, which are informed by human knowledge and intent. Therefore, the use of such a system underscores that the outcome remains heavily reliant on expert knowledge, as the AI's performance is ultimately shaped by the quality of the data and the guidance provided by human researchers. However, Huggett (2021) does raise concerns about the potential for digital technologies to disrupt the balance between autonomy and agency within archaeology, suggesting that these technologies may ultimately '(...) subvert and subdue human decisions, to the point where humans themselves may be shaped and used by the technology' (p. 423). Although this concept is clearly not the case in contemporary archaeology, they still highlight important aspects of the debate regarding algorithmic agency, accountability, responsibility, and ethics.

4.2 Archaeological space, place, and site mapping

Another aspect that has been impacted by digital archaeology and relevant for this thesis is archaeological site mapping. Therefore, the following paragraphs will discuss theories related to archaeological critical mapping and the conceptualisation of space and place. The mapping of archaeological features has been an important component within the research of archaeology. Unsurprisingly, as most of the data and remains retrieved by archaeologists are spatial in nature or include a spatial component (Wheatley & Gillings, 2003; Gillings *et al.*, 2020; Verschoof-van der Vaart, 2022). The introduction of Big Data and computer vision techniques in archaeology is transforming the perception of spatiality both practically and conceptually (Wheatley & Gillings, 2003, Zubrow, 2006; Bodenhamer, 2012; Dunn, 2017). This is predominantly due to the enhanced ability to create, process, and analyse large datasets, and the way in which maps are easily generated and visualised. Furthermore, debates surrounding the meaning and active agency of maps within archaeological research has revitalised the way in which maps are perceived and utilised. The following subchapters will discuss this alterations in perception, and will underscore some shortcomings of the dataset that is used within this thesis.

4.2.1 Digitisation of space and place

Before delving into the concept of mapping and the use of maps as a data proxy for archaeological research, it is essential to engage with the core theoretical debate of “space” versus “place”. This debate profoundly influences how spatial data is interpreted and its relationship to human experience and cultural significance. In short, the concept of “space" can be considered a more abstract, quantitative dimension, often associated with measurable and objective features. In contrast, the term "place" encompasses the experiential, subjective, and symbolic aspects, imbued with cultural meanings, memories, and social significance (Gillings *et al.*, 2020). This distinction is crucial for understanding the limitations and possibilities of spatial data in archaeology The concept of “space” in archaeology refers to the abstract and geometric framework within which physical objects and sites are located. It is fundamentally concerned with distances, coordinates, and the measurable aspects of the environment. Space is often perceived as a container within which human activities occur, an empty stage that can

be mapped and analysed using objective and quantitative methods. This perception aligns with a positivist approach, which emphasises the ability to observe, measure, and analyse phenomena in a detached and neutral manner.

Digital archaeology has transformed how archaeologists engage with this concept. GIS technology allows for the precise measurement and mapping of spatial relationships, creating detailed models of archaeological sites and landscapes. These models can reveal patterns and connections that are not immediately visible on the ground, providing new insights into how ancient societies organised their space. Traditionally, the handling of space in archaeology leaned towards positivist perspectives, emphasising objective measurements and quantifiable data. This means that the spatial component was seen as a neutral object of study, and therefore could be examined objectively. For example, the distribution of artifacts within a site can be mapped to understand the organization of activities, or the proximity of sites to natural resources can be analysed to infer settlement patterns. Such approaches rely on the assumption that space is an impartial framework, within which human behaviour and cultural phenomena can be systematically studied and understood. However, as underscored by Wheatley (2004) this perspective '(...) effectively substitutes a mathematical equation for the meaningful bit of human actions', and that archaeologists nowadays should '(...) recognise that the behaviour of human beings is not simply produced automatically from environmental stimuli' (p. 7). This means that one cannot study spatial relationships or handle spatial data, without considering the cultural, social, and cognitive contexts that influence human behaviour. 'In the case of space, one needs to distinguish clearly between spatial reality –that phenomena in which organisms, exist, move, and subsist - and the cultural construction of space. Even maps are not the disembodied view rather they are located in culture, space and time' (Zubrow, 2005, p. 2). Accordingly, a common thread in these critiques is the concern that maps, and their subsequent inferences, reduce the human cultural component to quantifiable variables, potentially oversimplifying the richness of human experience and cultural context. This reductionist approach risks interpretations that prioritise measurable data over the nuanced and multifaceted nature of human societies.

Consequently, the concept of "place" has been increasingly incorporated into digital archaeology, drawing heavily from its foundational use in landscape archaeology. The aforementioned perspectives often fail to capture the complexities of "place," extending beyond the mere physical dimensions of "space." Place is not just a location but a lived experience, filled with personal and communal significance, reflecting how individuals and communities perceive, navigate, and imbue meaning into their surroundings. In other words, space is not merely a physical container but as a construct imbued with social and cultural meanings; place is seen as the intersection of these meanings with specific locations. According to Blake (2007) '[a] human element is implicit in the very idea of place, of the conscious demarcation of space. World views emerge from, and are embedded in, the always-situated practices. This leads to a revalorisation of space, not as an inert backdrop, but as an active component of human activities and lifeworld' (p. 231). Or differently put, '[s]paces may be abstract, geometric and synchronous, but places have histories and biographies as well, and it is places that are inhabited by meaningful human actors' (Wheatley, 2004, p. 6). This abstraction of "place" goes hand in hand with phenomenological perspectives on the perception of human experience—a body of theory that underscores the significance of personal and embodied interactions with environments. Such theories emphasise how individuals and communities experience and interpret space not merely as a backdrop to events, but as an active, meaningful component of their lived experiences and cultural practices. Furthermore, it should be emphasised that place is not necessarily spatial, and many aspects cannot be mapped. Archaeologists have traditionally focused on the visible rather than the invisible, due to the inherent emphasis on tangible and measurable evidence in the discipline. However, what is considered visible or invisible extends beyond mere sight and into the realm of cultural perception. Something may be physically visible but not culturally recognised or understood (Daly & Evans, 2006).

While this debate does not fundamentally alter the core methodology of this thesis, it highlights important considerations for understanding the constraints and opportunities within the data. 'If space is a fluid, emergent, profoundly relational and highly contextual phenomenon, then the identification, representation and analysis of spatial patterns poses significant challenges that in turn require new methods to address' (Gillings *et al.*,

2020). In other words, is digital archaeology, a culmination of statistical analyses to study cultural behaviour, able to address these complexities of "place" and the nuanced nature of human experience within spatial contexts? Unfortunately, this thesis does not have the time, experience, and resources to fully grasp and solve this debate, as even the concept of time itself could be up for discussion. Still, acknowledging the constraints of digital approaches does not diminish their significance but rather underscores the importance of integrating these methods with a broader consideration of cultural and experiential dimensions of "place." The methodology in this thesis will ultimately acknowledge these limitations by focusing on the strengths of digital methods while recognising the need for future integration of qualitative and holistic analyses. Furthermore, when interpreting the results and conclusions that are drawn from this research, it is important to emphasise that the findings are rooted in the quantitative framework provided by digital tools and should be considered as part of a larger, ongoing discourse. The results offer insights into spatial patterns and relationships within archaeological contexts; however, they must be understood within the bounds of their methodological limitations. This thesis will present its findings with an awareness of these constraints, demonstrating how DL algorithms can effectively identify and detect prehistoric granaries. It will highlight the potential of these digital analyses to reveal spatial organisation and distribution patterns, while also acknowledging that they may not fully encompass the cultural and experiential dimensions inherent to the concept of "place."

In summary, while this thesis leverages the strengths of digital archaeology to analyse spatial data, it is mindful of the need for future research to address the complex interplay between space and place. The aim is to contribute meaningfully to the field while acknowledging that a complete understanding of human experience and cultural significance requires a multidimensional approach that extends beyond the capabilities of this methodology alone.

4.2.2 Mapping in archaeology

Building upon the above paragraph, archaeological mapping is another factor that is important to consider within the theoretical framework of this thesis. Researchers, both in the past and today, have explored various perspectives on the meaning, objectivity, and use of archaeological site maps. In general, there are two contrasting views in how

map-making is defined within archaeology, '[t]he first (...) rests upon the assumption that mapping is an objective data-gathering procedure. Mapping is, in this statement, a taken-for-granted part of archaeological fieldwork' (Flexner, 2009, p. 7) The second interpretation is more nuanced, '(...) rather than simply "recording", mapping "translates" and "mediates"; maps are subject to the subjectivities of their creators' (Idem, p. 8).

One of the main critiques of archaeological site maps is that they only record what is visible during the archaeological excavation. This limitation is not only problematic on a site level, but also on a broader scale. As emphasised by Wheatley (2004) any map '(...) that is based on the known distribution of archaeological sites is actually an embodiment of the visibility, bias, and historical accidents that have formed that record. Such a map is therefore predicting the bias in the known record' (p. 9). Maps, as rigid demarcations of archaeological features, suggest that their depiction is definitive and objective. However, in archaeology, this is never the case. The data used to create maps are a translation of various factors, including the archaeologists' interpretation, visibility, methodologies, and other contextual influences. This means that, maps should instead commonly be regarded as '(...) simplifications of reality – powerful simplifications – but simplifications, nevertheless, created according to rules of scale and projection. A perfect one-to-one map is a second reality and probably cannot exist. Digital maps are not the disembodied view from nowhere; rather they are located in culture, space, and time' (Zubrow, 2006, p. 18).

4.3 Key concepts in this thesis

The preceding discussions have provided an overview of theoretical issues pertinent to this thesis. The primary aim of this thesis is to address specific challenges related to the detection and identification of archaeological structures on GIS maps using DL algorithms. To achieve this, it is crucial to shift focus from theoretical discourse to practical implementation, exploring how these concepts translate into actionable methodologies and tangible results. Therefore, the paragraphs below outline and acknowledge direct issues and biases present within this thesis' methodology and dataset. This approach not only provides a clear understanding of the limitations and

potential impacts of the current research but also sets the stage for refining and enhancing future studies by identifying areas for improvement and potential avenues for further investigation.

4.3.1 Biases in the dataset mapping

When examining the data used within this thesis closely, it becomes apparent that biases permeate its origin and structure, even though it is collected with the intention of absolute objectivity and accuracy.

Firstly, the data is collected with a GPS system during an archaeological excavation. This means that it has been assembled by various individuals under different conditions and measurement systems, hence the data already fundamentally carries the signature of its creators. This is a common critique or footnote that is added to archaeological data in general: '(...) the data creator articulates their knowledge to identify and categorise information, and that information is atomized within a digital environment to create data. Data in these terms are therefore theory-laden, process-laden, and purpose-laden, and not raw in any sense' (Huggett, 2022, p. 59). Besides obvious ontological issues with making categorisations, such as imposing preconceived notions onto the data, there are also practical challenges related to data consistency and reliability. Variations in individual interpretation and measurement techniques can lead to inconsistencies within the dataset, further complicating the process of data analysis and interpretation.

Secondly, the data is subject to the limitations of the technology used for its collection. The accuracy of GPS systems can vary, and this variation can introduce additional uncertainty into the dataset. Furthermore, the GPS offers several methods in how certain shapes are measured during the excavation. In the context of the dataset collected from Archol bv, experience has shown that it is possible to both measure circular features, such as post-holes, with the spline option as well as a perfect circle. Where the first option is created through the use of multiple points that form a smooth curve, the latter is a geometrically perfect circle defined by a single centre point and radius. The choice between these two methods can significantly impact the recorded shape and size of the feature. The spline option, while more flexible, is subject to the individual's judgement in placing the points and can result in a less accurate representation if not done carefully.

On the other hand, the perfect circle option, while more consistent, may not accurately capture the true shape of the feature if it deviates from a perfect circle in reality. Furthermore, the circle option results in faster measurements, as the settings only have to be set one time, and, with that, multiple features can be measured in a quick and consecutive manner. Therefore, when there are multiple features of the same size present, as is the case with small structures, it is easy to measure these consecutively, resulting in a different outlook of structures on the extrapolated GIS map (figure 16).

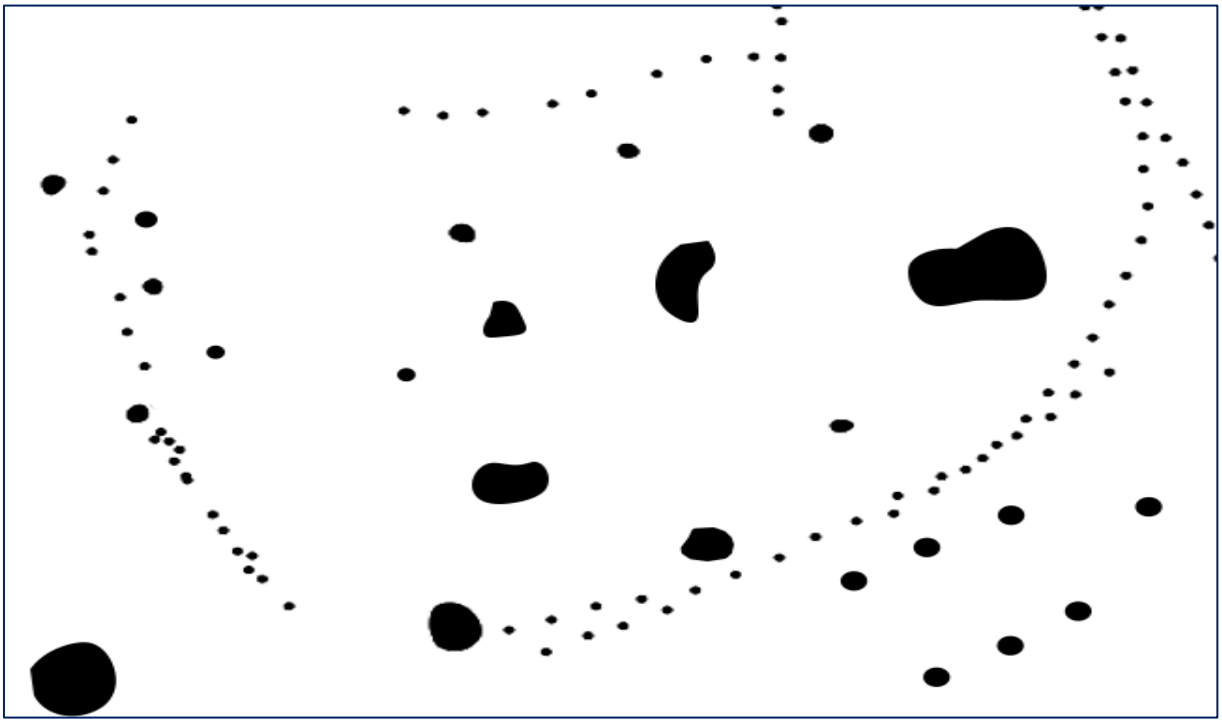


Figure 16: This image illustrates the contrast between the spline and circular feature settings of the GPS. The granary, located in the lower right corner, exhibits a distinct shape compared to the surrounding features measured using the spline method. This discrepancy could potentially bias the algorithm, causing it to prioritise this difference over other factors (Archol bv, n.d.).

Unfortunately, no clear guidelines are present during commercial excavations in the Netherlands, resulting in both methods being used interchangeably. Therefore, it is quite difficult to say how often the circular method is used as opposed to the spline function. All that can be said that the dataset seemingly had 84 instances where the circular method was used during the documenting of granary structures specifically. This approximates about 17% of the instances present within the entire dataset. This, in turn, might affect the algorithms selection criteria, as the archaeologists that already have preconceived notions about the features might unintentionally influence the data collection process. This bias could then be propagated through the algorithm, leading it

to favour certain shapes or sizes based on the initial input data, instead of looking at the structural characteristics itself. This problem goes hand in hand with the “black box” issue, as it cannot be said how reliant the model will be on this particular aspect. All in all, the impact of these discrepancies on the analysis will likely be minor, but it is something that should be discussed prior to the methodology to underline the inconsistency in the data collection.

Thirdly, the presence and exact location of archaeological features are often challenging to determine, a common issue with all archaeological data. This difficulty arises from various factors that affect visibility, including preservation conditions, the nature and age of the remains, the methodologies used during excavation, weather conditions, and the experience of the archaeologists involved. While strategies are oftentimes employed to mitigate these problems, the inherently fragmentary nature of the archaeological material record makes it impossible to completely eliminate these biases. This issue also goes hand in hand with the difficulty to recognise granaries in the field as well as later in the GIS environment. Although this thesis tries to offer a methodology to aid with this problem, there is a vicious circle that by training the model on imperfect data the model learns that these possibly unrecognised features are not features at all. This is a significant challenge as it can lead to a self-perpetuating cycle of bias in the model’s predictions. The model is trained on a dataset that includes only recognised features, and as a result, it learns to identify and classify these features as accurately as possible. However, if there are unrecognised features present within the dataset, the model may learn to classify these as non-features, thereby reinforcing the initial bias in the data. This dilemma underscores the importance of continual validation and refinement of the model, as well as the need for comprehensive data collection practices in the field. Thus, while every effort has been made to ensure the accuracy and comprehensiveness of the training data, the possibility of unrecognised features being misclassified cannot be entirely ruled out.

All in all, this shows that the data and methodology used within this thesis is biased by its very nature. The benchmark dataset used within this thesis is not a “golden standard”.

5. Methodology

5.1 Data collection and preparation

In the following paragraphs the methodology will be outlined that has been used for this thesis. The data cleaning has been done in a QGIS 3.26 environment. The practical implementation of this methodology will be conducted in a Python 3.11 environment. The choice is due to Python's extensive support for ML and DL libraries, making it a popular language for such tasks. Python 3.11 offers improved performance over previous versions, contributing to efficient execution of complex DL models. In this environment, key libraries, such as PyTorch for DL and data augmentation techniques, Ultralytics for YOLOv8s-specific functionalities, and other essential packages like NumPy, (Geo)Pandas, Matplotlib, and OpenCV for basic data manipulation and image processing will be used. All the datasets have been anonymised as instructed by the provider.

For the methodology, a workflow has been created to visualise the procedure developed in this thesis for the model (figure 17). This was done to provide a clear and structured overview of the process, ensuring that each step is transparent and reproducible. The structure of this chapter will therefore follow the steps outlined within this workflow.

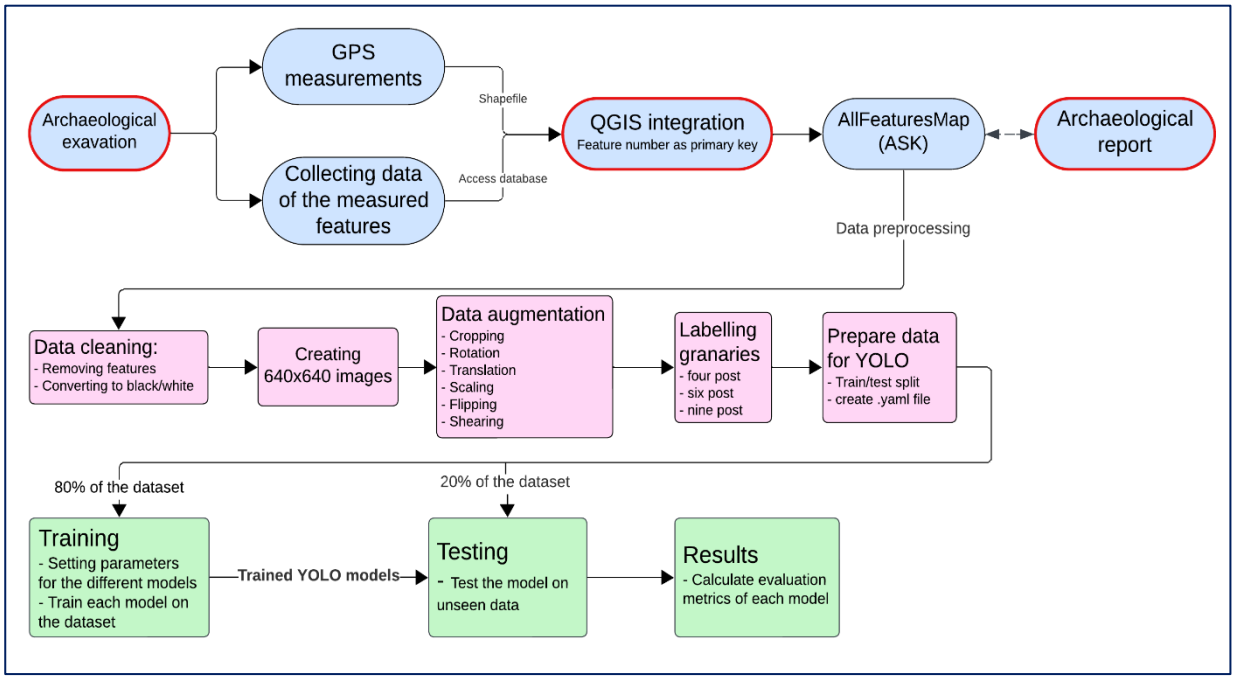


Figure 17: The workflow developed for this thesis. The blue boxes represent the steps carried out by archaeologists during the creation and collection of the data. The red borders indicate the identification phases of granary structures. The purple boxes illustrate the data preprocessing stages, while the green boxes depict the final implementation of the model.

5.1.1 Data cleaning and preprocessing

As the data was sourced from different excavations and shapefiles, ensuring consistency and quality across the dataset required comprehensive data cleaning and preprocessing procedures. This involved a series of steps to standardise the visual attributes of the images and eliminate any inconsistencies that could impact the accuracy of the DL model. This process has been annotated in the pink boxes in figure 17.

One of the primary challenges in working with multi-source data is the variation in colour profiles resulting from differences in GIS layout styles. While colour is often an important feature in many DL applications in archaeology, particularly those using remote sensing data where colour and spectral properties aid in classification, this research focuses on shape and spatial arrangement rather than chromatic information. Therefore, colour was deemed non-essential in this specific context.

To standardise the datasets and eliminate any potential variability introduced by differing colour schemes, all polygons were converted to black, and the background was set to white. This high-contrast approach enhances the visibility of the polygons and aligns with the research goal of analysing spatial patterns and geometric features, rather than relying on visual attributes such as colour or texture. By reducing the data to black-and-white, the preprocessing step simplifies the input for the model, focusing it entirely on the spatial and structural aspects of the polygons. This methodological choice is motivated by the data format that was used within this project, namely shapefiles, which in this case represent vector polygons. Since polygons do not inherently encode spectral or colour information beyond what is assigned for visual representation in GIS software, relying on colour would not provide meaningful input for the model.

To further refine the dataset, efforts were made to remove artifacts and noise that could interfere with the automated feature detection. This includes the removal of mapped irrelevant disturbances (e.g. fallen trees, animal based features, explicit modern features) that are clearly not related to the prehistoric granaries. However, features that were marked as “possibly natural” or had any other interpretative doubt were kept in the dataset, as they might have been unrecognised / degraded post-holes which could be relevant to the structure. Furthermore, features that had multiple sequential fillings due

to later modifications (as illustrated in figure 18) were analysed, and if secondary fillings were deemed irrelevant they were removed from the dataset. For instance, large pits which were dug to remove the posts were removed as they cannot be part of the original granary (figure 4). Although these fillings are generally important for archaeological interpretations, such as establishing the decline and recycling of construction materials, they are irrelevant for the purpose of this particular research. Furthermore, other data that was collected during the excavations but are not part of the detected features were removed. For instance, the demarcating trench borders, surface height measurements, coupe lines, and find locations were subsequently removed. Lastly, measurements that can be associated with the underlying stratigraphic horizon were also removed.

Finally, as the dimensions were sometimes too large within the geospatial datasets, the maps were split into sections of 640 by 640 pixels, as this is the default format to which the YOLOv8 architecture is compatible. An overlap of 10% was added to the images in order to avoid edge effects. All in all, the data cleaning ensures that the amount of noise on the maps was reduced, and thus the algorithm will be better capable of looking at the relevant aspects of the measurements. This will enhance the algorithm's ability to discern meaningful patterns and features within the data, ultimately improving its performance in its object detection task

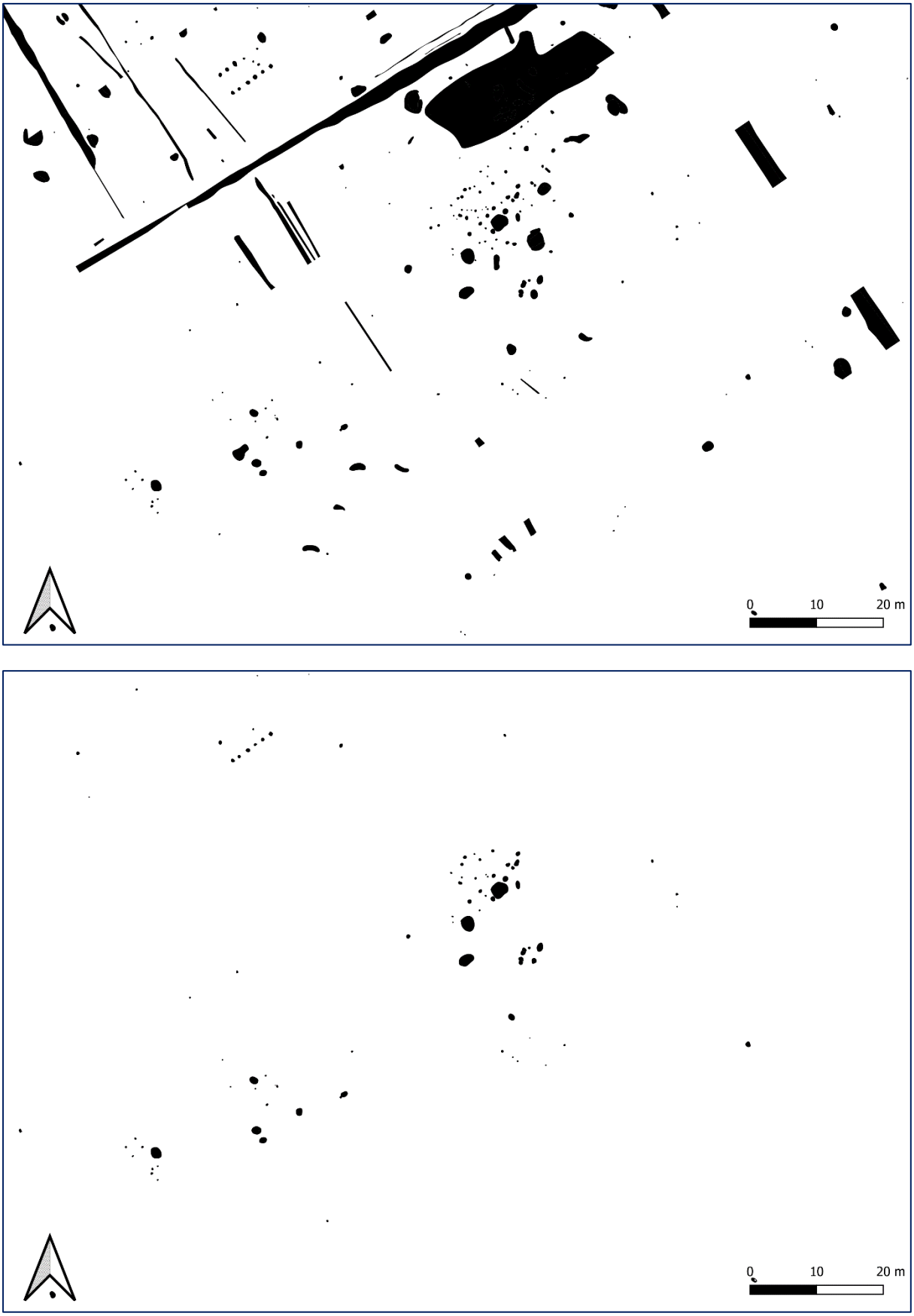


Figure 18: Example of the data cleaning process, showing the removal of irrelevant disturbances and secondary fillings. The image above is the original feature map, whereas the image below is the same location, but includes the data cleaning.

5.1.2 Data augmentation techniques

After the data cleaning process described above, the resulting maps were subjected to various data augmentation techniques to ensure a larger training dataset. In general, data augmentation techniques are applied to reduce the chances of overfitting the model to the limited training data, and thus to improve its versatility in handling unseen or novel inputs during inference.

There are many data augmentation techniques available which are relevant to all kinds of datasets (Haba, 2023). In this particular case, the dataset was subjected to the following data augmentation techniques (figure 19):

1. Cropping: can simulate variations in the scale and position of structures within the image. By randomly selecting different regions of the image to retain and discarding the rest, cropping can effectively alter the composition and layout of the granaries. Cropping focuses the model's attention on specific regions of interest within the image, potentially improving its accuracy in detecting granaries of varying scales or in cluttered environments.
2. Rotation: helps the model become invariant to the orientation of structures. By rotating the images by various degrees, the model is exposed to the granaries from different angles, improving its ability to generalise.
3. Translation: involves shifting the cleaned image along the x and y axes. This can simulate changes in the position of granaries within the image, making the model more robust to variations in their location.
4. Resizing: Scaling alters the size of the structures relative to the image. You can randomly resize the images, making the structures appear larger or smaller, which helps the model learn to detect structures of different sizes.
5. Flipping: Horizontal or vertical flipping can provide additional variations in the appearance of structures. This helps the model learn to recognise structures regardless of their orientation.
6. Shearing: Shearing involves skewing the image along one of its axes. This can simulate perspective distortions and variations in the viewpoint of the structures

These techniques were implemented with the popular computer vision OpenCV library[2] which easily applies these transformations to the input data. Leveraging OpenCV's functionalities resulted in a diverse set of training images with varying dimensions, rotations, and other transformations. The total number of training images generated through these augmentation techniques was 1201. All in all, although more is oftentimes better, this amount was deemed sufficient enough for accurate training in this context.

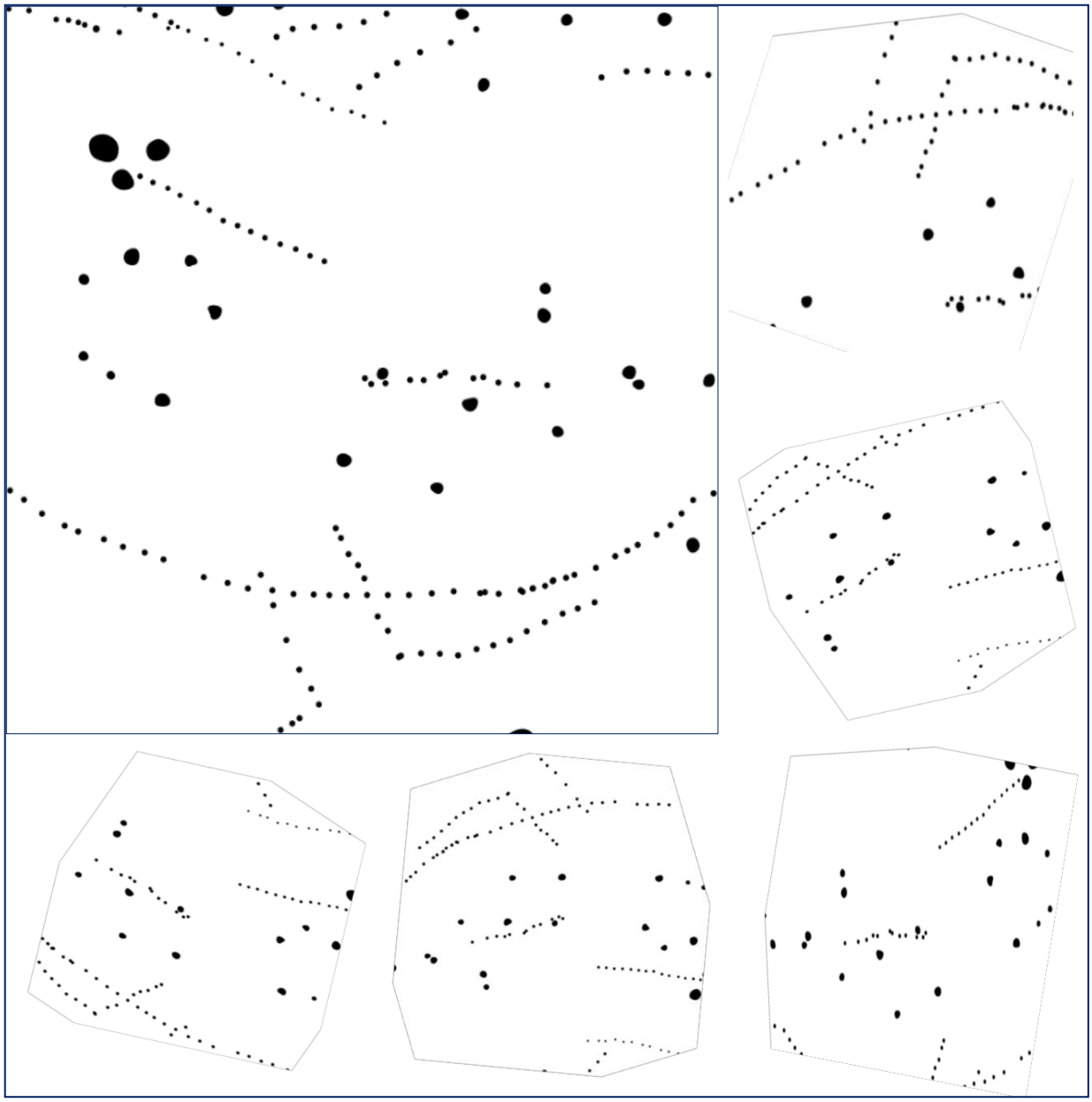


Figure 19: Example of the data augmentation process, where the upper left shows the original image, and the other images display augmented versions.

[2] The library by the OpenCV team can be found here: https://github.com/opencv/opencv

5.1.3 Data labelling

After the data cleaning and augmentation the data labelling process was carried out manually. As mentioned in several discussions above, these labels are based upon the publication of the excavation sites. Therefore, it can be assumed that the detected granaries are somewhat competently identified, although a completely accurate result might always be elusive (a comprehensive debate on this can be found in chapter 4). In order to have a relatively efficient data labelling process, the tool "LabelMe" was used to streamline this data preparation process[3]. This open source tool is specifically designed for computer vision tasks and compatible with YOLOv8. Furthermore, as opposed to other popular labelling tools, this particular software is able to label the examples in a polygon format, instead of constricting bounding boxes. This means that the tool is able to annotate the image with more accuracy, and is able to deal with the rotation of the structures that is present in the dataset (see figure 20).

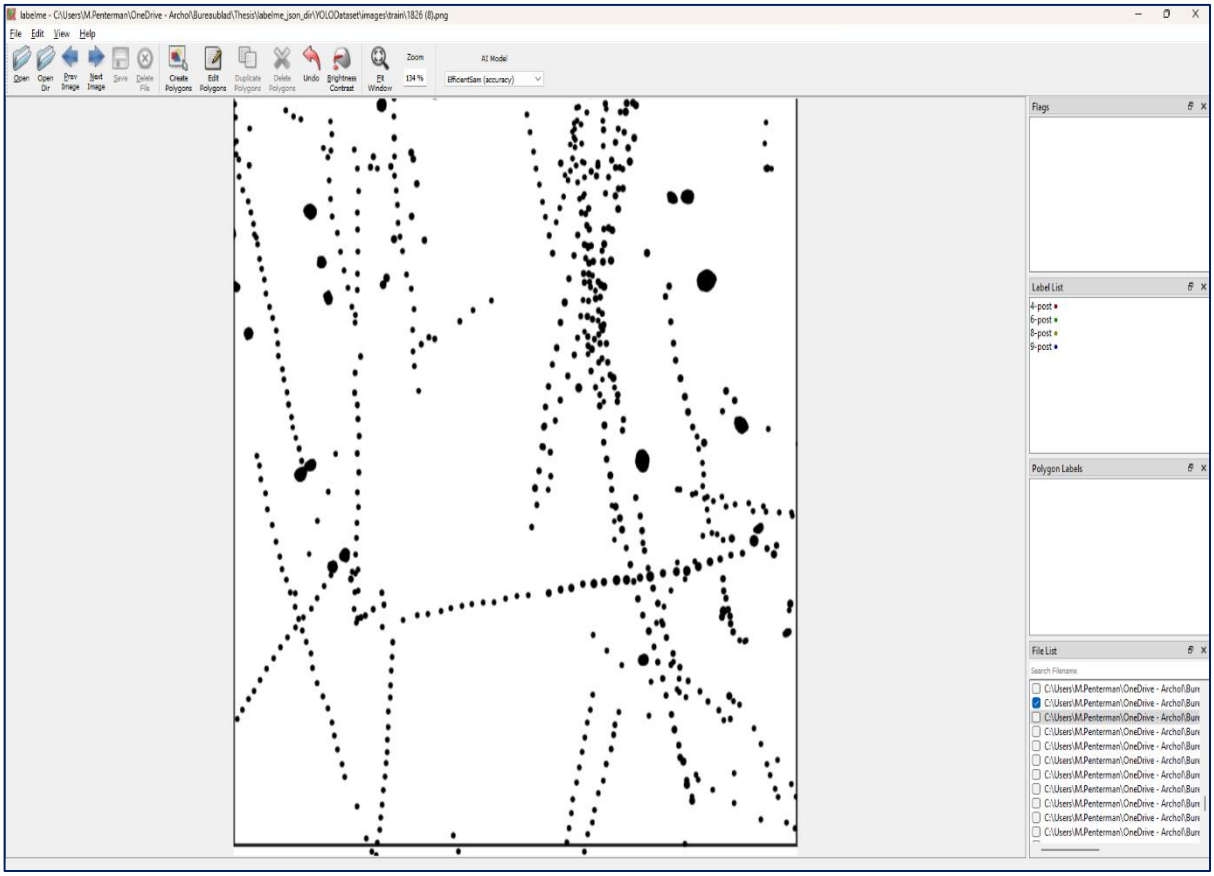


Figure 20: Example of the LabelMe interface incorporated into the integrated development environment of Visual Studio Code (Russel, 2024).

[3] This tool, maintained and created by Russel can be found here: https://github.com/labelmeai/labelme

A total of three different categories were labelled within the dataset: Granary (4), Granary (6), and Granary (9). Five-post and eight-post granaries were not considered due to their absence in the dataset. Ultimately, the total number of instances can be seen in table 2:

Labelling category	Number of instances	Percentage of the dataset
Granary (4)	382	80.1%
Granary (6)	59	12.3%
Granary (9)	36	3.6%
Total	477	100%

Table 2: Number of labelled instances for each granary label within the resulting training dataset. Where it is clear that some labels are more prevalent than others.

In practical terms, the labelling process involved creating polygons around the post-holes of the granary. These post-holes were identified using the metadata in the dataset, where each post-hole in the shapefile was assigned a structure number corresponding to a specific granary (Figure 14). This process required comparing the GIS map with the corresponding images, and then labelling the post-holes based on their alignment and identification in both sources. The labelling entailed precisely positioning points around each post-hole to delineate its boundaries within the polygon. Typically, three or four points were placed for each post, ensuring the circular shape of the hole was represented. To maintain consistency this process was kept similar to the greatest extent possible (figure 21).

Granaries that were incomplete due to the data augmentation methods (which was the case in 581 images) were not labelled, as these partial representations could potentially mislead the model during training. This decision aimed to maintain the integrity and reliability of the labelled dataset, ensuring that only representative instances were included for training the model. However, even though this seems unfortunate, many of these images were used as negative examples within the training dataset. The determination of whether these images were considered negative examples primarily depended on the visibility and extent of the features captured. If the granary structure was sufficiently obscured or fragmented to the point where it could not be confidently classified as a complete instance, those images were included as negative samples. For example, if the image only displayed one post-hole out of the original four, it would be

labelled as a negative sample due to the incomplete representation of the granary structure. This ensures that the model is exposed to a diverse range of scenarios where the target objects are absent, helping it to learn to distinguish between the presence of granaries and irrelevant background features.

Similarly, granaries identified by archaeologists during excavation, yet displaying missing features due to disturbances or other various factors, underwent a different labelling strategy. In these instances, if a missing post happened to be one of the corner posts of the granary, the instance was excluded from labelling. This decision was primarily influenced by the significant alteration it would cause to the structure's shape.

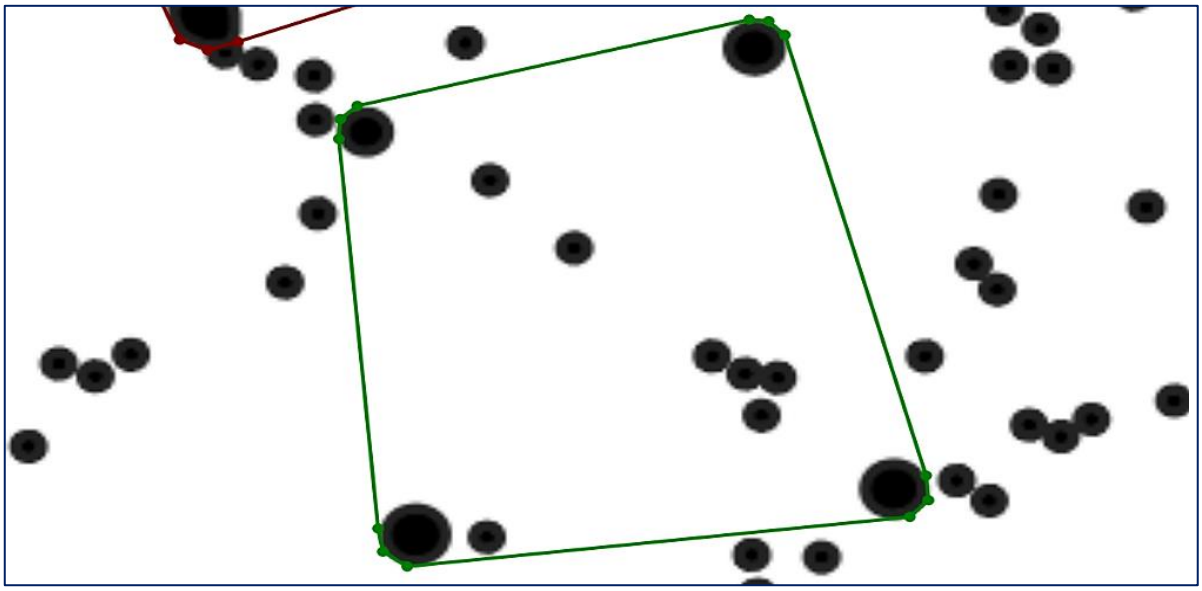


Figure 21: Example of the labelling of a four-post granary. The points of the polygon are carefully placed around the post-holes to ensure that the entire feature is within the polygon.

Conversely, if the missing post occurred along the longitudinal side, the image was labelled, as the fundamental shape of the structure remained intact within the polygon. While this strategy of excluding instances where corner posts are missing from granary labelling is deemed necessary to maintain structural integrity and facilitate model training, it presents a significant limitation. Unfortunately, this approach is less than ideal due to the prevalence of granaries with missing posts in the archaeological record. Granaries exhibiting degrees of deterioration, disturbances, or structural alterations are abundant, making it challenging to ignore such instances during labelling. Consequently, the exclusion of granaries with missing corner posts may result in a dataset that inadequately reflects the full spectrum of archaeological realities. This limitation underscores the complexity inherent in archaeological data annotation and highlights

the need for nuanced approaches to accommodate the diverse range of structural conditions observed in the archaeological record. Still, for this research the consideration of structural completeness during labelling was deemed necessary to ensure the reliability of the dataset for training archaeological models.

Overlapping structures were carefully addressed during the labelling process to ensure accurate representation within the dataset. When multiple structures overlapped in an image, each structure was still individually annotated to delineate its boundaries, even if partially obscured by other features. This approach would ensure that the model could possibly distinguish and understand the spatial relationships between overlapping structures, thereby improving its ability to accurately identify them. Ignoring overlapping structures would be problematic, as the archaeological record often presents complex scenarios where structures intersect or overlap spatially. Failing to account for these overlapping features could lead to incomplete or inaccurate interpretations of the archaeological context.

All in all, the labelling of the dataset is never as straightforward as it sounds, and strategies have to be employed to maintain a relative standardised workflow throughout the process. As is the case with all archaeological models, despite the cautious efforts during the data labelling, achieving complete accuracy in identifying and categorising archaeological features will remain impossible due data incompleteness, ambiguity, and subjectivity. Furthermore, the choices and strategies implemented during the labelling process inevitably introduce biases, whether implicitly or explicitly. However, with this strategy in mind, the final amount of images that were created with the labelling as well as the augmentation step can be seen in table 3 below:

	Number of images	Percentage of the dataset
Successfully labelled	620	51.6%
Negative examples	450	37.5%
Excluded images	131	10.9%
Total	1201	100%

Table 3: The amount of images created during the image augmentation and data labelling steps, detailing the total number of images included in the final dataset. This table outlines the various categories of images, including those that were successfully labelled, those excluded due to incompleteness, and those utilised as negative examples.

5.2 YOLOv8 algorithm development

5.2.1 Used model architecture

The specific architecture used for this thesis is the YOLOv8s detection model[4]. This model has been selected because it can be considered the best balance between speed and accuracy within the context of this particular learning problem (table 4). This is ideal as this project does not include too many training examples as opposed to other case studies. Larger models, such as YOLOv8m, YOLOv8l, and YOLOv8x, require quite a lot of computational power, which will drastically increase the training time. These models are therefore often used in applications that require high accuracy and where computational resources are not a constraint. Therefore, these larger models are more suitable for instances with convoluted datasets and scenarios where there are numerous object classes, complex backgrounds, or a high density of objects in each image. Although the research problem appears to have some of these characteristics, the overall level of complexity and the number of training examples are relatively moderate compared to other case studies. Consequently, the dataset and problem scope do not justify the resource demands of larger models. Given these factors, YOLOv8s is more suitable, providing a balance between performance and efficiency without overburdening the computational resources or extending the training time excessively. Therefore, while YOLOv8x or YOLOv8l might be required for more demanding tasks, YOLOv8s is adequate for this thesis's objectives, delivering efficient object detection without compromising on speed or requiring high-end computational hardware.

On the other hand YOLOv8n, is designed for scenarios where low computational power and fast inference times are paramount. However, this model has been excluded from this thesis due to its lower accuracy compared to YOLOv8s. Although YOLOv8n is optimised for speed and minimal resource usage, it often sacrifices some detection performance, making it less suitable for tasks that require a higher degree of precision or involve a moderate number of training examples. Given the research objectives of this thesis, which prioritise a balanced approach to speed, accuracy, and computational efficiency, YOLOv8s emerges as a more appropriate choice. Its enhanced accuracy over

[4] The model and more detailed information can be found here: https://docs.ultralytics.com/tasks/detect/

YOLOv8n ensures that the detection outcomes are reliable, while its computational requirements remain manageable. This balance aligns better with the dataset's size and the need for dependable object detection.

Model	Size (pixels)	mAP^{val}	Speed ONNX	Speed TensorRT	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

Table 4: Different model architecture of the YOLOv8 detection model (Ultralytics, 2024). These detection models are all pretrained on the COCO dataset.

1. Size: the input image resolution, consistent at 640 pixels for all models.
2. mAP^{val}: Mean Average Precision on the validation set, measuring detection accuracy; higher values indicate better performance.
3. Speed ONNX / TensorRT: Inference speed in frames per second when using the ONNX or TensorRT format; higher values indicate faster processing.
4. Params: number of parameters in the model, reflecting its complexity; larger models typically offer better accuracy.
5. FLOPs: Floating Point Operations per second, indicating the computational demand

5.2.2 Transfer-learning

The YOLOv8s algorithm is commonly pretrained on the COCO dataset. The COCO (Common Objects in Context) dataset is a large-scale object detection dataset widely used in the field of computer vision. It contains over 330,000 images, including more than 200,000 labelled images with over 1.5 million object instances across 80 object categories. COCO is designed to enable the development and evaluation of algorithms capable of detecting objects in complex scenes. The dataset is characterised by its diversity, containing objects in various contexts and backgrounds, which helps models trained on COCO generalise well to real-world scenarios. This diversity has established COCO as the standard dataset for transfer learning in computer vision research. Although this might seemingly no be relevant to this particular context, as the data and subsequent labelling categories are highly dissimilar to the goal of this thesis, it is still

commonly understood to be a helpful foundation for any computer vision algorithm. The pretraining on COCO helps the model develop a general understanding of object features, patterns, and relationships, which can then be fine-tuned to the specific requirements of a new task. This approach significantly reduces the amount of task-specific data needed for training and accelerates the convergence of the model, leading to improved performance even in specialised applications like this thesis. Therefore, the models that can be employed through the Ultralytics packages are already pretrained on this dataset.

Furthermore, training the model on other instances of archaeological data might also be beneficial for this case study, as it could further refine the model's ability to work with archaeological datasets. However, unfortunately, this is currently outside the scope of this thesis. For now, there are no readily available models specifically tailored to this niche application, which necessitates the use of more general pretrained models like YOLOv8s. Future research could explore the development of specialised datasets and models for this purpose, potentially improving detection accuracy in archaeological contexts.

5.2.1 Optimising recall instead of precision

Archaeology uses deep learning somewhat differently compared to other major disciplines. When creating deep learning models, the designer must always choose their parameters based on the specific context and objectives of the task. As a result, in many fields, precision is oftentimes prioritised over recall to minimise the risk of false positives and ensure that only highly probable instances are identified. For instance, in medical diagnostics, reducing false positives is critical to avoid unnecessary treatments or interventions. However, in archaeology, the priorities seemingly shift.

In automated object detection within the discipline of archaeology this conventional preference for precision is somewhat inverted. Here, the cost of missing a relevant finding—such as a prehistoric granary—can have a greater impact on the overall understanding of a site or landscape than incorrectly identifying a non-granary feature as one. Archaeologists tend to place greater value on ensuring that all possible features are flagged for further investigation, even at the risk of false positives. In the case of this

thesis, this concept is echoed, as the objective is to ensure that all potential granaries are identified for further analysis, allowing expert interpretation to guide the final conclusions.

To prioritise recall in these models, several measures were implemented to ensure the identification of as many potential granaries as possible, even at the cost of some false positives. First, the confidence threshold was set relatively low across the models. By decreasing this threshold, the model becomes more sensitive to detecting objects, allowing it to flag a larger number of potential granaries, even if they are less certain. This reduces the likelihood of missing a relevant finding. Additionally, the IoU threshold was set at a moderate level to balance detection accuracy and recall. A lower IoU threshold means the model can classify overlapping or near-miss objects as potential granaries, which is important in archaeological contexts where features might not always be perfectly delineated. Finally, weight decay values were adjusted to prevent overfitting, ensuring that the model generalises well across various data inputs, further supporting recall by not overly favouring false negatives. The combination of these adjustments reflects a deliberate strategy to maximise the number of features detected for further investigation.

5.2.2 Training procedure

During the training of the model, several choices were made to ensure optimised performance and an effective learning process. First of all, a train-test split was performed on the dataset to accurately assess the model's ability. This split was done according to the common 80-20 split (856 x 214 images). This was done with the random sampling technique that, simply put, shuffles the dataset and randomly arranges each image in their respective category. The training set was used to teach the model, allowing it to learn patterns and relationships within the data, while the test set provided a definitive evaluation of the model's performance.

Ultimately, a total of three models were trained in order to assess whether certain parameters were more suitable to this particular research context (table 5). The following parameters were altered over the course of the training procedures to analyse the effectiveness of each respective model:

A. Epochs: The number of times the entire training dataset passes through the model. More epochs can improve performance but also risk overfitting.
B. Batch Size: The number of training samples processed before the model's weights are updated. Larger batch sizes can lead to more stable training, while smaller batch sizes can make the model more responsive to changes in the data.
C. IoU Threshold: The threshold for deciding whether predicted bounding boxes overlap sufficiently with ground truth boxes. Adjusting this can affect precision and recall.
D. Confidence Threshold: The minimum confidence score required for a detection to be considered valid. Tuning this helps balance between false positives and false negatives.
E. Weight Decay: A regularisation technique that adds a penalty for larger weights to prevent overfitting and improve generalisation.

This list is not intended as an exhaustive overview of all the parameters that can be altered within a DL model, instead these specific settings were chosen as they are the most relevant to this particular case study. Furthermore, to keep the model's training procedure within the limited scope of this research, some parameters were not introduced due to constraints in time and computational resources.

All in all, as there is no clear baseline or set of rules to which a model must conform to function appropriately, the best approach to developing a DL model is through trial and error. Therefore, this thesis developed three trained models that are designed to address different levels of caution and thoroughness in detection. As outlined in the chapters above, the model will aim to optimise recall while balancing precision, as it is preferable in archaeological research to find more potential detections, even at the cost of introducing some false positives. The nature of archaeological work often involves the identification of subtle and infrequent features, where the risk of overlooking structures can have considerable implications for research outcomes. Therefore, a model that errs on the side of caution is essential. This also complements the idea that this model is intended to serve as a valuable tool alongside human expertise rather than as a definitive ground truth.

The parameters for each respective model can be seen in table 5 below. Furthermore, the specific code that was used in the python environment can be seen in appendix 1.

Parameter	Model_1	Model_2	Model_3
Epochs	100	100	150
Batch size	16	32	32
IoU threshold	0.4	0.5	0.5
Confidence threshold	0.2	0.3	0.4
Weight decay	0.0001	0.0005	0.001

Table 5: The outline of the different settings per parameter for the three models developed in this thesis. Each model is configured with varying values for epochs, batch size, learning rate, IoU threshold, confidence threshold, dropout rate, and weight decay.

5.2.3 Testing procedure

After the training procedure, it is important to assess the models' overall performance. This is done during the testing phase. The testing is an important aspect of the development process, as it serves as a final checkpoint to thoroughly understand the models performance. By critically monitoring the models' behaviour on the testing set, it can be ensured that the model is not only learning effectively but also generalising well to unseen data. Additionally, the test set simulates the real-world scenario where the model encounters data it has never seen before. As mentioned in the training procedure, the dataset was split into training and test sets using an 80-20 ratio, with 856 images for training and 214 for testing. The testing was selected using random sampling to ensure a diverse and representative subset of the dataset. This random sampling was deemed appropriate to prevent bias in the testing process and to ensure that the test set accurately reflected the distribution of the full dataset.

As discussed in chapter 2.2.2, the evaluation metrics used to assess the models' performance include precision, recall, and mAP^{50}. These metrics were calculated to quantify the models' strengths and areas for improvement. Precision and recall provide insight into the models' performance in terms of false positives and false negatives, respectively, while mAP^{50} offers a summary metric that balances precision and recall across different threshold levels. These metrics together form an overall evaluation framework, enabling a thorough assessment of the model's performance.

6. Results

This chapter presents the testing results obtained from the three models trained to detect and classify instances of granaries. The performance of each model was evaluated using precision, recall, and mean Average Precision at 50% Intersection over Union (mAP^{50}) across different classes of granaries, labelled as Granary (4), Granary (6), and Granary (9). For examples of the model's actual predictions, please refer to appendix 2, where a selection of predicted images from model_3 are presented to provide an overview of its practical performance. Additionally, a more detailed analysis of specific cases and observations will be discussed in chapter 7.

6.1 Evaluation model_1

Model_1 was trained for 100 epochs and tested on a set of 214 images (20% of the entire dataset) containing 213 instances of granaries across all classes and 78 negative examples. The overall performance metrics for this model are in table 5 below. The model completed the 100 epochs in 0.485 hours, with an average processing time per image of 0.9ms for preprocessing, 6.9ms for inference, and 1.5ms for post-processing. The overall evaluation metrics can be seen in table 6.

	Images	Instances	Precision	Recall	mAP50
Granary (4)	214	157	0.909	0.624	0.804
Granary (6)	214	50	0.862	0.62	0.793
Granary (9)	214	6	1	0.99	0.995
All classes	214	213	0.924	0.745	0.864

Table 6: The overall performance metrics calculated during the testing of model_1. Here each labelled class is represented, and the overall summarised metrics for all classes combined.

In figure 22 the precision-recall (PR) curve was plotted. The pr-curve is a plot that visualizes the trade-off between precision and recall for different threshold settings of a classification model. It is particularly useful for evaluating models on imbalanced datasets, as it highlights how well the model identifies positive instances while balancing false positives and false negatives. With model_1 the curve reveals that the model exhibits high precision at lower recall levels. Furthermore, the model performs best with

the Granary (4) class and has poorer performances with the other two. More specifically, a lower area under the curve (AUC) suggests that the model has lower precision and recall compared to the curve with a higher AUC-PR, meaning it is less effective at identifying positive instances accurately across various threshold settings with these particular categories.

Furthermore, the Granary (9) class seems to be invisible due to the line being plotted on perfect precision and recall inference, meaning that the precision and recall have been nothing other than 1.0. This indicates that, theoretically speaking, the Granary (9) category performs in a perfect manner, however, as will be discussed in chapter 7, these metrics are not indicative of a perfect performance, but are probably the result of a low instance amount and overfitting of the training model.

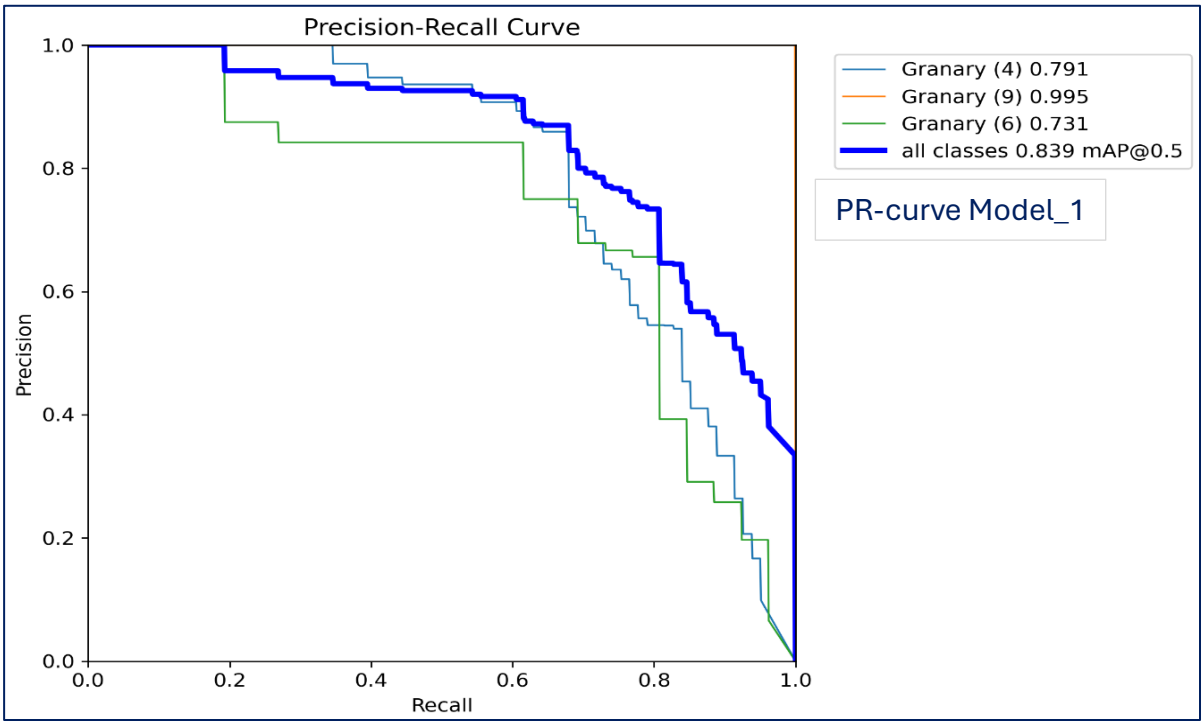


Figure 22: The precision-recall curve plotted for model_1. As can be seen, the model performs better with the class Granary (4) as opposed to other categories.

Another aspect that can be seen from the AUC-PR is the apparent decline in precision when stepping around the 0.8 recall. It seems to be the case that all the classes, as well as the average, decline downwards. This indicates that the model struggles to maintain high precision as it attempts to capture more true positives, suggesting that the instances it identifies beyond this recall threshold are increasingly likely to be false positives. The steep decline in precision at this point highlights a trade-off where improving recall

comes at the cost of accuracy in identifying true positives. This could imply that the model's performance is better at lower recall levels but starts to falter as it tries to generalise further. This trend can inform decisions about where to set the threshold depending on whether precision or recall is more critical for the task at hand. Therefore, in simpler terms: as the model becomes more aggressive in identifying positive cases, it also starts getting more unreliable, and more of its positive predictions turn out to be wrong.

A raw and row-wise normalised confusion matrix for model_1 were generated based on the testing procedure (figure 23 and 24). This raw matrix provides an overview of the predicted versus actual classifications for all instances in the testing set. The results show that 119 instances were correctly identified as four-posted granaries (TPs). However, the remaining instances were misclassified, either as Granary (6) or, more frequently, as negative examples (background) (FNs). Furthermore, the model correctly identified 6 instances of nine-posted granaries (TPs), and 1 as a Granary (4) and 6 as negative examples (FNs). For the Granary (6) class, 38 instances were accurately detected (TPs). Among the misclassifications, 7 were incorrectly labelled as Granary (4),

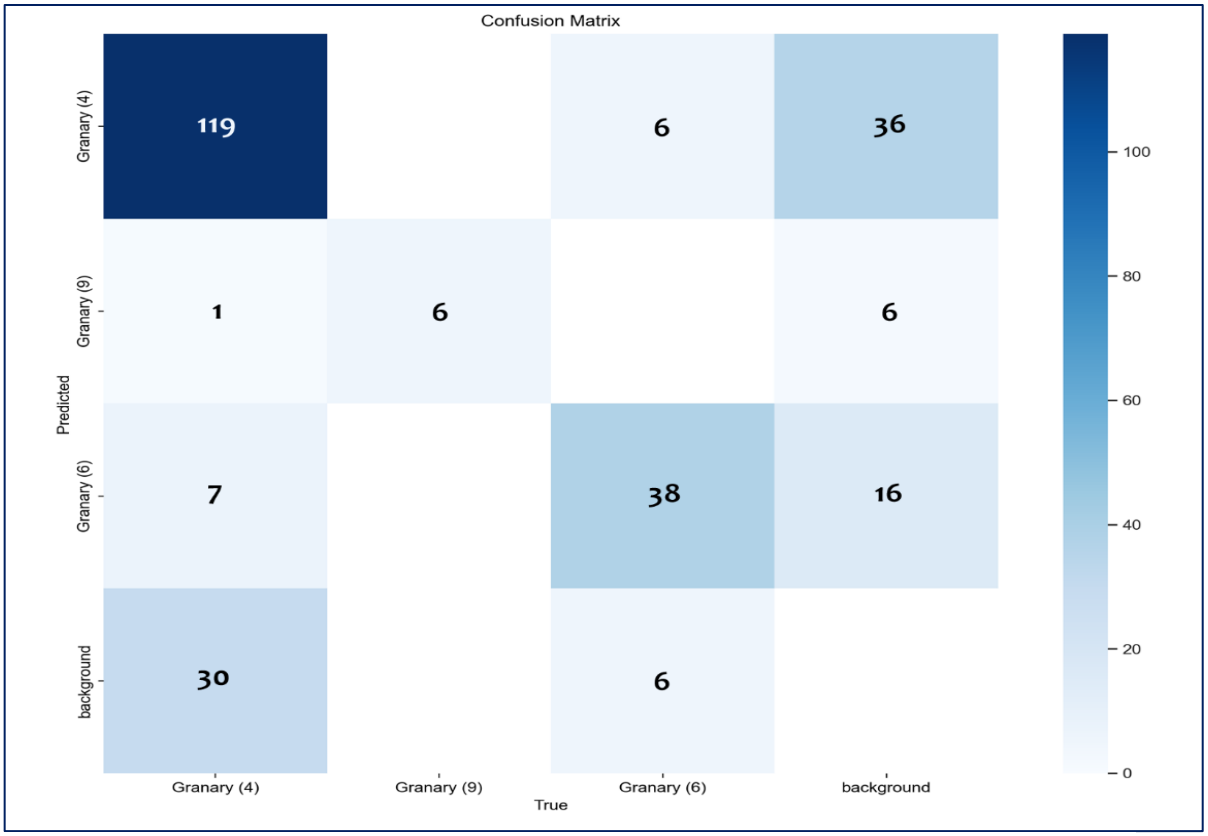


Figure 23: The raw confusion matrix plotted for model_1.

and 16 were mistaken for negative examples (FNs). Additionally, the model incorrectly classified 30 background instances as Granary (4) and 6 background instances as Granary (6) (FPs). In total, 36 instances were falsely predicted as granaries when they were actually background. Lastly, as can be seen, the [background, background] cell shows 0. However, this does not reflect a failure of the model to recognise background regions. Instead, it is due to the fact that the confusion matrix does not track true negatives (TN) for the background class.

The normalised matrix presents similar results, but with percentages rather than raw numbers. This allows for clearer insights into the model's performance. The percentages highlight that the model generally achieves a high rate of positive predictions for each class. However, it also reveals a significant issue: a relatively high percentage (0.67) within the predictions, which are background are incorrectly predicted as Granary (4). Perhaps, the same can also be said with the Granary (6) class being misinterpreted due

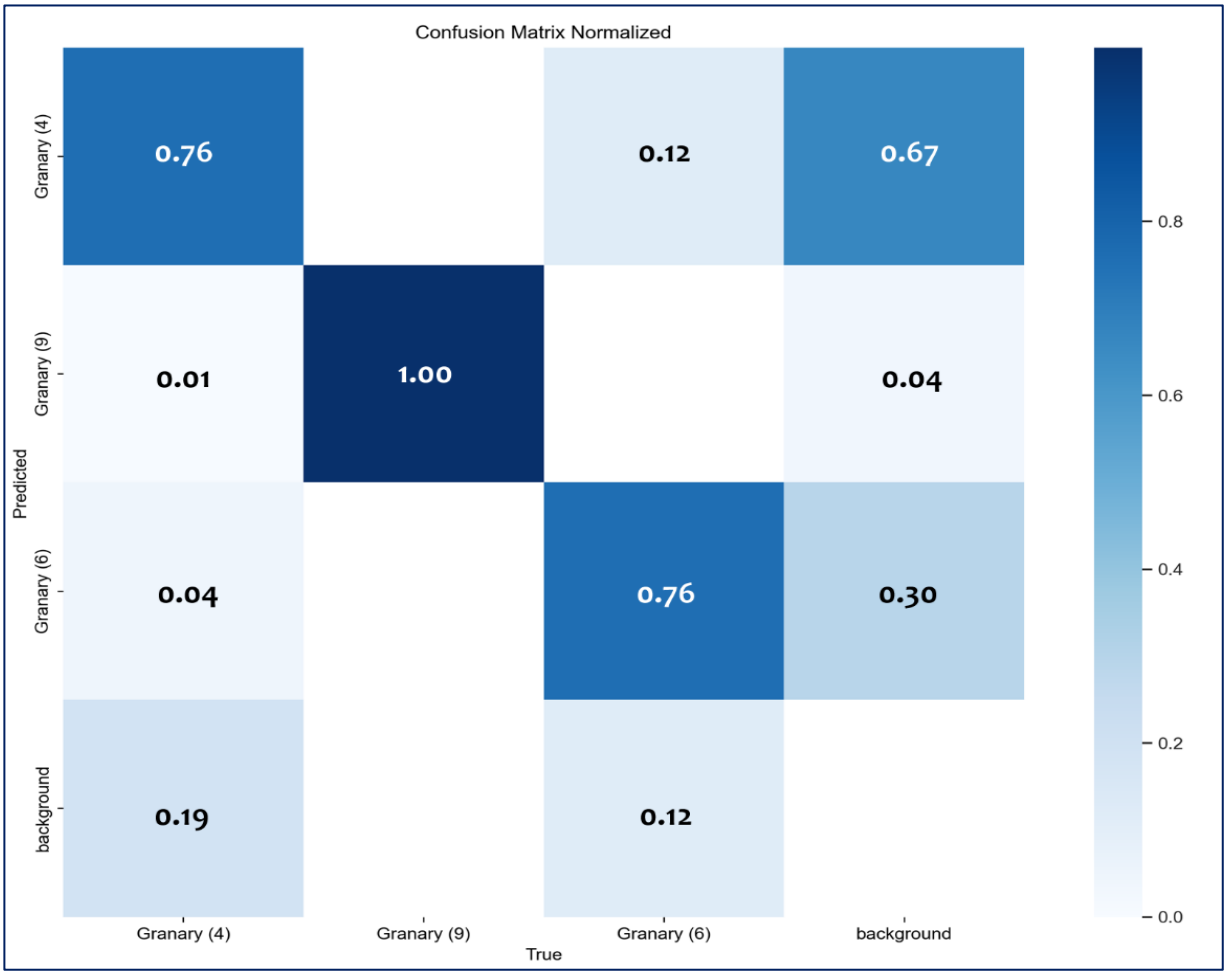


Figure 24: The row-wise normalised confusion matrix for model_1. It shows the overall proportion of each true class that was classified as each predicted class

to background noise (0.3). Generally, this suggests that the model struggles with background noise, leading to frequent misclassifications in this category and thus FPs.

All in all, model_1 shows a good overall performance. However, challenges with background noise led to frequent misclassifications, impacting the accuracy for other granary classes. The precision-recall analysis revealed that while the model maintains high precision at lower recall levels, it struggles to sustain accuracy as it tries to identify more positive instances. These findings highlight the model's effectiveness but also point to areas needing improvement, particularly in managing background noise and balancing precision with recall.

6.2 Evaluation model_2

Model_2 had several different metrics parameters as opposed to model_1, with a slight increase in the batch size, IoU threshold, confidence threshold, and weight decay. The model trained over 100 epochs in 0.534 hours. Furthermore, model_2 had an average processing times of 0.9ms for preprocessing, 6.9ms for inference, and 1.5ms for post-processing. The outcome of the evaluation metrics can be found in table 7 below:

	Images	Instances	Precision	Recall	mAP50
Granary (4)	214	157	0.835	0.656	0.787
Granary (6)	214	50	0.862	0.749	0.84
Granary (9)	214	6	0.644	1	0.942
All classes	214	213	0.78	0.802	0.856

Table 7: The overall performance metrics calculated during the testing of model_2. Here each labelled class is represented, and the overall summarised metrics for all classes combined.

All in all, model_2 exhibits distinct differences from model_1 in terms of performance metrics across various classes. While model_2 has a slightly lower precision and mAP^{50} for Granary (4) and Granary (6) compared to model_1, it shows an improved recall for these classes, indicating better overall sensitivity in detecting granaries. In contrast to model_1, which achieved perfect precision for Granary (9), model_2 has lower precision for this class but maintains a high recall, suggesting that it identifies nearly all instances of Granary (9) while being less accurate in its predictions. Aggregately, model_2 has a lower precision but higher recall than model_1, which reflects a shift in focus from

precision to a more balanced detection capability, evidenced by its comparable mAP50 score.

When examining the AUC-PR for model_2 (figure 25), the curve is quite similar compared to model_1. Notably, the performance for the Granary (6) class has improved. However, the Granary (9) category remains positioned on the outer axis of the graph, apart from one sudden drop the 0.8 recall threshold. Although hardly distinguishable, the Granary (4) class seems to have a slight decrease in performance. Additionally, the graph reveals a similar decline in precision around the 0.8 recall threshold. Overall, while the distribution of the PR-curves for model_2 is relatively similar to that of model_1, there is a slight increase in performance for the Granary (6) class, a similar unstable result for the Granary (9) class, and a slight decrease in performance for the Granary (4) class.

Once again, a raw and normalised confusion matrix for model_2 were plotted (figure 26 and 27). Overall, while model_2 demonstrates improvements in some areas, particularly for Granary (6), the persistent issues with Granary (9) and the increased misclassification of Background instances as Granary (4) or (6) highlight areas where further enhancements are needed. In other words, the incremental changes in model_2, while beneficial in some respects, have not yet fully addressed these persistent issues.

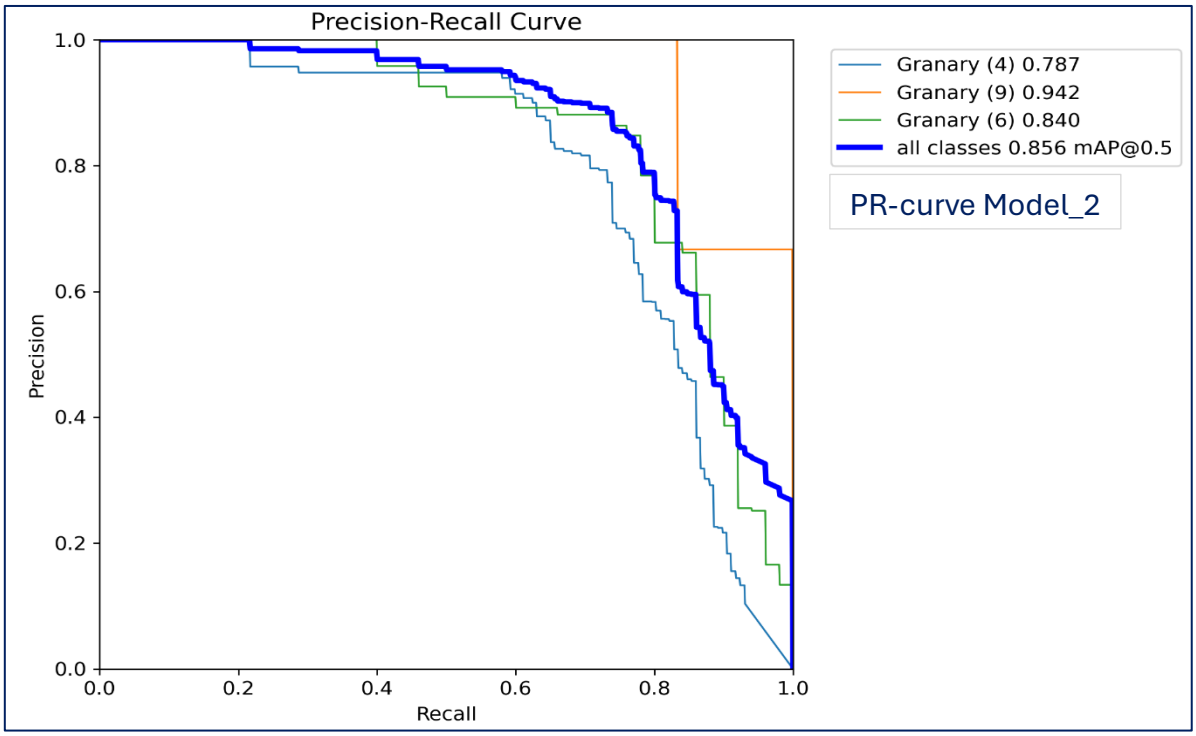


Figure 25: The precision-recall curve plotted for model_2. A slightly better performance of Granary (6) can be observed.

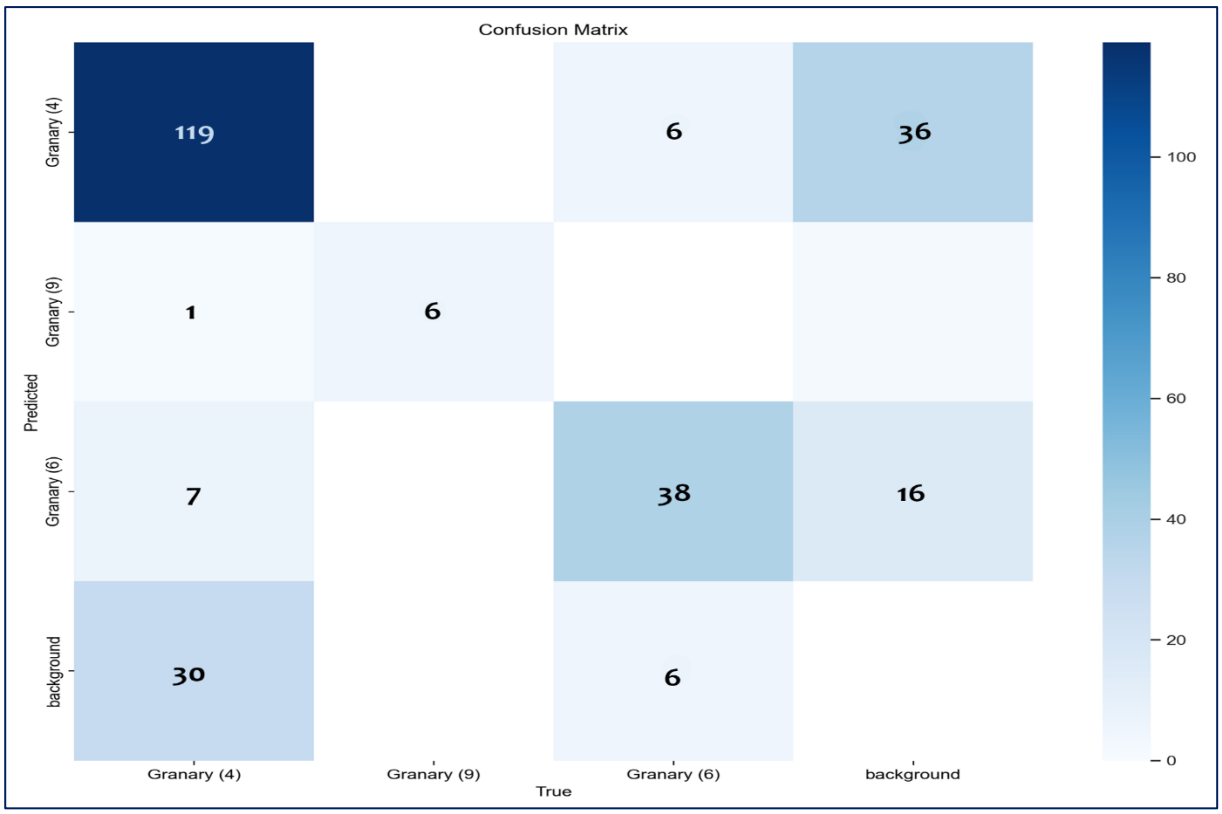


Figure 26: The raw confusion matrix plotted for model_2.

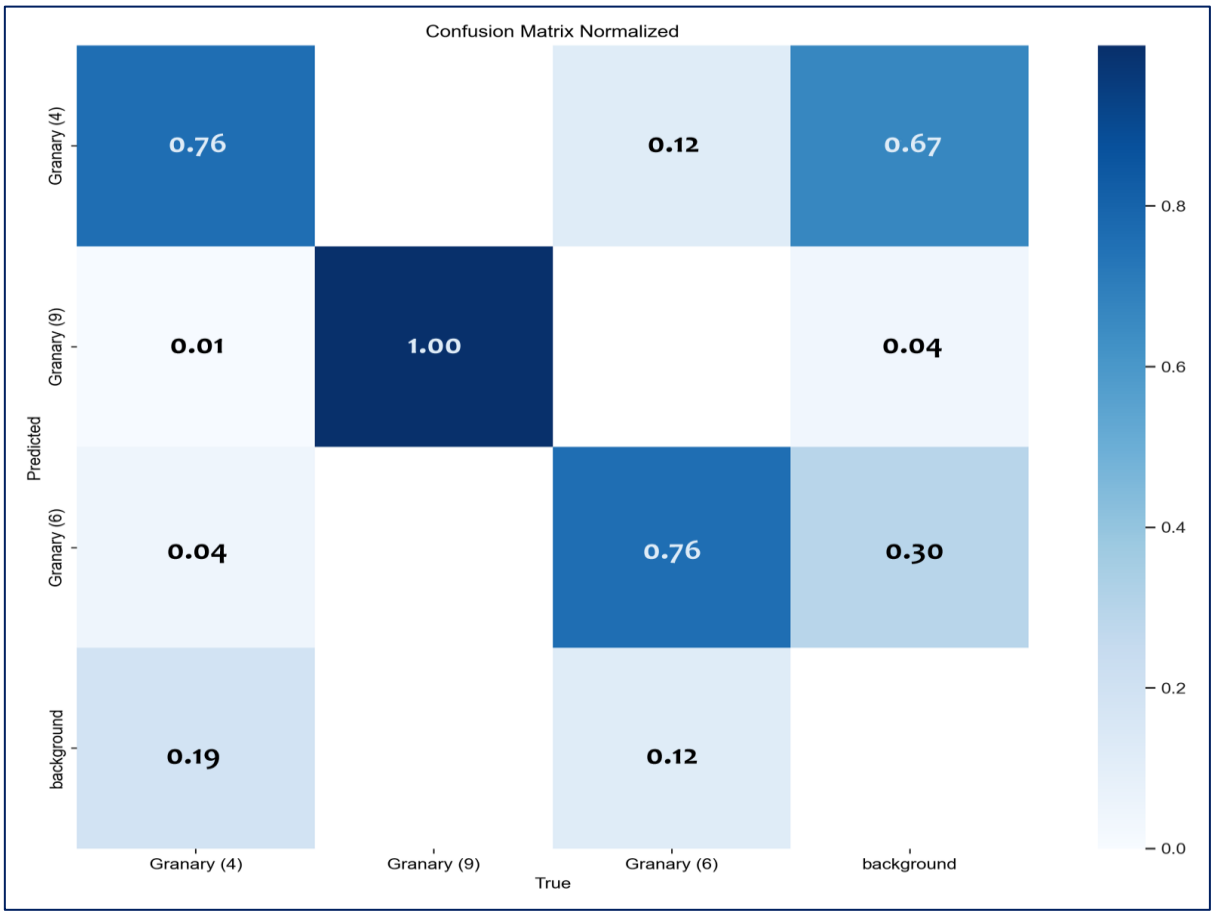


Figure 27: The row-wise normalised confusion matrix plotted for model_2.

6.3 Evaluation model_3

Model_3 had, once again, several different metrics parameters as opposed to model_1 and model_2, with a slight increase in the confidence threshold, amount of epochs, and weight decay. The model trained in a total of 150 epochs in 0.195 hours. Correspondingly, model_3 had an average processing times of 0.5ms for preprocessing, 6.9ms for inference, and 1.5ms for post-processing. The outcome of the subsequent evaluation metrics can be found in table 8:

	Images	Instances	Precision	Recall	mAP50
Granary (4)	214	157	0.891	0.643	0.808
Granary (6)	214	50	0.897	0.698	0.825
Granary (9)	214	6	0.723	1	0.955
All classes	214	213	0.837	0.781	0.863

Table 8: The overall performance metrics calculated during the testing of model_3. Here each labelled class is represented, and the overall summarised metrics for all classes combined.

All in all, model_3 demonstrates improvements compared to model_2 and shows varied performance in relation to model_1. Compared to model_2, model_3 exhibits higher precision for Granary (4) and Granary (6), reflecting more accurate predictions for these classes. Additionally, model_3 achieves a higher recall for Granary (6) than model_2, indicating better detection capabilities for this category. However, while model_3 shows improved performance over model_2 in several metrics, its precision for Granary (9) is lower than that of model_1 but is balanced by a high recall, suggesting that it is more sensitive in detecting Granary (9) but with slightly lower precision.

Overall, model_3 shows the best results in recall and precision. Especially when ignoring the Granary (9) class, as these evaluation metrics seem to be highly influenced by limitations in the dataset. Model_3 achieves the highest precision and recall for both Granary (4) and Granary (6), along with the best mAP^{50} scores for these classes. This suggests that Model_3 provides a strong balance of accuracy and sensitivity for the remaining granary categories.

The AUC-PR showcases a similar shape as the curve for model_2 (figure 28). However, the curves for all classes have an overall higher bend which indicates an overall better performance than the other models. Therefore, based upon the curve as well as the overall metrics, it can cautiously be said that model_3 has the best performance compared to the previously trained models. Similar to the other examples, the Granary (9) class is on the "perfect" 1.0 to 1.0 ratio indicating a flawless performance on all the precision and recall thresholds. Once more, although this ratio is the ideal metric, this outcome is the result of clear overfitting of the Granary (9) class within the entire dataset. The drop in precision at higher recall thresholds can also be observed in this particular graph. Lastly, and perhaps the most notable difference can be seen when looking at the difference between the Granary (4) and Granary (6) class. Whereas within the other models the Granary (4) class or Granary (6) class outperform one another, here it seems that they are relatively similar. The Granary (6) class has a higher curve and seems to be performing better across the varying precision and recall thresholds as opposed model_1, and similar things can be said for Granary (4) in model_2. This can be attributed to several factors. The last raw and normalised confusion matrices were plotted (figure 29 and 30).

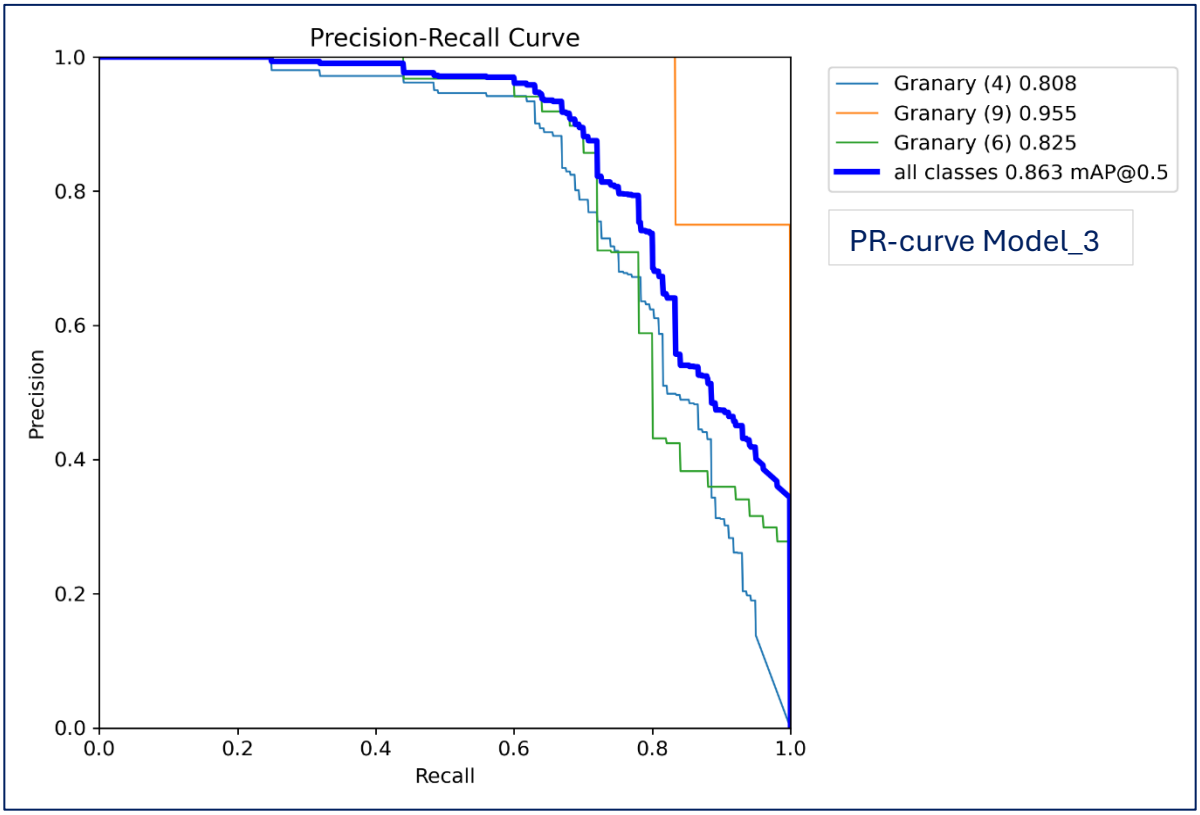


Figure 28: The precision-recall curve plotted for model_3.

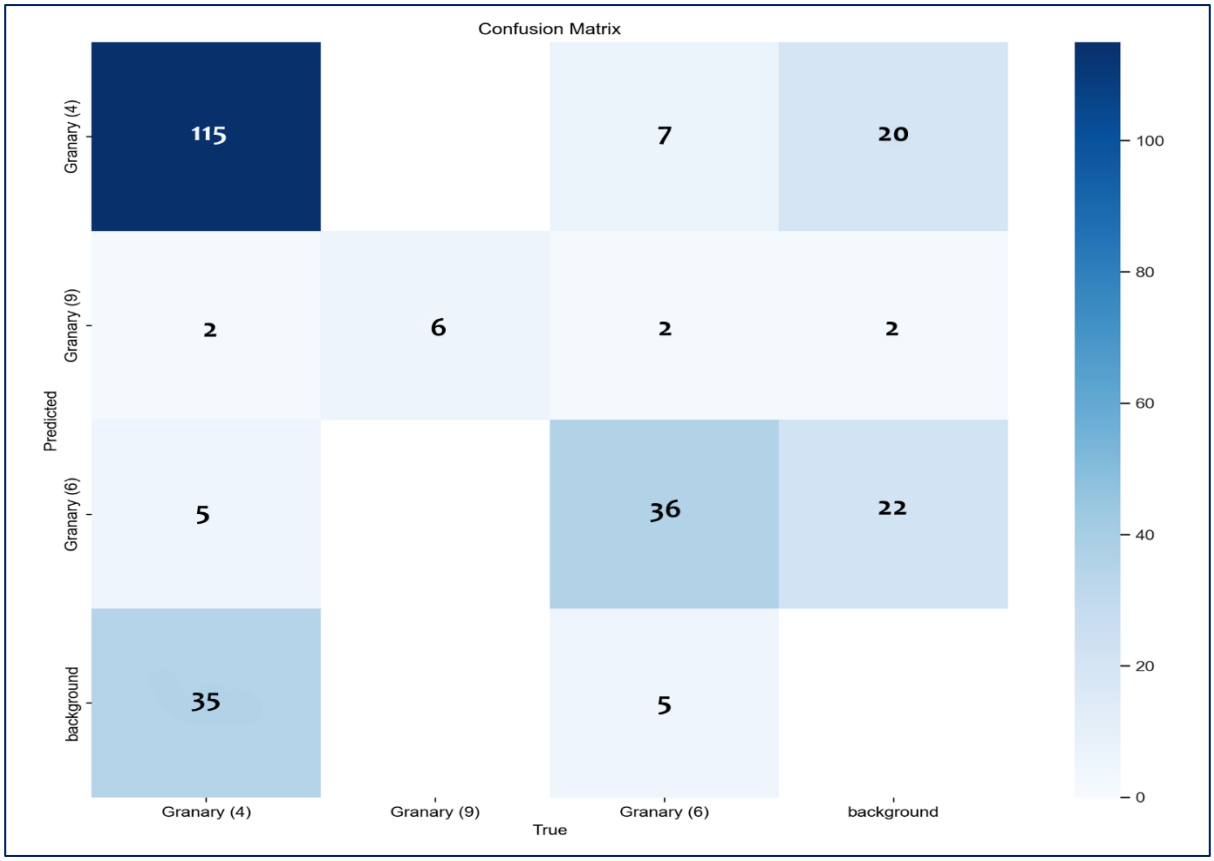


Figure 29: The raw confusion matrix plotted for model_3.

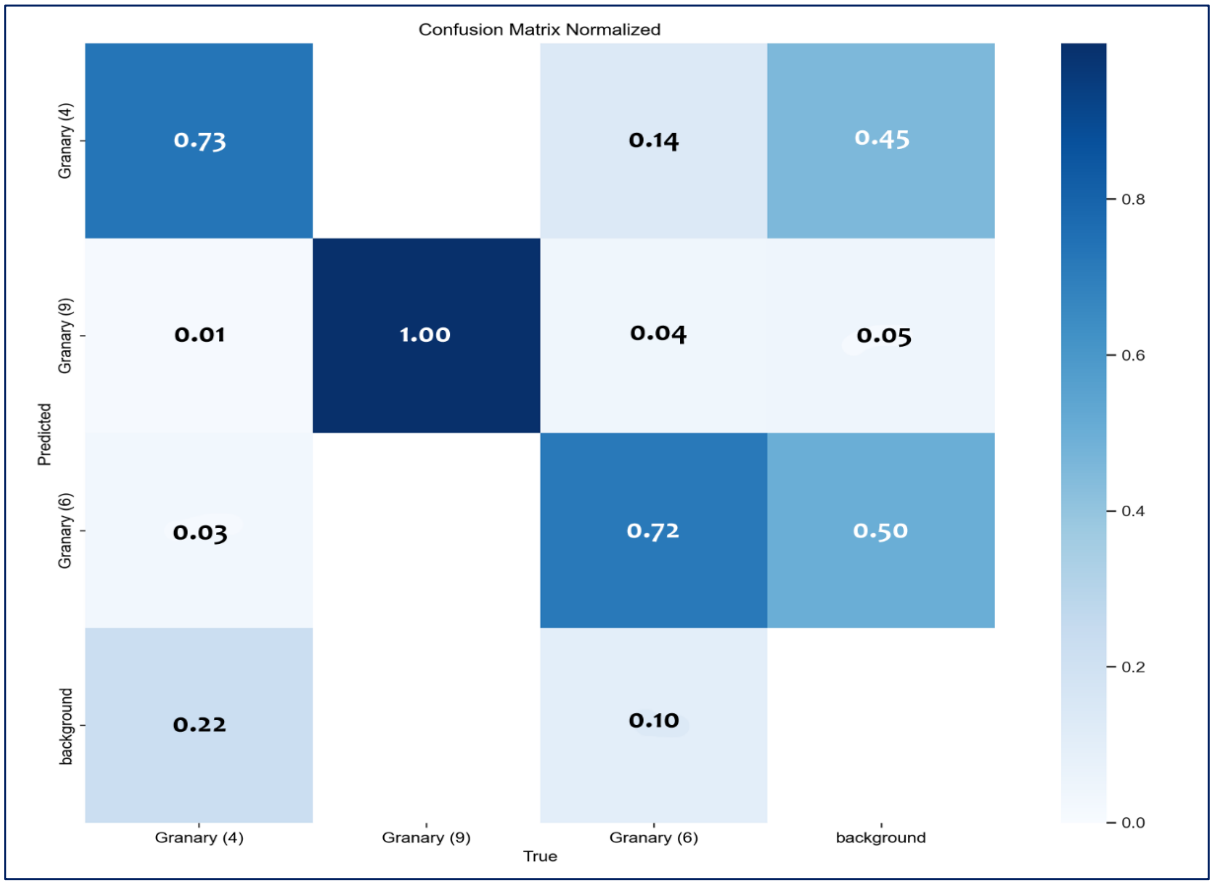


Figure 30: The row-wise normalised confusion matrix plotted for model_3.

6.4 Comparison of models

To facilitate a side-by-side comparison of the models, bar plots were generated, allowing for a clear visualisation of their overall performances.

6.4.1 Precision

All the models can be deemed reliable when evaluating precision, as each shows strong performance in different contexts and with varying parameter settings. The overall precision scores for all models are notably high, reflecting the effectiveness and robustness of these DL models in distinguishing relevant instances from irrelevant ones. This level of precision indicates that, despite their individual variations, all models exhibit a commendable ability to accurately predict positive instances, demonstrating their potential for practical applications in diverse scenarios.

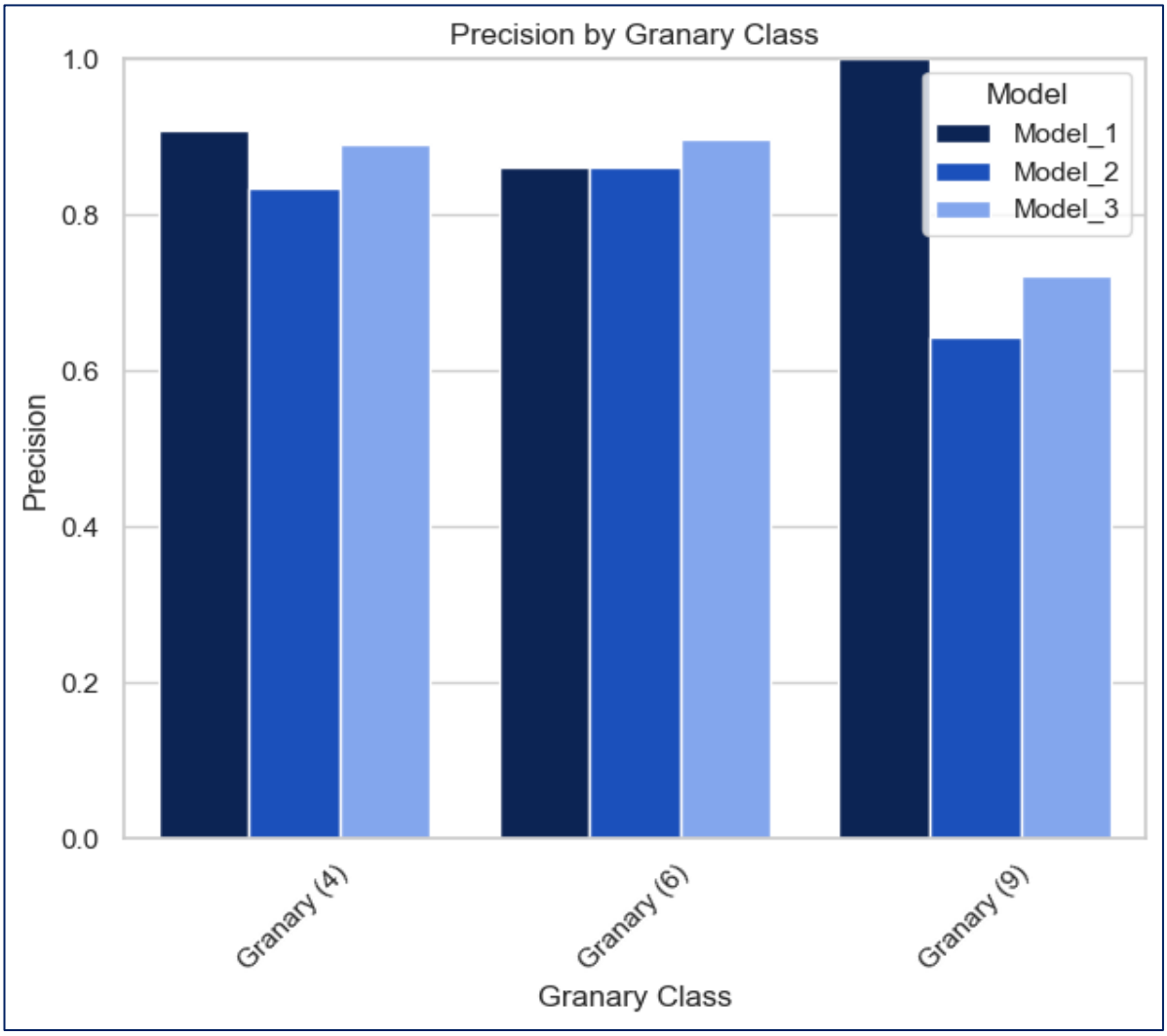


Figure 31: Bar graph of precision score for each model and their respective classes

6.4.2 Recall

The recall values across all models are moderate. The values observed across the board indicate that all models struggle to detect a substantial portion of relevant instances. This challenge is particularly evident when examining the confusion matrices for each model, where it becomes clear that background noise, and hence the difficulty in distinguishing between instances and the background are factors contributing to the recall scores.

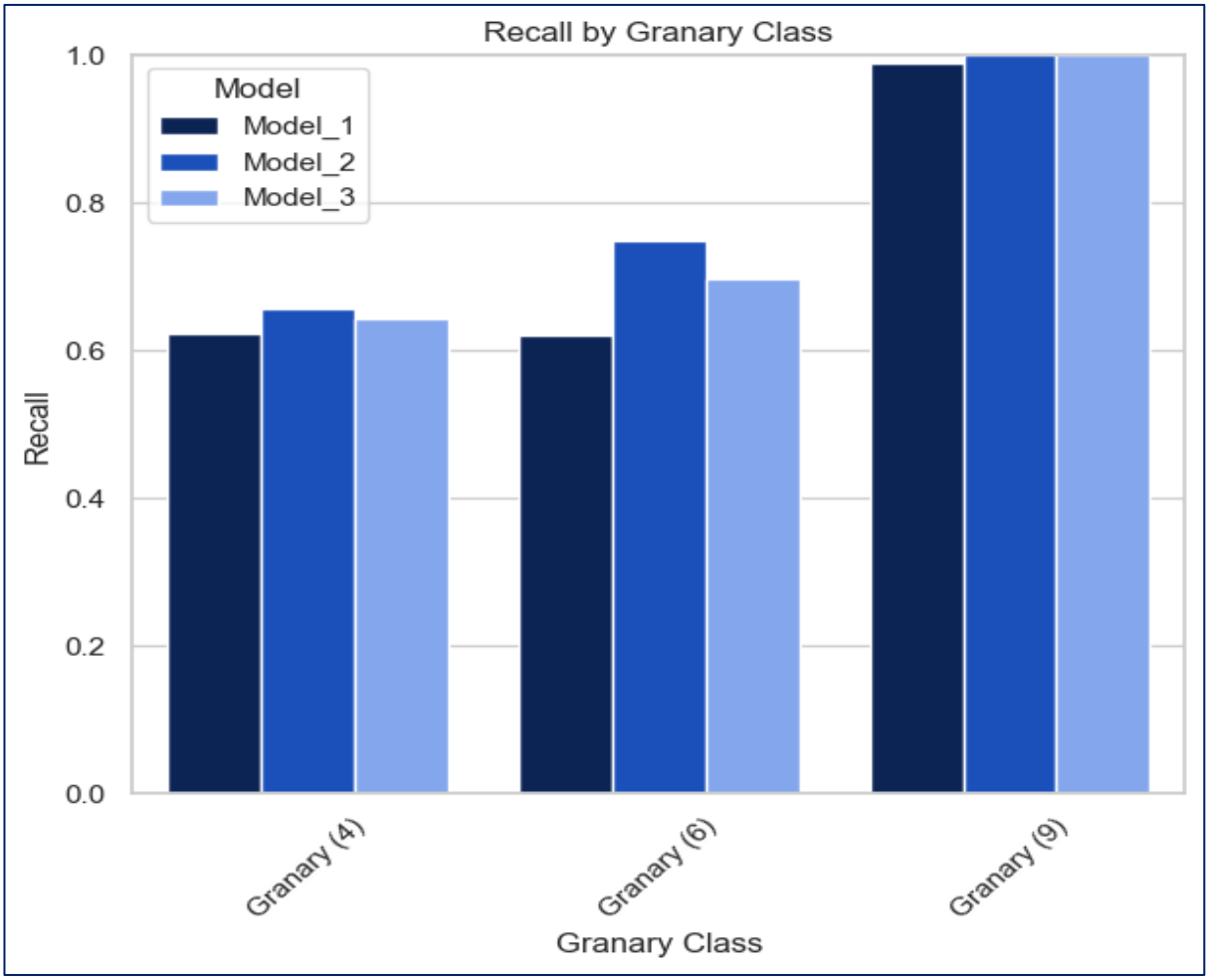


Figure 32: Bar graph of recall score for each model and their respective classes

6.4.3 mAP^{50}

While the mAP^{50} values across all models reveal varying degrees of performance in balancing precision and recall, they collectively demonstrate that the models are quite effective. Model_3 stands out with the highest mAP^{50} scores, reflecting its ability to maintain a good balance between precision and recall across different datasets.

Model_1 and model_2 also show solid performance, with moderate to high mAP^{50} values indicating that they handle precision and recall effectively. Despite the challenges posed by the small sample size of Granary (9) and the moderate recall values, the overall results suggest that the models are performing well and have the potential for further refinement to enhance their accuracy and reliability.

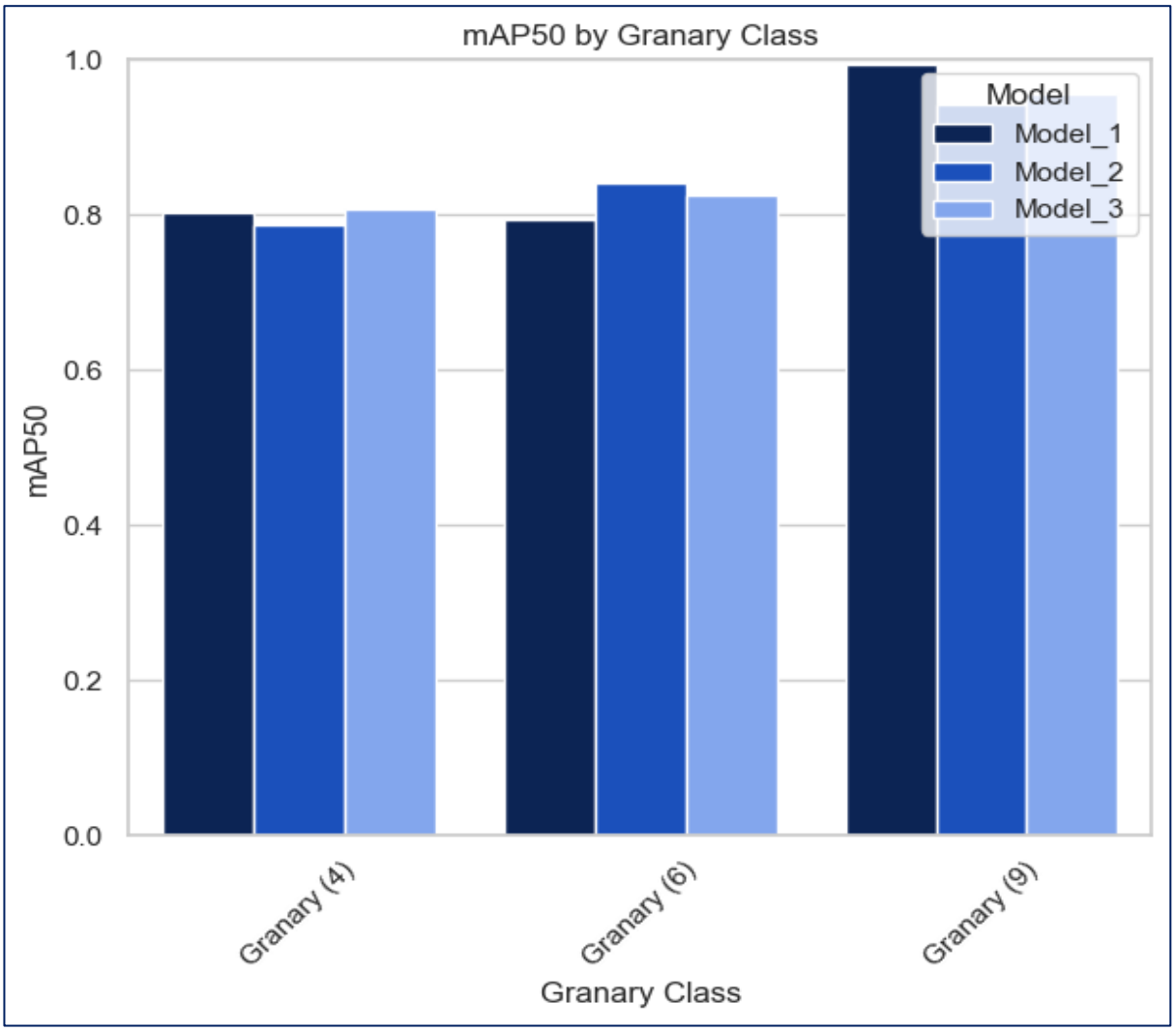


Figure 33: Bar graph of mAP^{50} score for each model and their respective classes

7. Discussion

The chapter will delve into the key results, evaluating the performance of the YOLOv8 algorithm and identifying the factors that influenced its recognition capabilities. This will be followed by a discussion on the potential implications of these findings for archaeological interpretation, considering how the algorithm's performance may impact the analysis and understanding of archaeological data. Next, the general methodology of this study will be scrutinised, highlighting both the strengths and limitations encountered during the research process.

7.1 Interpretation of results

As reflected in the chapter 6, all the models trained within this thesis perform quite well in identifying prehistoric granary structures on archaeological excavation maps (figure 34). With an average mAP^{50} of 0.861 for every model across all categories, it can be said that the models demonstrate a moderately high level of accuracy in their overall predictions. However, it is important to note that these evaluation metrics are not

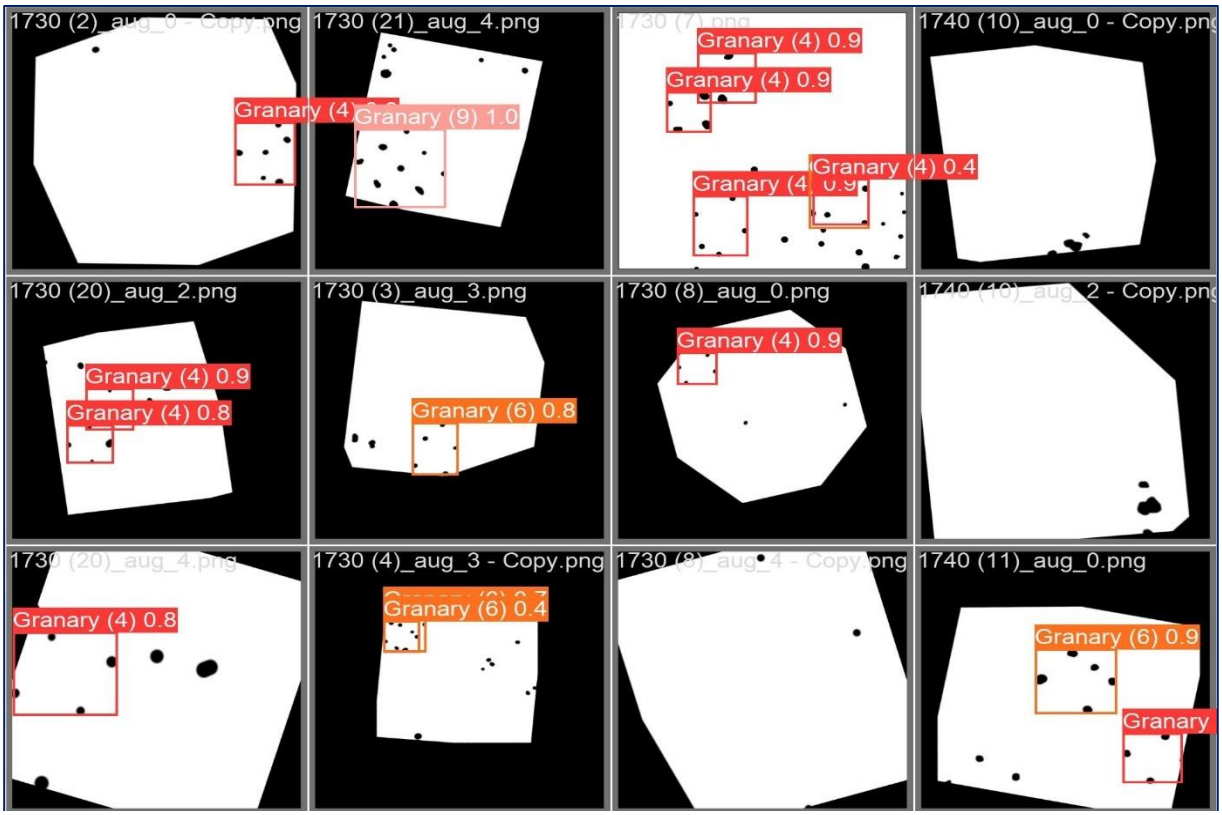


Figure 34: Some examples of identifications by the model_3 on the test data set. For more examples consult appendix 2.

exhaustive indicators of the models' true effectiveness. Several critical factors must be considered to fully understand, interpret, and apply these metrics.

7.1.1 Nine-post granary

The category of “Granary (9)” frequently achieves a perfect precision and recall in its performance metrics, indicating that the models are able to identify all the nine-post granaries relatively flawlessly on the unseen testing data. While these results may seem ideal at first glance, they raise questions about the models’ overall performances. Several factors could be the reason behind this outcome. First of all, the overall uniformity and distinctiveness of the Granary (9) category might contribute to these metrics. In general, the nine-post granaries have relatively consistent features that are easily distinguishable from other categories. All the instances used within the dataset include all the posts, have a low amount of background noise, and are not intersected by other granaries (figure 35). Which means that all the images that are used are “perfect” examples of nine-post granaries, which is not representative of a realistic archaeological dataset. This could mean that the models are not necessarily generalising well, but are instead learning to identify a category that is less complex or more homogeneous compared to the others.

Secondly, there is most probably an issue of overfitting where the models have memorised the training data rather than learning to generalise from it. This overfitting could be particularly prevalent as the granary category is represented disproportionately in the training and test set, with only 36 instances which comprise approximately 4% of the entire dataset. This small size of instances immediately raises concerns about the reliability of these metrics, as the test data may not adequately represent the broader variability of granaries, which could skew the performance evaluation.

Therefore, given the current results, it is challenging to definitively assess whether the model is overfitting or if it is genuinely performing well. It is clear that the evaluation metrics are unrealistic and unreliable to provide a complete picture of the models’ capabilities. To gain a clearer understanding of the models’ true generalisation capabilities, it is crucial to expand the dataset by incorporating more diverse examples. Increasing the number of instances will provide a better evaluation and help determine whether the models’ performance are effective or if any further adjustments are needed.

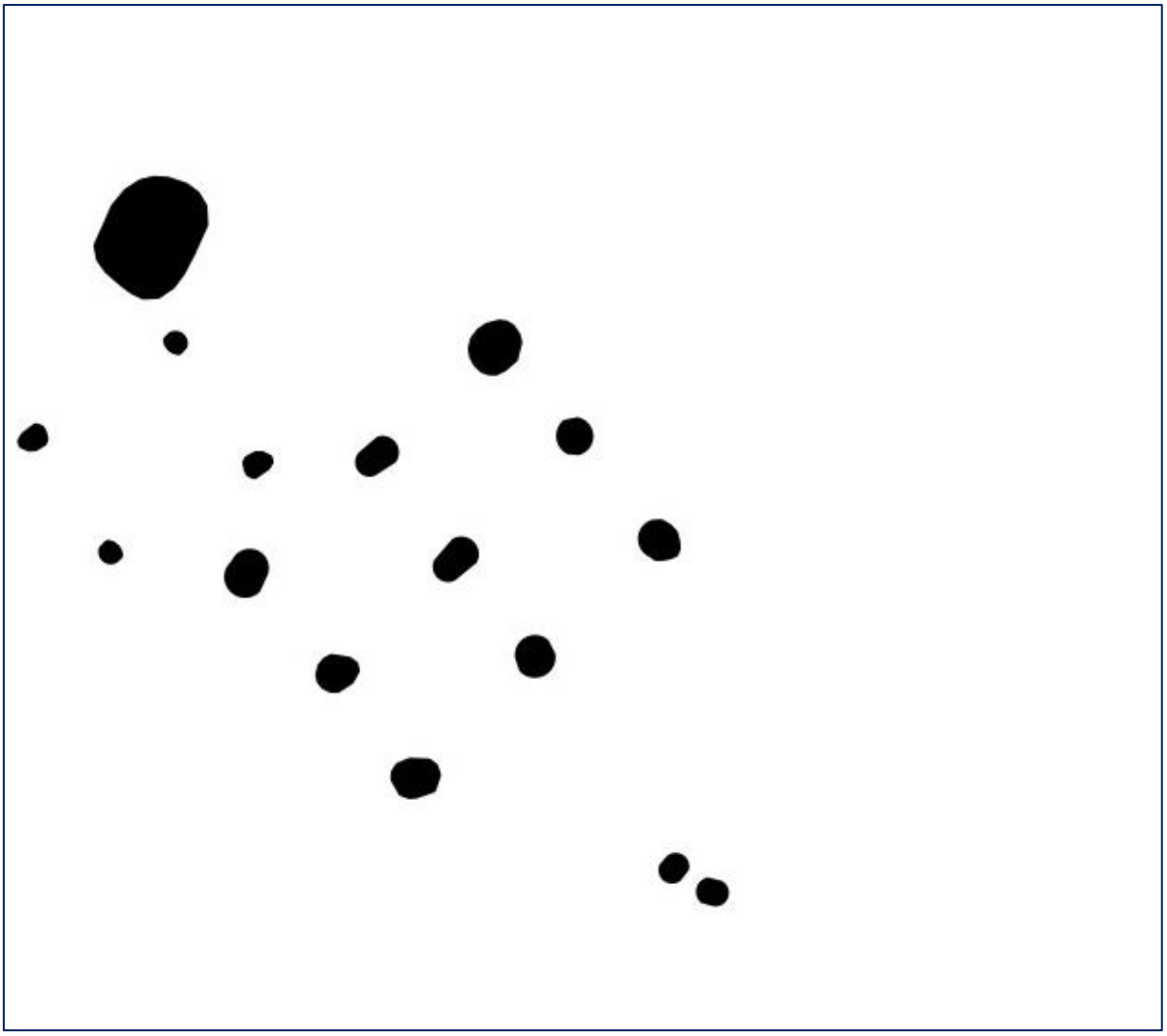


Figure 35: Example of a nine-post granary instance in the dataset. As can be seen, the image is relatively clean, and the granary is easily distinguishable.

7.1.2 4-post and 6-post granaries

When taking a broader view of the performance metrics by also looking at the Granary (4) and Granary (6) categories, the models can be considered reliable. Where the average metrics without the Granary (9) category are the following: precision of 0.876, recall of 0.665, and mAP^{50} of 0.810. The precision metric, in particular, is noteworthy. A precision of 0.876 indicates that a significant majority of the predicted granaries are true positives, highlighting the models' effectiveness in minimising false positives. The recall of 0.665 demonstrates the models' ability to identify a substantial portion of the actual granary instances. While this is somewhat lower than the precision, it still shows that the models are reasonably good at detecting most of the relevant granaries. The mAP^{50} of 0.810 provides a balanced view of both precision and recall, indicating overall strong

performance in both identifying and localising granaries. This metric suggests that the models achieve a good compromise between correctly identifying granaries and minimising missed detections.

All in all, the model can be summarised as having a high precision, but a lower recall. This means that the models are effective at correctly identifying granaries among the instances it detects, with a high proportion of true positives. However, the lower recall indicates that the model may miss up to 35% of the actual granaries. Although the original intention was to prioritise recall to ensure that as many granaries as possible are identified, the outcome has inadvertently favoured higher precision. Even with ensuring a low confidence and IoU threshold, the models' recall remains lower than desired. The possible reasons for this will be addressed in the subchapter 7.1.2, however the lower recall has implications for the models' usefulness in practical archaeological research. While high precision is beneficial because it means that the instances detected by the model are mostly true positives, thus reducing need for excessive validation, a lower recall can be less advantageous.

In short, while high precision reduces the burden of false positives, in archaeological contexts where discovering and reviewing as many potential examples as possible is important, a model with lower precision but higher recall could be more beneficial. This approach would ensure that fewer granaries are missed, even if it means dealing with a higher volume of potential false positives for expert validation.

7.1.3 False positives

As discussed in chapter 6, model_3, the best-performing model, identified a total of 40 instances of granaries within the background class. According to the confusion matrix (figure 23), the model recognised 35 four-post granaries and 5 six-post granaries among previously unlabelled instances in the dataset. This observation raises two possible explanations: either the model made incorrect predictions, or it successfully uncovered previously undetected granaries within the dataset.

To evaluate these possibilities, a closer analysis of these specific instances was conducted. First of all, some of the instances are clear misclassifications, caused by background noise. A total of 14 FP classifications were made on prehistoric house plans

present within the dataset (figure 36). Three house plans were partially visible within the images, which led to confusion in the models' classification process. The house plans, which exhibit a rectangular patterns, created strong visual similarities with the granaries. This similarity caused the models to struggle in distinguishing between granaries and house plans, resulting in these erroneous identifications. Although this issue clearly skews the results of the models' performance, it is also an interesting reference for future research, as it demonstrates the models' potential to identify a wider range of categories. In future, expanding the classification categories to include more categories, such as house plans, could be an interesting avenue of research. Furthermore, by adding additional categories, the model could learn to distinguish between granaries and other structures more effectively, leading to better classification.

Approximately 20 of the FP's made by the model seem to be simple mistakes. When looking at the output, there are no clear indications that these could be actual granaries (figure 37). This is mostly based upon the fact that there is no structural resemblance at first glance, and when looking at the metadata of these particular post-holes they seem to be too dissimilar to be flagged as granaries. Either the post-hole depth, colour, or approximate dating of the archaeological features does not sufficiently correspond.

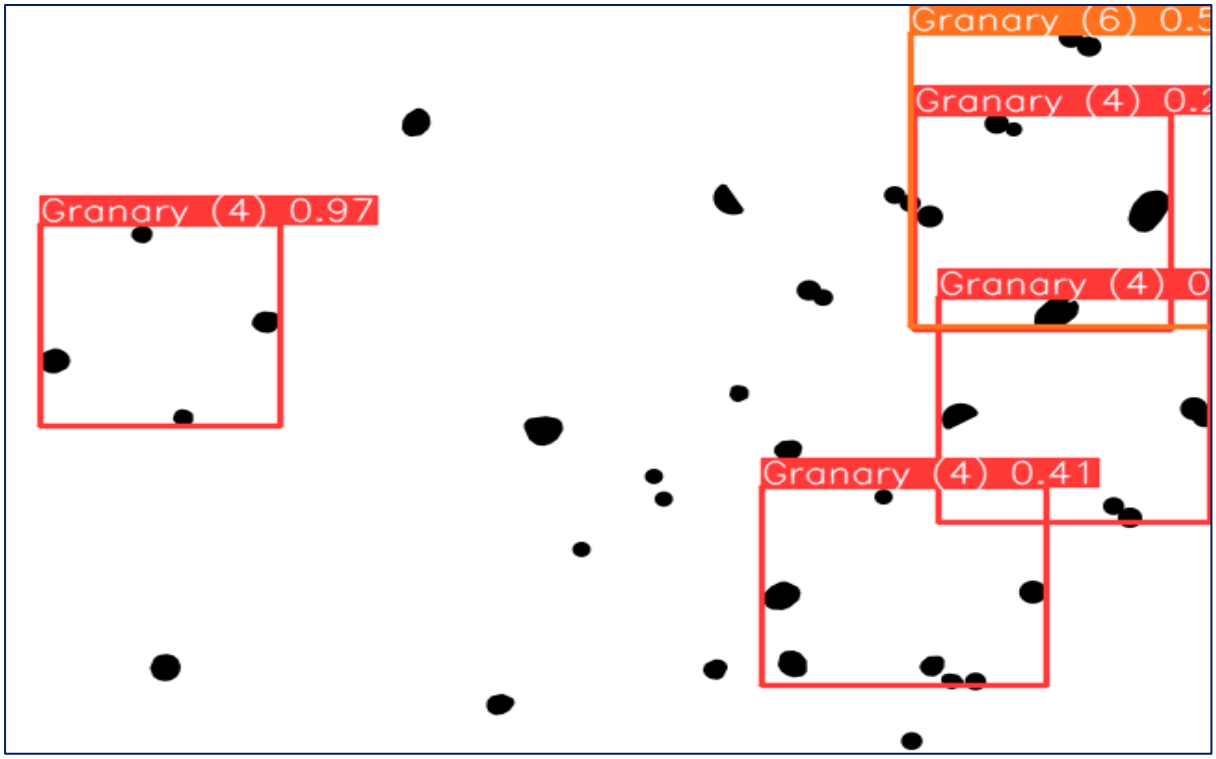


Figure 36: Misclassification of a house plan as multiple granaries. This example illustrates how similar visual features between house plans and granaries can lead to false positives.

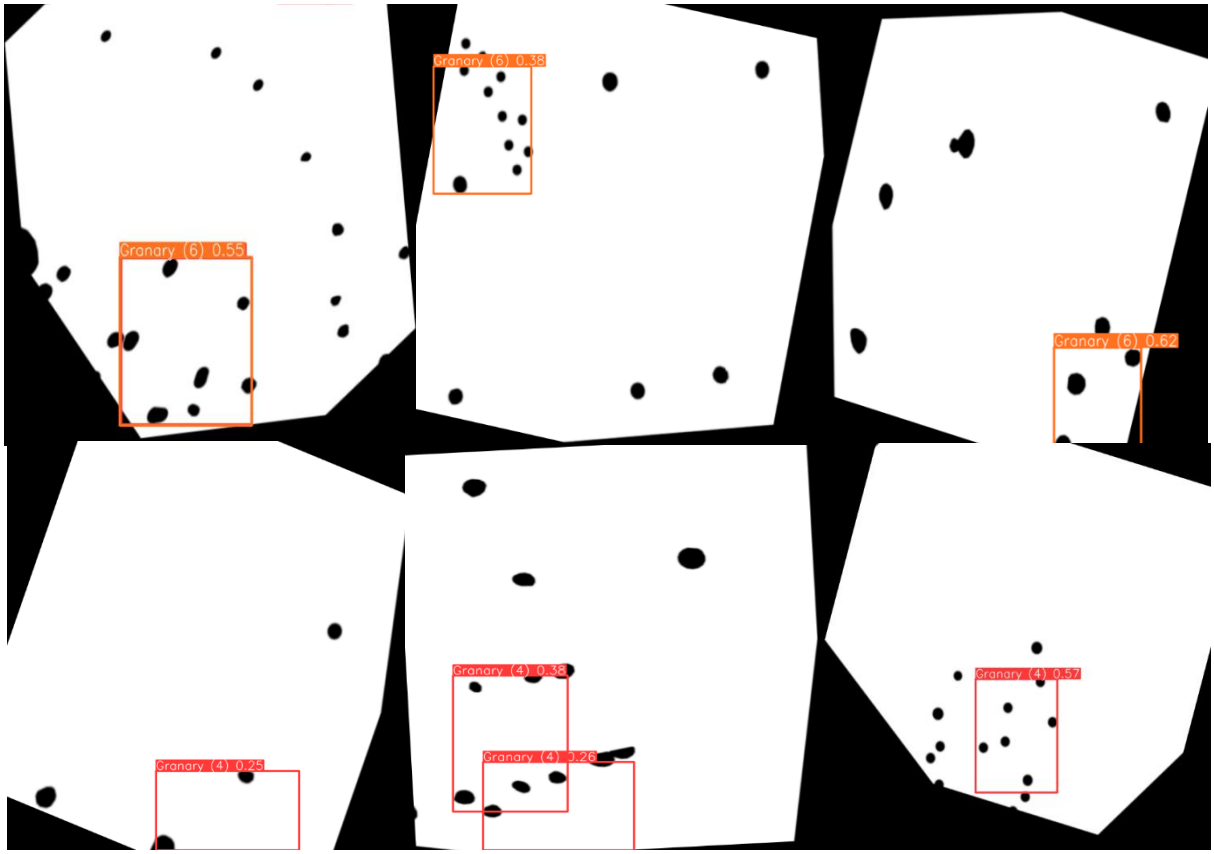


Figure 37: False positives that seem to be clear misclassifications, based upon the structural dissimilarity and the metadata of the corresponding post-holes.

The remaining 6 FP's, however, are more interesting to consider (figure 38). Although it is difficult to definitively identify something as a granary as outlined by the "human black-box" issue outline in chapter 4. These FP's might actually give examples of newly identified granaries.

Where in image (A) the initial labelling was empty as this was a negative example, it is clear that a six post granary could be identified here. Although the bounding box predicted by the model is slightly off-centre, the outermost vertically aligned post-holes exhibit consistent characteristics in feature depth, colour, and approximate dating.

Image (B) reveals another interesting example of a potential new four-post granary. While the original dataset annotation acknowledged the presence of the six-post granary, the model's prediction suggests an additional, intersecting four-post granary. Closer examination of the excavation data supports this hypothesis, revealing that the northern two post-holes of the earlier four-post granary were later repurposed for the six-post granary constructed in a subsequent phase. This overlap indicates a chronological

sequence of use, with the four-post granary representing an earlier phase of activity at the site. Despite this, the dataset annotations only recorded the six-post granary, overlooking the earlier structure.

Image (C) reveals two potential four-post granaries that were not included in the original dataset annotations. The uppermost structure appears consistent with a typical granary, but could also be considered a five-post variant due to the presence of an additional post-hole near the centre. The post-holes in this structure exhibit similar dimensions, depth, approximate dating, and spatial distribution, lending strong support to its identification. In contrast, the lower structure, while sharing comparable post-hole characteristics, presents a less conventional alignment, with the posts spaced approximately three meters apart—wider than the standard configuration for four-post granaries. This atypical spacing introduces some ambiguity, making the classification less certain. While it is plausible that the lower structure represents a variation of a granary, further analysis is required to confirm this interpretation.

Image (D) presents two additional potential granaries, though both exhibit features that make them somewhat atypical. The left granary shows a slight misalignment in its post-hole arrangement, which deviates from the more precise configurations typically associated with granary structures. Meanwhile, the rightmost structure is characterised by significant variations in post-hole depth. While these inconsistencies might suggest a different function or partial degradation over time, the overall spatial arrangement and other characteristics still align with known granary features.

All in all, the findings highlight both the strengths and limitations of the models in. While a significant portion of the false positives can be attributed to background noise and misclassifications the model's ability to recognise potential structures is noteworthy.

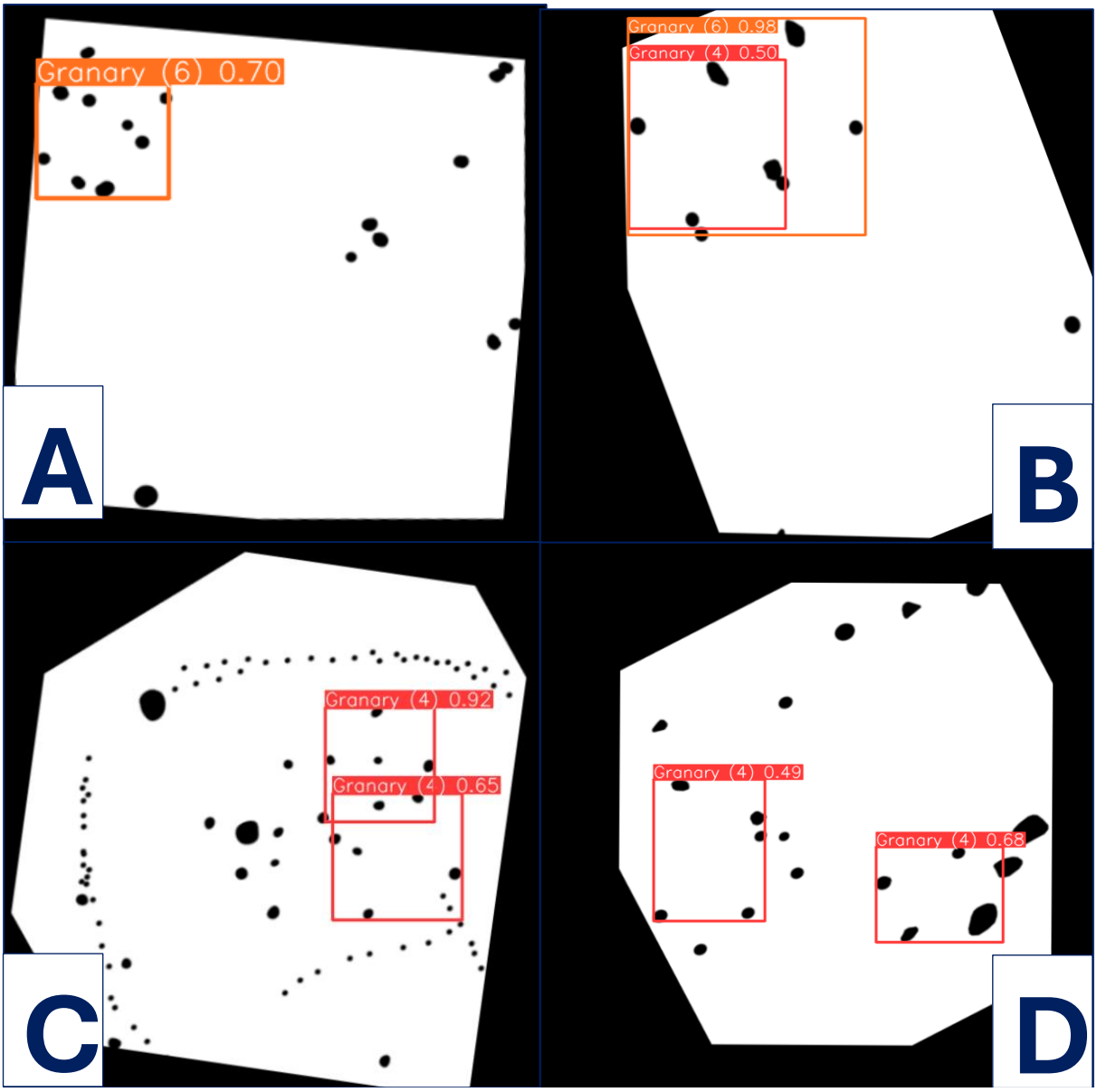


Figure 38: False positives identified by model_3 that could be previously unseen granaries.

7.2 Factors influencing recognition performance

To gain an understanding of the model and its potential applications in practical archaeological contexts, it is crucial to examine the other factors that may impact recognition performance.

7.2.1 Background

Overall, the models seem to be working well on the Granary (4) and Granary (6) categories. Although the direct recognition criteria are not explicitly outlined, often referred to as the "black-box issue," the model demonstrates effective performance in

distinguishing these categories. Logically speaking, the models are recognising these categories on the basis of the amount of post markers and the distance between them. However, there are also instances within the dataset, particularly with the six-post granaries, where one of the posts is obscured or missing. Even though that is the case, the model seems to recognise these instances as well, suggesting that it can accommodate variations in post visibility and still correctly classify the granaries. This indicates that the models are somewhat flexible in handling incomplete or partially obscured data. While this is a positive aspect of the models' performance, it may also contribute to a challenge: as indicated in the former paragraph, there is a common issue where the background class or negative examples are frequently misclassified as granary instances.

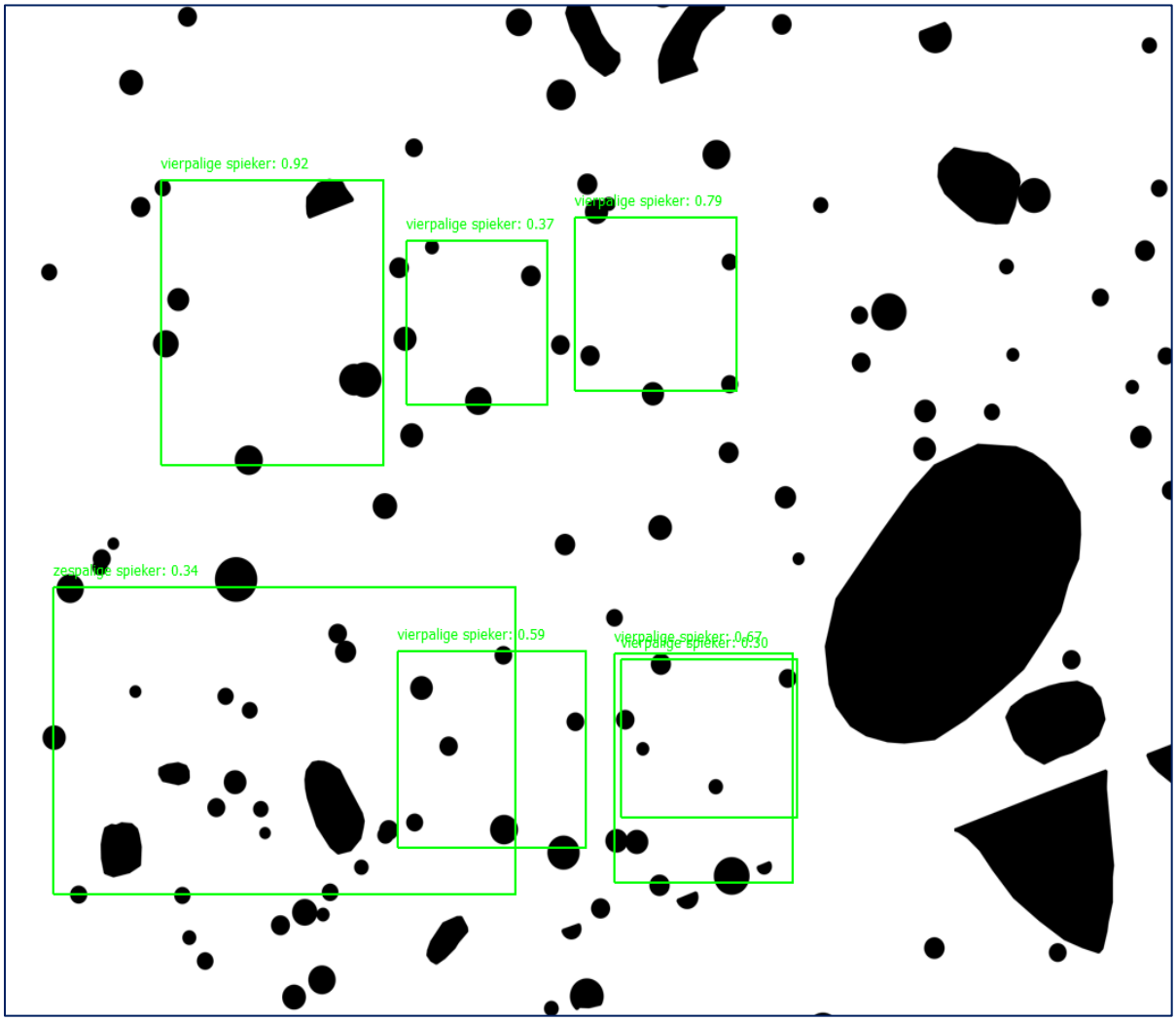


Figure 39: Model_3, when applied to the excavation of Schipluiden, which is known for its high levels of post-hole noise, demonstrates limitations in handling such extensive background noise. As observed, the model struggles to manage the overwhelming amount of noise, resulting in diminished performance.

This issue is not particularly surprising, as the background in the excavation maps showcase similar properties as the instances that the models are supposed to find black circular features on a white background. This similarity in appearance between the background and the target granaries likely contributes to the models' difficulty in distinguishing between them. The presence of visually similar elements presents a challenge for the models, complicating the task of distinguishing granaries from the background. Additionally, the significant amount of background noise in some images exacerbates this issue. The difficulty becomes even more pronounced when the model is applied to excavation maps with substantial background noise, as illustrated in figure 39.

In other words, while the models perform well on relatively clean imagery, their performance declines significantly when confronted with increased background noise. Clearly, this is not ideal for the models' intended purpose of helping archaeologists disentangle complex datasets. The models' inability to effectively handle noisy or cluttered data undermines its utility in distinguishing and categorising archaeological structures accurately. To be truly valuable, the model needs to be capable of managing and interpreting challenging datasets, which is crucial for providing meaningful insights and supporting archaeological research. Therefore, this represents a significant shortcoming in the models' capabilities and is a critical factor that must be addressed. The main possible reason for this shortcoming is discussed in chapter 7.3.

7.2.1 Overlapping granaries

The output of the models also indicates that they struggle with disentangling and accurately classifying overlapping or incomplete features, as well as differentiating between categories that appear visually similar. There are several misclassifications that seem to occur throughout all the models. While the primary distinction between a four-post, six-post, and nine-post granary lies in the number of posts, the visual similarities between these structures often lead to confusion. The models often fail to recognise that a feature classified as a six-post granary cannot simultaneously be a four-post granary. In other words, components of larger categories, such as six-post and nine-post granaries, are frequently misclassified as smaller individual categories, leading to significant confusion and inaccuracies in classification (figure 40). Although the confidence levels

are often lower in these specific cases, the models still frequently misclassify features, indicating a persistent problem in accurately distinguishing between visually similar or overlapping categories. While this issue presents challenges, it is not entirely detrimental to the overall functionality of the model. Despite the inaccuracies, the model still successfully identifies and classifies granaries, which is the primary objective. However, the problem primarily affects performance metrics, leading to potential discrepancies in accuracy, precision, and recall. Furthermore, the output can become skewed as the model may identify more granaries than actually present in a single instance, which complicates the interpretation of results. This skewed output makes it more difficult for archaeologists to accurately interpret the data, as the models' overidentification of granaries can obscure the true distribution and characteristics of the features. Thus, while the model meets its core objective, the challenges in performance metrics and data interpretation highlight the need for further refinement to ensure clearer insights for archaeological analysis.

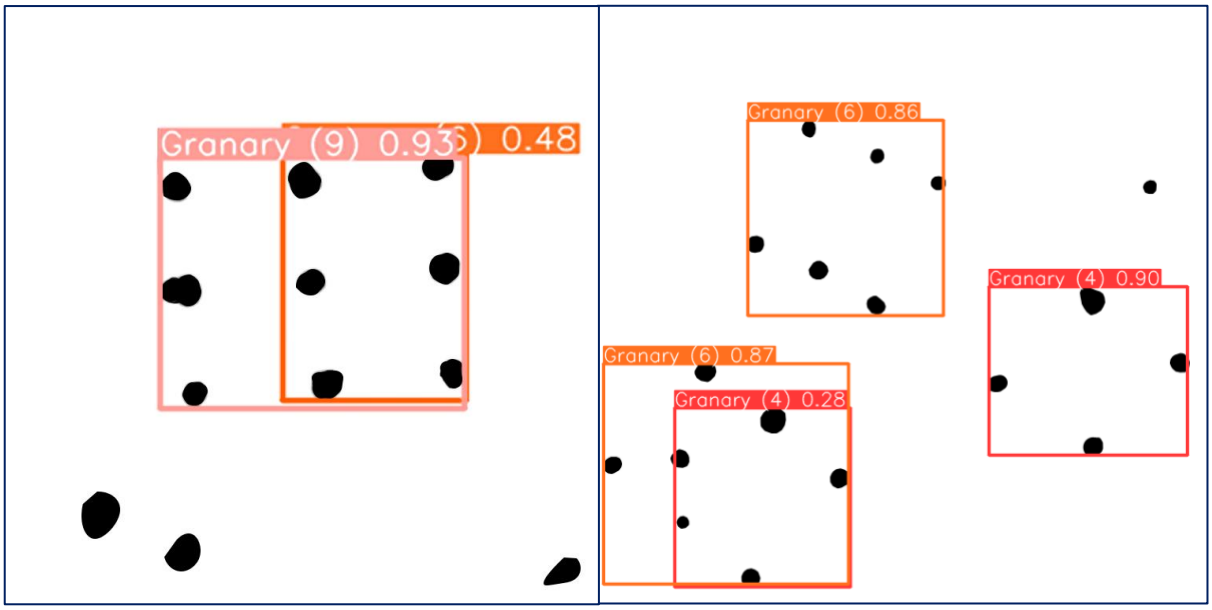


Figure 40: **Left:** Model_2 erroneously classifies the nine-post granary as a six-post granary as well. **Right:** Misclassification made by model_3 where the bottom-left six-post granary is also identified as a four-post granary.

A related issue is the problem of missing post-holes within the Granary (6) category. The confusion matrices reveal that several six-post granaries were incorrectly classified as four-post granaries. An examination of the model outputs indicates that this misclassification primarily stems from the absence of posts in some six-post granaries. During the development of the labelling strategy, granaries were labelled as six-post even

if the middle post along the longitudinal axis was missing, based on the assumption that the main shape of the granary remained intact. However, this approach has possibly led to several six-post granaries being erroneously classified as four-post granaries (figure 41). This issue is likely due to the low number of examples that include this particular abnormality, resulting in the models being unable to learn from such cases.

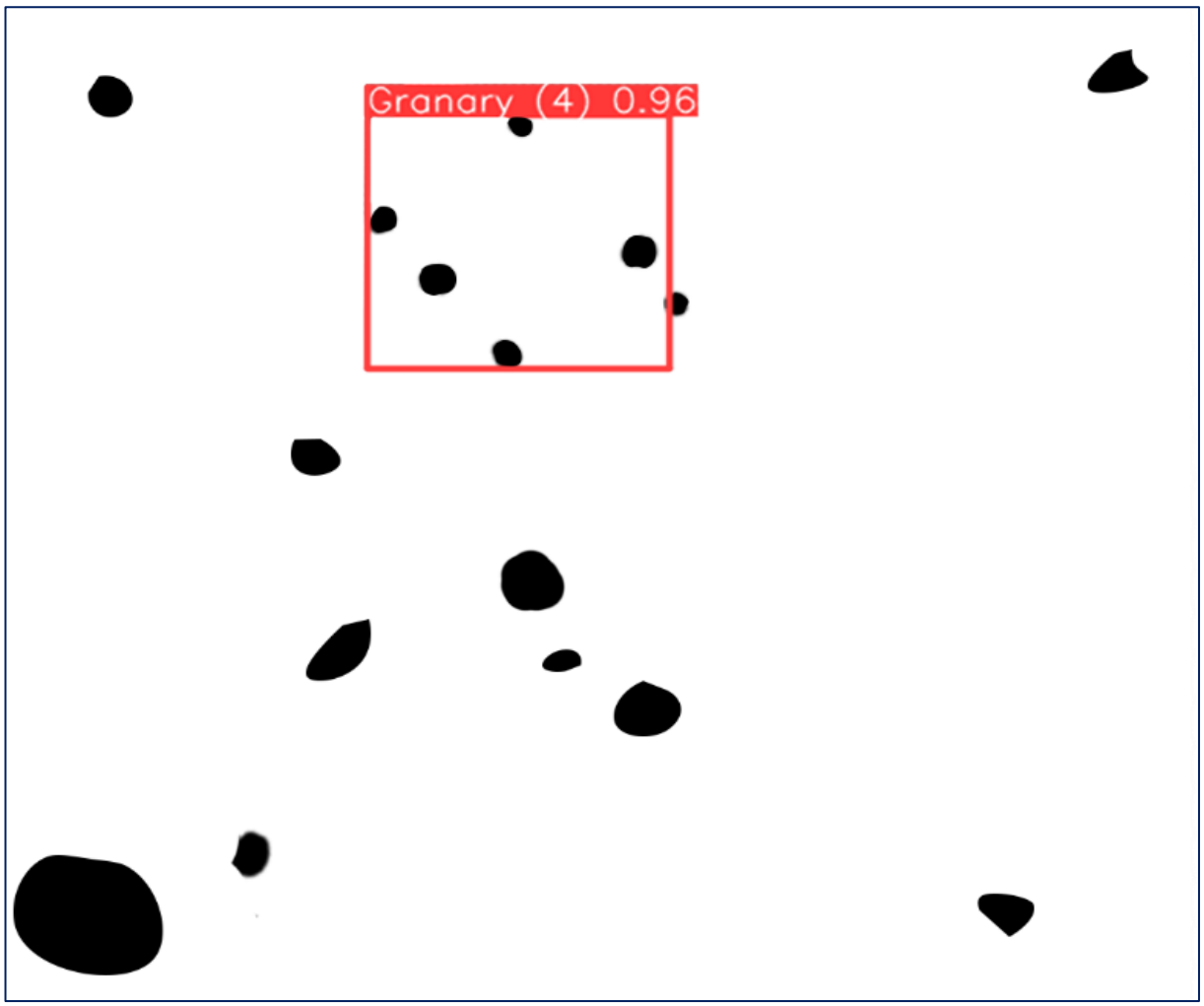


Figure 41: Misclassification of a six-post granary with a missing post. The missing post causes the granary to visually resemble a four-post structure, which confuses the model and results in a misclassification.

In conclusion, it is clear there are several biases and implications present within the dataset that impact the performance of the models. The dataset's inherent biases, such as the visual similarities between several different labelling categories, likely contribute to the models' misclassifications.

7.3 Methodological mistakes

Many of the previously highlighted issues related to the challenges in the models' performance can be attributed to the overarching methodology employed in this thesis. A combination of inexperience with DL model creation and the limited availability of similar models within this specific context has introduced notable flaws into the methodological framework. These shortcomings became evident only after the models were developed and their results analysed, leaving limited opportunity to refine the methodology or test potential improvements within the scope of this research. Consequently, the following section aims to evaluate the methodological framework.

7.3.1 Manual intervention

First of all, something that is echoed throughout the entirety of this thesis is the amount of manual intervention done during the creation of the models. While these interventions were implemented with the intention of optimising conditions to enable the DL model to perform at its best, they inadvertently introduced significant levels of modification that undermined this very objective. By excessively tailoring the environment and inputs, the approach probably compromised the models' generalisability and scalability, limiting its ability to perform well in less controlled, real-world scenarios.

Most of the concerns lie within the data preprocessing methods employed in this thesis. While the dataset was meticulously cleaned to remove noise from the background imagery, this process likely introduced unintended consequences. Specifically, the model trained on such pristine data struggles to handle background noise effectively, as most of the images used during training were unrealistically clean. This presents a limitation when applying the model to real-life archaeological data, which rarely, if ever, conforms to such ideal conditions. In practical scenarios, archaeological data is inherently noisy, often containing clutter, environmental disturbances, and variations in lighting or composition. Consequently, the models' applicability in real-world settings is compromised, as it lacks the robustness needed to perform reliably on unprocessed, authentic datasets.

Correspondingly, the removal of granary instances where posts were missing, although initially deemed necessary to maintain dataset consistency and ensure model accuracy, introduced additional limitations. This decision excluded valuable variations that might have helped the model develop a broader understanding of granary structures. By eliminating these imperfect examples, the model was deprived of the opportunity to learn from incomplete or atypical instances that are often encountered in real-life archaeological contexts. As a result, the models' performance in recognising or interpreting granaries with missing posts in authentic datasets may be significantly hindered, further reducing its practical applicability.

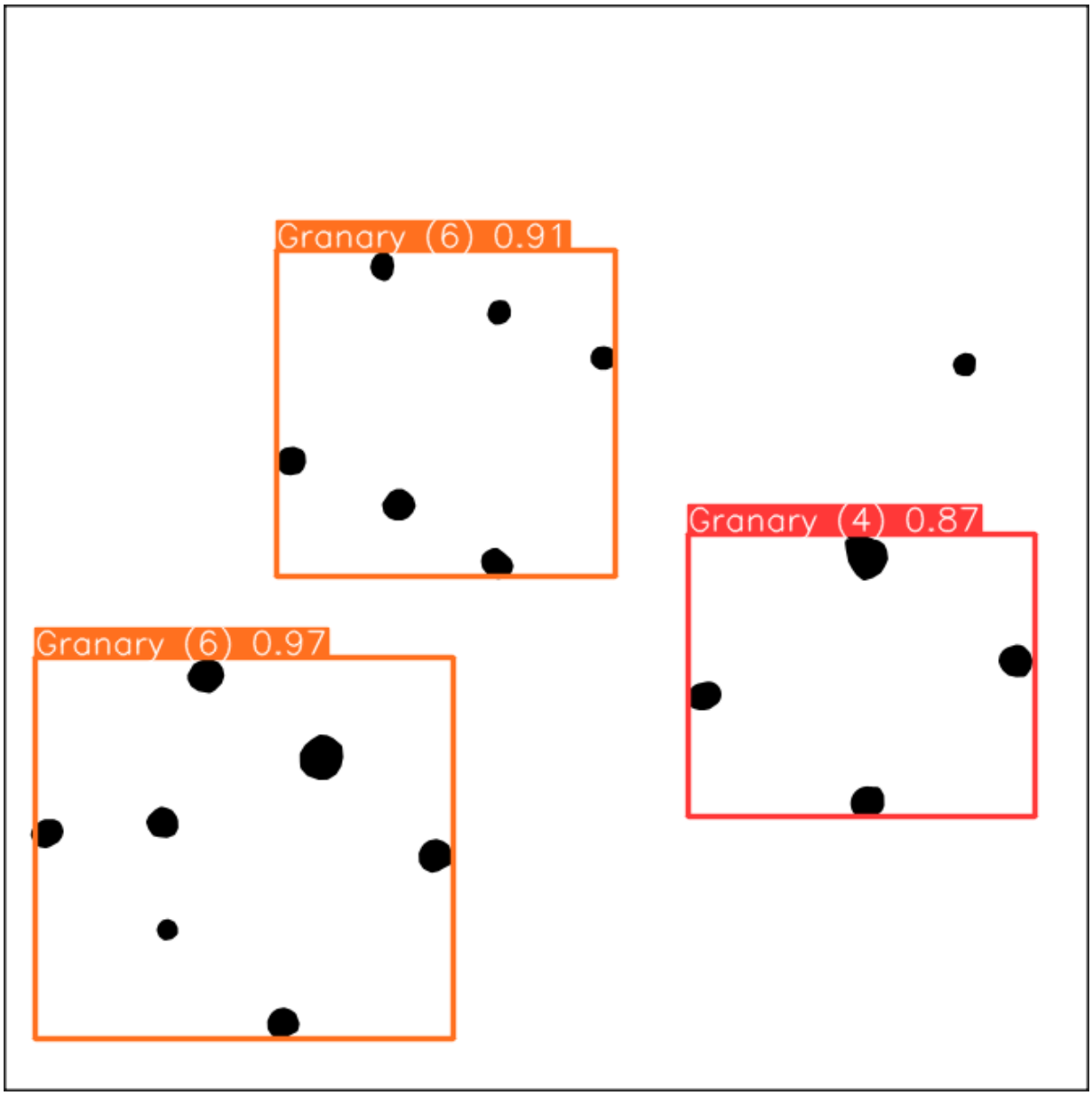


Figure 42: Example of a "perfect" image in which the model could easily identify the granaries. This image is not representative of a regular archaeological dataset.

7.3.2 Labelling categories

Another significant factor that compromised the performance of the DL models was the introduction of the labelling categories. Initially, these categories were considered advantageous, as they were designed to represent distinct structures with varying characteristics and perspectives. The intent was to provide the model with well-defined distinctions to enhance its learning and predictive capabilities. However, upon further reflection, it became evident that these categories introduced unintended complexities. Instead of simplifying the models' learning process, the categories may have fragmented the data in ways that hindered their ability to generalise effectively. This over-compartmentalisation likely resulted in decreased model adaptability and performance, particularly as the granaries are, in hindsight, quite similar in outlook and could therefore have been clustered together in a single category.

When introducing a single "granary" category some of the issues highlighted before would probably have become less problematic. For instance, the double identification of the Granary (9) class into multiple smaller Granary (4) identifications might not have occurred. Therefore, the labelling strategy within this thesis is something that would possibly have great benefits for the results of the model.

7.3.3 Transfer learning

Another issue within the methodology is related to the transfer learning process employed with the pretrained YOLOv8 model. As outlined in the methodology, this model was pretrained on the widely used COCO dataset, which is a common practice in deep learning applications. However, after further reflection and review of recent research, it appears that transfer learning in this context may have been counterproductive (Pires de Lima & Marfurt, 2019; Verschoof-van der Vaart, 2022).

Verschoof-van der Vaart (2022) summarises the issue: '[b]ecause the image characteristics of both types of imagery, grayscale LiDAR versus RGB-coloured photographs, are quite different, it has been suggested that the effectiveness of transfer-learning declines as primary data, i.e., the images used for pre-training, and secondary data, i.e., the images used for fine-tuning, become less similar' (p. 128). While this thesis

deals with different data sources, the principle remains relevant. Black-and-white images from GIS datasets are fundamentally different from the RGB images used in the COCO dataset, both in terms of colour characteristics and the overall image structure. This discrepancy between grayscale and colour imagery may have hindered the performance of the pretrained YOLOv8 model.

Furthermore, the COCO dataset, includes traditional or "normal" objects which are often large, prominent, and reliably orientated (Verschoof-van der Vaart, 2022). Whereas archaeological DL models, including this thesis, are oftentimes looking for small and scarcely distributed objects, where '(...) one class is over-represented (Idem, 2022, p. 125). Therefore, aside from the colour difference, the COCO dataset is fundamentally distinct from the archaeological datasets used here, both in terms of the object characteristics and their distribution. Thus questioning that the applicability of this transfer-learning method for this particular context. Unfortunately, as previously discussed, no large GIS datasets with labelled archaeological structures exist (Opitz & Herrmann, 2018), which further complicates the issue.

Still, whether or not this pretraining approach would benefit the models' performance is uncertain, and remains a point of contention in the general DL discipline (Trier *et al.*, 2019; Zoph *et al.*, 2020; Verschoof-van der Vaart, 2022).

7.3.4 Evaluation metrics

Lastly, during the final phases of this thesis, a new evaluation metric came to light that could have provided a more insightful analysis of the models' performance. This metric, the centroid-based measure was introduced by Fiorucci *et al.* (2022). In essence, this measure was created because '(...) standard evaluation measures are ill-fitted for [archaeological] tasks due to inherent differences between archaeological objects and more common objects and their disregard of geospatial information' (p. 15). The centroid-based methodology assesses whether the centroid of a predicted object falls within the closest ground truth bounding box, reflecting the importance of detecting objects within their expected locations. As they explain '(...) a prediction is considered as a TP if the predicted object's centroid falls inside the area of (at least) one ground truth's bounding box; otherwise, the prediction is considered as an FP. The association between a ground

truth object and its eventual prediction is exclusive: whenever a TP prediction is associated with a ground truth, the latter cannot be associated with any further predictions' (p. 6).

This evaluation measure emphasises the importance of precise locations, as '(...) this crucial aspect is not grasped by the IoU that computes the extension of the overlap between the predicted and the ground truth bounding boxes' (Ibidem). This metric would have been beneficial for this thesis, as it places a higher priority on the spatial accuracy of object detection, an aspect that is crucial in archaeological contexts where the exact positioning of structures is often as important as identifying them. By focusing on centroids rather than just the overlap of bounding boxes, this approach could have been a better evaluation metric within this thesis.

Algorithm 1: Centroid-based measure

Input: Annotation and prediction bounding boxes
Output: True Positive (TP), False Positive (FP), False Negative (FN)

```
for each class do
    prediction_shapes ← compute_shapes_from_predictions
    ground_truth_shapes ← compute_shapes_from_annotations
    ground_truth_shapes_copy ← clone(ground_truth_shapes)
    for p in prediction_shapes do
        for g in ground_truth_shapes_copy do
            if g contains centroid_of p then
                add p to the set of predicted_TP
                remove g from ground_truth_shapes_copy
        if p is not in the set of predicted_TP then
            add p to the set of predicted_FP
for each class do
    for g in ground_truth_shapes do
        for p in prediction_shapes do
            if g contains centroid_of p then
                add g to the set of ground_truth_TP
                remove p from predicted_shapes
        if g is not in the set of ground_truth_TP then
            add g to the set of predicted_FN
```

Figure 43: Algorithm of the centroid-based evaluation metric (Fiorucci *et al.*, 2022, figure 1).

8. Conclusion

This thesis research has demonstrated the potential of automated detection techniques in archaeological research, specifically through the application of Deep Learning algorithms. By focusing on the YOLOv8s model for identifying Bronze and Iron Age granaries on archaeological GIS excavation maps, this research has illustrated both the strengths and limitations of using advanced AI technologies in this context. Utilising an image-based dataset, the models was trained to identify three primary categories of granaries in the Netherlands: four-post, six-post, and nine-post structures. With an average mean Average Precision at IoU 0.5 of 0.861 across all classes, this performance highlights the models' effectiveness in distinguishing these granary types. Although these research findings suggest that the YOLOv8 model shows considerable promise, in its current form, it is not yet adequate and usable for practical archaeological applications. The models' limitations and methodological problems, as highlighted in the discussion, emphasise the need for further refinement. Adequacy in this context requires not only high technical performance but also the ability to generalise across diverse archaeological contexts, handle incomplete datasets, and produce interpretable results that fit within broader archaeological narratives. Therefore, there is still a necessity for the solving of methodological issues, integration of contextual data, and the need for further recall refinement to enhance the model's accuracy and reliability in real-world applications.

8.1 Answer to the research questions

The research questions designed in chapter 2 relate to the evaluation of the effectiveness and implications of DL algorithms in the context of archaeological feature detection and analysis. These can be answered as follows.

8.1.1 Main question

How can the YOLOv8 algorithm be effectively employed to automate the detection and analysis of Bronze and Iron age granaries within archaeological excavation maps, and to what extent can this approach potentially enhance the efficiency and accuracy of archaeological site documentation and analysis?

Based on the preceding paragraphs, it can be carefully concluded that the use of a DL model for predicting the location of prehistoric granaries holds significant potential. The performance metrics are good, and with the added methodological improvements, a hypothetical future model might be considered adequate enough to be implemented on a practical level within archaeological research. However, the question remains whether, even with a theoretically functional model, DL can be deemed sufficiently adequate for this complex task.

Therefore the concept of model adequacy should be addressed. Determining when a model is adequate enough to be considered a practical tool for archaeological research is a complex issue (Huggett, 2020; Argyrou & Agapiou, 2022; Huggett, 2022; Kadhim & Abed, 2023). Adequacy in this context involves more than just achieving high precision and recall metrics during the training phase; it also encompasses the model's ability to generalise across different sites, its versatility in handling diverse and incomplete datasets, and its capacity to produce results that are not only accurate but also meaningful in the context of archaeological interpretation. Furthermore, an adequate model must consistently perform well under varying conditions, including different types of excavation maps, varying preservation states of archaeological features, and different temporal contexts. Most importantly, adequacy also relates to the interpretability of the model's outputs. Archaeologists need to understand the basis of the model's predictions to trust its results and to integrate them into broader archaeological narratives (Barredo Arrieta *et al.*, 2020; Labba *et al.*, 2023; Li *et al.*, 2023; Matrone *et al*. 2023; Tenzer *et al.*, 2024; Vadineanu *et al.*, 2024). A model that performs well on a technical level but lacks transparency in its decision-making process may be inadequate for practical use, as it could lead to misinterpretations or overlooked contextual factors (Gattiglia, 2022). Ultimately, deciding when a model is "adequate" requires a balance between technical performance and practical usability. It involves a continuous dialogue between the model creators and archaeologists, where the model's capabilities are aligned with the specific goals and challenges of archaeological research (Huggett, 2022). Only when a model meets these criteria can it be considered truly adequate for practical application in the field.

The model developed for this thesis currently falls short of several essential criteria, and, as such, cannot be regarded as a practical methodology for use. It is encumbered by numerous implications and biases introduced through the dataset and the methodology, and its lack of transparency regarding prediction weights hinders the ability to understand the decision-making mechanisms employed by the model. This means that the model, in its current form, is too opaque and unreliable to use for practical archaeological inferences.

However, it is important to note that the primary goal of this thesis was not to develop a fully functional model, but to evaluate whether this methodology has potential for future effective use. In response to that question, the conclusion of this thesis is a cautious yes. Even at its basic level, these models have demonstrated that the results are promising and suggest promising pathways for further development and refinement. Implementing more representative data, incorporating feedback loops, appropriate transfer-learning (Verschoof-van der Vaart, 2022), and exploring different labelling strategies are all factors that would potentially enhance the model's performance and utility in future applications.

If such a model were developed for this specific purpose, it would undoubtedly be advantageous for archaeological research. It could significantly reduce the time, energy, and resources required for detecting and analysing archaeological structures, thereby streamlining the research process and enhancing the efficiency of fieldwork and data analysis (Huggett, 2020). However, in this situation, it still important to heavily emphasise the agency such a model should have. The integration of cultural and historical context, and the accompanying subjective interpretations, into AI currently remains a distant goal. Therefore, while the model could greatly assist in identifying potential locations, it should be used in conjunction with expert knowledge and not as a replacement for the nuanced understanding that human researchers bring to archaeological analysis (Ibidem).

8.1.1 Subquestion 1

How can the Deep Learning YOLOv8 algorithm be adapted and optimised to effectively recognise diverse architectural features representative of Bronze and Iron age granaries on archaeological GIS excavation maps?

First, enhancing the quality and quantity of the training data is crucial. YOLOv8's performance is significantly influenced by the diversity of the data it is trained on. Therefore, expanding the dataset to include a broader range of granary examples, preservation states, and archaeological contexts is essential. This expansion will enable the model to learn to identify various granary features more accurately and improve its ability to generalise across different contexts. Additionally, addressing data imbalances by increasing the number of labelled examples for less common granary types, such as the nine-post granary, will help improve the model's recall and accuracy. Ensuring that the dataset includes diverse examples of obscured and overlapping granaries will also aid in reducing misclassifications and improving the model's performance in complex scenarios. Ultimately, reducing the amount of manual intervention introduced within this methodology. Instead of trying to make the "perfect" dataset, a model should work with the data that is out there.

Next, as outlined in the discussion, leveraging appropriate transfer learning by using pre-trained weights and adjusting them based on the granary dataset can also optimise the model's performance for this specialised task (Pires de Lima & Marfurt, 2019; Verschoof-van der Vaart, 2022).

Lastly, integrating expert knowledge and archaeological context is crucial for ensuring the predictions are contextually relevant. Collaborating with archaeologists provides valuable insights into the significance of various granary features and their historical context. Additionally, using contextual information as another data source, such as post-hole depth, texture, colour, shape, and stratigraphy, would offer the model a more holistic understanding of the granaries (e.g. Verschoof-van der Vaart *et al.*, 2020). This multi-faceted approach would enrich the model's capacity to differentiate between various features and nuances, moving beyond a purely pixel-based analysis to a more contextually informed methodology.

8.1.1 Subquestion 2

What are the limitations, biases, and challenges associated with implementing the Deep Learning YOLOv8 algorithm for automated feature detection within archaeological contexts, and how can these challenges be addressed to ensure the reliability and accuracy of the automated identification process?

One notable limitation of the model's performance is its inconsistency across different granary categories. For instance, the Granary (9) category, which frequently achieves perfect precision and recall metrics, highlights a potential issue with the model's capabilities. The disproportionate representation of Granary (9) in the training and test sets, combined with potential similarities introduced by data augmentation, raises concerns about the reliability of the performance metrics. To address this, it is crucial to expand the dataset with more diverse examples, including various granary types and conditions, to better assess the model's true generalization capabilities. Furthermore, adjusting the labelling strategy to include just one overarching "granary" category would simplify and improve the model's overall performance.

Further, while the models show high precision in detecting granaries, the recall is lower, indicating that a significant portion of actual granaries is missed. This trade-off between precision and recall is particularly impactful in archaeological research, where identifying as many potential granaries as possible is crucial. A model with higher recall but lower precision might be more beneficial, as it would generate a larger number of potential candidates for expert validation. To improve recall, strategies such as increasing the diversity of training data, enhancing data augmentation techniques, and adjusting detection thresholds should be considered.

Challenges also arise from the models' difficulty in handling background noise and distinguishing granaries from cluttered backgrounds. The performance of the model deteriorates with increased background noise, affecting its ability to accurately identify granaries in more complex datasets. This sensitivity to background noise underscores the need for the model to manage and interpret noisy data effectively to be useful in archaeological research. Therefore, less manual intervention and more realistic cluttered data would enhance the model's overall performance.

The models also struggle with classifying overlapping or incomplete features and differentiating between visually similar categories. Misclassifications between granary categories, such as confusing six-post granaries with missing posts for four-post granaries, highlight the need for a more all-encompassing training dataset. Including more examples of incomplete or atypical features can help the model better understand and classify such variations. Additionally, the presence of other structures, such as prehistoric house plans, within the dataset has led to confusion due to visual similarities with granaries. This issue suggests, once again, the potential for reducing the classification categories to include one overarching class.

Lastly, the concept of DL bias in general is something both the user and the creator of such a methodology should be highly aware of. It is essential to recognise that biases in training data or model design can significantly impact the accuracy and fairness of predictions. Therefore, the dataset should be thoroughly assessed for characteristics that could introduce bias, and measures should be taken to mitigate these issues. By ensuring the data and model are as unbiased and representative as possible, the reliability and equity of the automated identification process can be substantially improved.

All in all, there are several important considerations to keep in mind when designing and using a DL model in archaeological research. Being aware of these components and trying actively to mitigate them in the future will ensure that such methodologies are used responsibly and contribute meaningfully to the field.

8.2 Future research

Looking to the future of integrating DL methodologies in archaeological research, there is significant potential to enhance the understanding and analysis of archaeological structures. Ideally, this research has created a foundation for further exploration into refining and expanding DL models on archaeological GIS maps. This subchapter will therefore include suggestions for future research and adaptation to the methodology and its corresponding DL models. Besides the model improvements that have already been mentioned several times before.

One of the major issues within AI research is the "black-box issue," which refers to the lack of transparency in how deep learning models arrive at their predictions. Consequently, a major upcoming area of research is XAI (Barredo Arrieta *et al.*, 2020; Labba *et al.*, 2023; Li *et al.*, 2023; Matrone *et al*. 2023; Tenzer *et al.*, 2024; Vadineanu *et al*., 2024). This specific field focuses on developing methods and tools that make the decision-making processes of AI models more transparent and interpretable. For archaeological research, implementing XAI techniques can help bridge the gap between complex algorithms and human expertise, ensuring that model outputs are not only accurate but also understandable and justifiable within the context of archaeological inquiry. Practical examples of XAI for the YOLO model include heatmaps, saliency maps, and feature maps (Doran *et al.*, 2018; Dwivedi *et al.*, 2023; Moradi *et al.*, 2024). Heatmaps illustrate the regions of an image that the YOLO model deems most significant for its predictions. By highlighting areas with varying intensities of attention, heatmaps clarify which parts of the image influence the model's detection results. Saliency maps, highlight the specific pixels that most affect the model's decision-making. They show how changes in pixel values might impact the prediction, thus revealing the key features that drive object detection (Moradi *et al.* 2024). Feature maps display the activations of different layers within the YOLO network, providing insights into the features extracted at various processing stages (e.g. Vadineanu *et al*, 2024). By visualising these activations, feature maps help understand how the model interprets different aspects of the image and contributes to object detection (Moradi, 2024, p. 19). Together, these types of XAI tools help disentangle YOLO's prediction process, offering a clearer view of how input data is processed, and which components are pivotal in generating results. Although these methods are not completely all-encompassing, and sometimes difficult to interpret, they might aid in the understanding of the predictions the YOLO algorithm makes during the identification of granary structures.

Another potential future direction is the integration of this (working) DL algorithm into a GIS environment, such as through an installable plugin. This approach would offer significant advantages to users by facilitating the identification of granaries directly within their GIS platforms. It would also enhance accessibility by making the tool available to a broader range of software users, thereby fostering greater collaboration and

application across the archaeological research community (Batist & Roe, 2024). Openly available research tools and science are essential for advancing DL knowledge and fostering collaboration within the academic community (Schmidt & Marwick, 2020). By making such resources accessible to a wider audience, researchers can more easily build upon each other's work, validate findings, and contribute to collective progress. This open approach not only democratises access to technologies but also accelerates innovation and ensures that advancements in fields like archaeology benefit from diverse insights and applications.

Lastly, as previously discussed, incorporating multiple categories within the detection algorithm would be a significant enhancement. Expanding the model to recognise a broader range of features and classifications could greatly improve its utility in archaeological research by providing more holistic insights. Including house plans, grave structures, barrows, or any other category with a relatively standardised typology would enrich the methodological analytical capabilities and broaden its applicability. Although this would entail substantial amounts of annotated archaeological data and extensive training time (Opitz & Herrmann, 2018; Verschoof-van der Vaart, 2022), this addition is not insurmountable. Ultimately, because the addition of extra classes would eventually '(...) enhance the generalisation capabilities of Deep Learning approaches' (Verschoof-van der Vaart, 2022, p. 149). With advancements in computational resources and data management techniques, such challenges could potentially be addressed. The effort required to incorporate a wider range of categories is justified by the potential for significantly improving the model's accuracy and versatility, ultimately leading to more insightful and comprehensive archaeological analyses.

Works Cited

Anichini, F., & Gattiglia, G. (2022). Reflecting on artificial intelligence and archaeology: the ArchAIDE perspective. *European Journal of Postclassical Archaeologies, 12*(1), 69-86.

Antolín, F., & Schäfer, M. (2024). Insect pests of pulse crops and their management in Neolithic Europe. *Enviornmental Archaeology, 29*(1), 20-33. doi:10.1080/14614103.2020.1713602

Argyrou, A., & Agapiou, A. (2022). A review of artificial intelligence and remote sensing for archaeological research. *Remote Sensing, 14*(23), 1-23. doi:10.3390/rs14236000

Arnoldussen , S., & Fokkens, H. (2009). *Bronze Age settlements of the Low Countries*. Oxford: Oxbow Books.

Arnoldussen, S. (2008). *Living landscape: Bronze Age settlement sites in the Dutch river area (c. 2000-800 BC)*. [Doctoral dissertation, Leiden University].

Arnoldussen, S., & Theunissen, E. M. (2013). Huisplattegronden uit de late prehistorie in het rivierengebied. In A. G. Lange, E. M. Theunissen, J. H. Deeben, J. van Doesburg, J. Bouwmeester, & T. de Groot (Eds.), *Huisplattegronden in Nederland: Archeologische sporen van het huis* (pp. 115-142). Amersfoort: Barkhuis & Rijksdienst voor het Cultureel Erfgoed.

Bakels, C. (2009). *The western European loess belt: agrarian history, 5300 BC; AD 1000*. Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-9840-6

Bakels, C. (2014). The first farmers of the Northwest European plain: some remarks on their crops, crop cultivation and impact on the environment. *Journal of Archaeological Science, 51*(1), 94-97. doi:10.1016/j.jas.2012.08.046

Baredo Arrieta, A., Díaz-Rodríquez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*(1), 82-115. doi:10.1016/j.inffus.2019.12.012

Batist, Z., & Roe, J. (2024). Open Archaeology, Open Source? Collaborative practices in an emerging community of archaeological software engineers. *Internet Archaeology, 67*(1). doi:10.11141/ia.67.13

Benallou, M. (2021). Sporen van rituele depositie in nederzettingscontexten tijdens the ijzertijd in België: een vergelijkende studie met rituele praktijken in de aanpalende regio's. (Master's thesis). Gent University.

Binford, L. R. (1982). The archaeology of place. *Journal of Anthropological Archaeology, 1*(1), 5-31. doi:10.1016/0278-4165(82)90006-X

Blake, E. (2007). Space, spatiality, and archaeology. In L. Meskell, & R. W. Preucel, *A companion to social archaeology* (pp. 230-254). Oxford: Blackwell Publishing Ltd.

Blakely, S. (2023). Agency, affect, games, and gods: Archaeogaming and the archaeology of religion. . *Data Science, Human Science, and Ancient Gods: Conversations in Theory and Method, 3*(1), 283-288.

Bodenhamer, D. J. (2012). Space, time and place in the new digital age. In D. J. Bodenhamer, *The spatial humanities* (pp. 23-38). Indiana: Indiana University Press.

Bogaard, A. (2017). The archaeology of food surplus. *World Archaeology, 49*(1), 1-7. doi:10.1080/00438243.2017.1294105

Bogaard, A., Charles, M., Twiss, K. C., Fairbairn, A., Yalman, N., Filipović, D., . . . Henecke, J. (2009). Private pantries and celebrated surplus: storing and sharing food at Neolithic Çatalhöyük, central Anatolia. *Antiquity, 83*(321), 649-668. doi:10.1017/S0003598X00098896

Brijker, J., Deitch-van der Meulen, W., van Dinter, M., Drenth, E., Meikert, M. J., Moolhuizen, C., & Prangsma, N. M. (2012). *Prehistorische boerderijen onder de stal: een inventariserend veldonderzoek in de vorm van proefsleuven en een archeologische opgraving te Eefde Schurinklaan 49, Gemeente Lochem*. Amersfoort: ADC Archeoprojecten.

Brück, J. (1999). Houses, lifecycles and deposition on Middle Bronze Age settlements in southern England. *Proceedings of the Prehistoric Society, 65*(1), 145-166. doi:10.1017/S0079497X00001973

Bundzel, M., Jaščur, M., Kováč, M., Lieskovský, T., Sinčák, P., & Tkáčik, T. (2020). Semantic Segmentation of Airborne LiDAR Data in Maya Archaeology. *Remote Sensing, 12*(22), 3865. doi:10.3390/rs12223685

Campana, S. (2008). Archaeological site detection and mapping: some thoughts on differing scales of detail and archaeological 'non-visibility'. In S. Campana, & S. Prio (Eds.), *Seeing the unseen. Geophysics and landscape archaeology* (pp. 31-52). CRC Press. ISBN 978-0-415-44721-8.

Campana, S. (2017). Drones in archaeology. State-of-the-art and future perspectives. *Archaeologcial Prospection, 24*(4), 275-296. doi:10.1002/arp.1569

Campillo, X. R., Cela, J. M., & Cardon, F. X. (2012). Simulating archaeologists? Using agent-based modelling to improve battlefield excavations. *Journal of Archaeological Science, 39*(2), 347-356. doi:10.1016/j.jas.2011.09.020

Canedo, D., Fonte, J., Seco, L. G., Vázquez, M., Dias, E., Do Pereiro, T., . . . Neves, A. J. (2023). Uncovering archaeological sites in airborne LiDAR data with data-centric

Artificial Intelligence. *IEEE Access, 11*(1), 65608 - 65619. doi:10.1109/ACCESS.2023.3290305

Carabantes, M. (2019). Black-box artifcial intelligence: an epistemological and critical analysis. *AI & Society, 35*(1), 309-317. doi:10.1007/s00146-019-00888-w

Casagrande, L., Wolf, C., A.D, M., Nordlander, T., Yong, D., & Bessel, M. (2019). SkyMapper stellar parameters for galactic archaeology on a grand scale. *Montly Notices of the Royal Astronomical Society, 482*(2), 2770-2787. doi:10.1093/mnras/sty2878

Cheng, G., & Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing, 117*(1), 11-28. doi:10.1016/j.isprs.2016.03.014

Childe, V. G. (1942). *What happened in history.* London: Penguin.

Cowley, D., Verhoeven, G., & Traviglia, A. (2021). Archaeologcial remote sensing in the 21st century: (re)defining practice and theory. *Remote Sensing, 13*(1431), 1-4. doi:10.2290/rs13081431

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science, 40*(2), 130-139. doi:10.1080/15230406.2013.777137

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: politics, ethics, epistemology. *International Journal of Communication, 8*(10), 1663-1672.

Cunningham, P. (2011). Caching your savings: the use of small-scale storage in European prehistory. *Journal of Anthropological Archaeology, 30*(2), 135-144. doi:10.1016/j.jaa.2010.12.005

Dallas, C. (2015). Curating archaeological knowledge in the digital continuum: from practice to infrastructure. *Open Archaeology, 1*(1), 176-207. doi:10.1515/opar-2015-001

Daly, P., & Evans, T. L. (2006). *Digital archaeology: bridging method and theory.* New York: Routledge.

Darmangeat, C. (2020). Surplus, storage and the emergence of wealth: pits and pitfalls. In L. Moreau, *Social inequality before farming? Multidisciplinary approaches to the study of social organization in prehistoric and ethnographic hunter-gatherer-fisher societies* (pp. 59-70). Cambridge: McDonald Institute for Archaeological Research. doi:10.17863/CAM.60644

Davies, B., Romanowska, I., Harris, K., & Crabtree, S. (2019). Combining geographic information systems and agent-based models in archaeology: Part 2 of 3. *Advances in Archaeologcial Practice, 7*(2), 185-193. doi:10.1017/aap.2019.5

Davies, D. (2019). Object-based image analysis: A review of developments and future directions of automated feature detection in landscape archaeology. *Archaeologcial Prospection, 26*(2), 155-163. doi:10.1002/arp.1730

de Laet, V., & Lambers, K. (2009). Archaeological prospection using high-resolution digital satellite imagery: recent advances and future prospects. *Computer Applications and Quantitative Methods in Archaeology (CAA) Conference. 39*, pp. 9-17. Williamsburg: AARGnews - The newsletter of the Aerial Archaeology Research Group. Retrieved June 2024, from https://kops.uni-konstanz.de/entities/publication/5998ad63-ee0d-4ad4-843b-46f21aaa383f

Deeben, J., & Theunissen, E. (2013). De huisplattegrond als archeologisch studieobject: Een korte bespiegeling over de theoretische achtergronden. In A. Lange, E. Theunissen, J. Deeben, J. van Doesburg, J. Bouwmeester, & T. de Groot (Eds.), *Huisplattegronden in Nederland: Archeologische sporen van het huis* (pp. 5-17). Amersfoort: Barkhuis & Rijksdienst voor het Cultureel Erfgoed.

Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint*, 1-8. doi:10.48550/arXiv.1710.00794

Drennan, R. D. (2001). Numbers, models, maps: computers and archaeology. In D. R. Brothwell, & A. M. Pollard, *Handbook of archaeological sciences* (pp. 663-670). New York: John Wiley and Sons.

Dunn, S. (2017). Praxes of "the human" and "the digital": spatial humanities and the digitization of place. *Geohumanities, 3*(1), 88-107. doi:10.1080/2373566X.2016.1245107

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), 1-33. doi:10.1145/3561048

Fiorucci, M., Verschoof-van der Vaart, W. B., Soleni, P., Le Saux, B., & Traviglia, A. (2022). Deep Learning for Archaeological Object Detection on LiDAR: New Evaluation Measures and Insights. *Remote Sensing, 14*(7), 1694. doi:10.3390/rs14071694

Flexner, J. (2009). Where is reflexive map-making in archaeological research? Towards a place-based approach. *Archaeological Review from Cambridge, 24*(1), 7-21.

Forte, M. (2011). Cyber-Archaeology: Notes on the simulation of the past. *Virtual Archaeology Review, 2*(4), 7-18. doi:10.4995/var.2011.4543

François, P., Leichman, J., Laroche, F., & Rubellin, F. (2021). Virtual reality as a versatile tool for research, dissemination, and mediation in the humanities. *Virtual Archaeology Review, 12*(25), 1-15. doi:10.4995/var.2021.14880

Gatiglia, G. (2015). Think big about data: archaeology and the big data challenge. *Archäologische Informationen, 38*(1), 113-124. doi:10.11588/ai.2015.1.26155

Gattiglia, G. (2022). A postphenomenological perspective on digital and algorithmic archaeology. *Archeologia e Calcolatori, 33*(2), 319-334. doi:10.19282/ac.33.2.2022.17

Gent, H. (1983). Centralized storage in Later Prehistoric Britain. *Proceedings of the, 49*(1), 243-267. doi:10.1017/S0079497X00008008

Gillings, M., Hacigüzeller, P., & Lock, G. (2020). Archaeology and spatial analysis. In M. Gillings, P. Hacigüzeller, & G. Lock, *Archaeological spatial analysis: a methodological guide* (pp. 1-16). New York: Routledge.

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

Haba, D. (2023). *Data Augmentation with Python: Enhance deep learning accuracy with data augmentation methods for image, text, audio, and tabular data*. Birmingham: Packt Publishing Ltd.

Halstead, P., & O'Shea, J. (1989). Introduction: cultural responses to risk and uncertainty. In P. Halstead, & J. O'Shea, *Bad years economics. Cultural responses to risk and uncertainty* (pp. 1-7). Cambridge: Cambridge University Press.

Hastorf, C. A., & Foxhall, L. (2017). The social and political aspects of food surplus. *World Archaeology, 49*(1), 26-39. doi:10.1080/00438243.2016.1252280

Hermsen, I., & Haveman, E. (2009). *Op het spoor van de Holterweg: archeologisch en historish onderzoek van, onder en langs de Holterweg in Colmschate (Gemeente Deventer).* Deventer: Rapportages Archeologie Deventer.

Hodder, I. (2001). *Archaeological theory today.* Cambridge: Blackwell Publishers.

Huggett, J. (2015). Challenging Digital Archaeology. *Open Archaeology, 1*(1), 79-85. doi:10.1515/opar-2015-0003

Huggett, J. (2020). Is big digital data different? Towards a new archaeological paradigm. *Journal of Field Archaeology, 45*(1), 8-17.

Huggett, J. (2021). Algorithmic agency and autonomy in archaeological practice. *Open Archaeology, 7*(1), 417-434. doi:10.1515/opar-2020-0136

Huggett, J. (2022). Archaeological practice and digital automation. In E. Watrall, & L. Goldstein, *Digital Heritage and Archaeology in Practice: Data, Ethics, and Professionalism* (pp. 275-304). Gainesville: University Press of Florida.

Huggett, J. (2024). Changing Theory and Practice? CAA and Archaeology's Digital Turn. *Journal of Computer Applications in Archaeology, 7*(1), 316-331. doi:10.5334/jcaa.144.

Huggett, J., Reilly, P., & Lock, G. (2018). Whither digital archaeological knowledge? The challenge of unstable futures. *Journal of Computer Applications in Archaeology, 1*, 42-54. doi:10.5334/jcaa.7

Isaksen, L. (2013). O what a tangled web we weave: Towards a practice that does not deceive. In J. Knappett (Ed.), *Network analysis in archaeology: New approaches to regional interaction* (pp. 43-70). Oxford: Oxford University Press.

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), 7-14. doi:10.37934/araset.26.1.714

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A review of YOLO algorithm developments. *Procedia Computer Science, 199*(1), 1066-1073. doi:10.1016/j.procs.2022.01.135

Jiménez-Jáimez, V., & Suárez-Padilla, J. (2019). Understanding pit sites: storage, surplus, and social complexity in prehistoric western Europe. *Journal of Archaeological Method and Theory, 1*, 799-835. doi:10.1007/s10816-019-09429-7

Jongste, P. F., & Knippenberg, S. (2005). *Archol rapport 36: Terug naar Zijderveld*. Archol bv. Leiden: DANS Data Station Archaeology. doi:10.17026/dans-zev-38k4

Kadhim, I., & Abed, F. M. (2023). A critical review of remote sensing approaches and deep learning techniques in archaeology. *Sensors, 23*(6), 2918. doi:10.3390/s23062918

Kansteiner, W. (2022). Digital doping for historians: can history, memory, and historical theory be rendered artificially intelligent? *History and Theory, 61*(4), 119-133. doi:10.1111/hith.12282

Kirleis, W., Klooß, S., Kroll, H., & Müller, J. (2012). Crop growing and gathering in the northern Germanic Neolithic: a review supplemented by new results. *Vegetation History and Archaeobotany, 21*(1), 221-242. doi:10.1007/s00334-011-0328-9

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 14*(1), 1-12. doi:10.1177/2053951714528481

Knappett, C. (2020). Relational concepts and challenges to network analysis in social archaeology. In C. Knappett, *Archaeological networks and social interaction* (pp. 20-37). New York: Routledge.

Kowarik, K., Koch, A., Kutzner, T., & Eder, T. (2012). Agents in archaeology-agent based modelling (ABM) in archaeological research. *Geoinformationssysteme: Beiträge zum, 17*(1), 238-251.

Kuijt, I., & Finlayson, B. (2009). Evidence for food storage and predomestication of granariess 11,000 years ago in the Jordan Valley. *PNAS, 106*(27), 10966-10970. doi:10.1073/pnas.0812764106

Labba, C., Alcouffe, A., Crubézy, E., & Boyer, A. (2023). IArch: An AI Tool for Digging Deeper into Archaeological Data. *2023 IEEE 35th International Conference on*

Tools with Artificial Intelligence (ICTAI), (pp. 22-29). doi:10.1109/ICTAI59109.2023.00012

Lambers, K., Verschoof-van der Vaart, W., & Bourgeois, Q. (2019). Integrating Remote Sensing, Machine Learning, and Citizen Science in Dutch Archaeological Prospection. *Remote Sensing, 11*(7), 794-814. doi:10.3390/rs11070794

Lange, A., Theunissen, E., Deeben, J., van Doesburg, J., Bouwmeester, J., & de Groot, T. (2013). *Huisplattegronden in Nederland: Archeologische sporen van het huis*. Amersfoort: Barkhuis & Rijksdienst voor het Cultureel Erfgoed.

Lercari, N. (2017). 3D visualization and reflexive archaeology: A virtual reconstruction of Çatalhöyük history houses. *Digital Applications in Archaeology and Cultural Heritage, 6*(1), 10-17. doi:10.1016/j.daach.2017.03.001

L'Heureux, A., Grolinger, K., Elyamanay, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access, 5*(1), 7776-7797. doi:10.1109/ACCESS.2017.2696365

Li, P., Zang, Y., Wang, C., Li, J., Cheng, M., Luo, L., & Yu, Y. (2016). Road network extraction via deep learning and line integral convolution. *International Geoscience and Remote Sensing Symposium*, 1599-1602. doi:10.1109/IGARSS.2016.7729408

Li, X., Chen, D., Xu, W., Chen, H., Li, J., & Mo, F. (2023). Explainable dimensionality reduction (XDR) to unbox AI 'black box'models: A study of AI perspectives on the ethnic styles of village dwellings. *Humanities and Social Sciences Communications, 10*(1), 1-13. doi:10.1057/s41599-023-01505-4

Liu, X., Zhao, Z., & Jones, M. K. (2017). From people's commune to household responisbility: ethnoarchaeological perspectives of millet production in prehistoric northeast China. *Archaeological Research in Asia, 11*(1), 51-57. doi:10.1016/j.ara.2017.07.005

Maes, S. (2009). Voedselopslag tijdens the metaaltijden tussen Rijn en Seine. Een studie over spiekers en silo's. *Terra Incognita, 2*(1), 79-90.

Mallick, S. (2021). A deep learning & computer vision based approach to airborne laser scanning data: automated instance segmentation of Celtic fields in LiDAR data from the Veluwe, Netherlands, using mask r-CNN. (Master's thesis). Leiden University, Faculty of Archaeology.

Malrain, F., Matterne, V., & Méniel, P. (2002). *Les paysans gaulois (IIIe siècle – 52 av. J.-C.).* Paris: Editions Errance.

Maltas, T., Şahoğlu, V., Erkanal, H., & Tuncel, R. (2021). Prehistoric farming settlements in western Anatolia: Archaeobotaniccal insights into the Late Chalcolithic of the Izmir region, Turkey. *Journal of Mediterranean Archaeology, 34*(2), 252-277. doi:10.1558/jma21981

Marçal, D., Câmara, A., Oliveira, J., & de Almeida, A. (2024). Evaluating R-CNN and YOLO V8 for Megalithic Monument Detection in Satellite Images. *International Conference on Computational Science. 14834*, pp. 162-170. Málaga: Springer Nature Switzerland AG. doi:10.1007/978-3-031-63759-9_20

Marche, S. (2012). Literature is not data: against digital humanities. *LA Review of Books, 28*(1), 8-31.

Matrone, F., Felicetti, A., Paolanti, M., & Pierdicca, R. (2023). Explaining AI: Understanding Deep Learning Models for Heritage Point Clouds. *29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage:*, (pp. 207-216). Florence. doi:10.5194/isprs-annals-X-M-1-2023-207-2023

Moore, J. A., & Keene, A. S. (1983). Archaeology and the law of the hammer. In J. A. Moore, & A. S. Keene, *Archaeological hammers and theories* (pp. 3-13). New York: Academic Press. doi:10.1016/B978-0-12-505980-0.50007-0

Moradi, M., Yan, K., Colwell, D., Samwald, M., & Asgari, R. (2024). Model-agnostic explainable artificial intelligence for object detection in image data. *arXiv preprint*, 1-26. doi:10.48550/arXiv.2303.17249

Morales, J., Rodríguez-Rodríguez, A., González-Marrero, M. D., Martín-Rodríguez, E., Henríquez-Valido, P., & del-Pino-Curbelo, M. (2014). The archaeobotany of long-term crop storage in northwest African communal granaries: a case study from pre-Hispanic Gran Canaria. *Vegetation history and Archaeobotany, 23*(1), 789-804. doi:10.1007/s00334-014-0444-4

Morgan, C. (2009). (Re)building Çatalhöyük: Changing virtual reality in archaeology. *Archaeologies, 5*(1), 468-487. doi:10.1007/s11759-009-9113-0

Morgan, C. (2019). Avatars, Monsters, and Machines: A Cyborg Archaeology. *European Journal of Archaeology, 22*(3), 324-337. doi:10.1017/eaa.2019.22

Morgan, C. (2022). Current digital archaeology. *Annual Review of Anthropology, 51*(1), 213-231. doi:10.1146/annurev-anthro-041320-114101

Morgan, C. (2022). Current Digital Archaeology. *Annual Review of Anthropology, 51*(1), 213-231. doi:10.1146/annurev-anthro-041320-114101

Morgan, C. L. (2012). Emancipatory digital archaeology. [Doctoral dissertation, University of California].

Moscati, P. (2021). How Big is Big Data? *Big Data and Archaeology: Proceedings of the XVIII UISPP World Congress (June 2018, Paris, France)* (pp. 8-22). Oxford: Archaeopress.

Nikolakopoulos, K., Soura, K., Koukouvelas, I., & Argyropoulos, N. (2017). UAV vs classical aerial photogrammetry for archaeological studies. *Journal of Archaeological Science, 14*(1), 758-773. doi:10.1016/j.jasrep.2016.09.004

Ollich, I., Rocafiguera, M., Ocaña, M., Cubera, C., & Amblàs, O. (2012). Experimental archaeology at l'Escquerda - crops, storage, metalcraft and earthworks in mediaeval and ancient times. In I. Ollich, *Archaeology, new approaches in theory and techniques* (pp. 205-228). Rijeka: INTECH Open Access Publisher. doi:10.5772/38790

Opitz, R., & Herrmann, J. (2019). Recent trends and long-standing problems in archaeological remote sensing. *Journal of Computer Applications in Archaeology, 10*(1), 19-41. doi:10.5334/jcaa.11

Orengo, H. A., Conesa, F. C., Garcia-Molsosa, A., Lobo, A., Green, A. S., Madella, M., & Petrie, C. A. (2020). Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *Proc. Natl. Acad. Sci. USA, 117*(31), 18240-18250. doi:10.1073/pnas.2005583117

Out, W. A. (2009). Sowing the seed? Human impact and plant subsistence in Dutch wetlands during the late Mesolithic and early and middle Neolithic (5500-3400 cal BC). *Archaeological Studies Leiden University, 18*(1), 99-214.

Papadopoulos, N. (2021). Shallow offshore geophysical prospection of archaeological sites in eastern Mediterranean. *Remote Sensing, 13*(7), 1237-1249. doi:10.3390/rs13071237

Peña-Chocarro, L., Pérez Jordà, G., Morales Mateos, J., & Zapata, L. (2015). Storage in traditional farming communities of the western Mediterranean: ethnographic, historical, and archaeological data. *Environmental Archaeology, 20*(4), 379-389. doi:10.1179/1749631415Y.0000000004

Perry, S., Taylor, J. S., Matsumoto, M., & Uleberg, E. (2018). Theorising the digital: a call to action for the archaeological community. *Oceans of Data: Proceedings of the 44th conference on computer applications and quantitative methods in archaeology (CAA)* (pp. 11-22). Oxford: Archaeopress.

Pires de Lima, R., & Marfurt, K. (2019). Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing, 12*(1), 86-106. doi:10.3390/rs12010086

Rasalle, T. (2018). *Archaeogaming: An introduction to archaeology in and of video games*. Berghahn Books.

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263-7271. doi:10.48550/arXiv.1612.08242

Redmon, J., & Farhadi, A. (2018, April 8). *YOLOv3: An incremental improvement*. doi:10.48550/arXiv.1804.02767

Reinhard, A. (2018). *Archaeogaming: an introduction to archaeology in and of video games*. New York: Berghahn Books.

Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). Real-Time Flying Object Detection with YOLOv8. *arXiv preprint*, 1-10. doi:10.48550/arXiv.2305.09972

Reynolds, A. (2009). The archaeology of buildings: introduction. *World Archaeology, 41*(3), 345-347. doi:10.1080/00438240903148470

Richardson, L., & Lindgren, S. (2017). Online tribes and digital authority: what can social theory bring to digital archaeology? *Open Archaeology, 3*(1), 139-148. doi:10.1515/opar-2017-0008

Risbøl, O., Bollandsas, O., Nesbakken, A., Orka, O., Naesset, E., & Gobakken, T. (2013). Interpreting cultural remains in airborne laser scanning generated digital terrain models: effects of size and shape on detection success rates. *Journal of Archaeologcial Sciences, 40*(12), 4688-4700. doi:10.1016/j.jas.2013.07.002

Rowley-Conwy, P., & Zvelebil, M. (1989). Saving it for later: storage by prehistoric hunter-gatherers in Europe. In P. Halstead, & J. O'Shea, *Bad year economics. Cultural responses to risk and uncertainty* (pp. 40-56). Cambridge: Cambridge University Press.

Schinkel, K. (1994). *Zwervende erven. Bewoningssporen in Oss-Ussen uit de Bronstijd, IJzertijd en Romeinse tijd*. [Doctoral dissertation, Leiden University].

Schmidt, S. C., & Marwick, B. (2020). Tool-driven revolutions in archaeological science. *Journal of Computer Applications in Archaeology, 3*(1), 18-32.

SIKB. (n.d.). *BRL SIKB 4000 protocols*. Retrieved from https://www.sikb.nl/archeologie/richtlijnen/brl-sikb-4000

Smith, N., Passone, L., Al-Said, S., Al-Farhan, M., & Levy, T. (2014). Drones in archaeology: integrated data capture, processing, and dissemination in the al-Ula Valley, Saudi Arabia. *Near Eastern Archaeology, 77*(3), 176-181.

Sobotkova, A., Kristensen-McLachlan, R. D., Mallon, O., & Ross, S. A. (2024). Validating predictions of burial mounds with field data: the promise and reality of machine learning. *Journal of Documentation*. doi:10.1108/JD-05-2022-0096

Somrak, M., Džeroski, S., & Kokalj, Z. (2020). Learning to Classify Structures in ALS-Derived Visualizations of Ancient Maya Settlements with CNN. *Remote Sensing, 12*(14), 2215. doi:10.3390/rs12142215

Soroush, M., Mehrtash, A., Khazraee, E., & Ur, J. A. (2020). Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq. *Remote Sensing, 12*(3), 500. doi:10.3390/rs12030500

Stopp, M. P. (2002). Ethnohistoric analogues for storage as an adaptive strategy in northeastern subarctic prehistory. *Anthropological Archaeology, 21*(3), 301-328. doi:10.1016/S0278-4165(02)00004-1

Tenzer, M., Pistilli, G., Brandsen, A., & Shenfield, A. (2024). Debating AI in Archaeology: applications, implications, and ethical considerations. *Internet Archaeology, 67*(1). doi:10.11141/ia.67.8

Terven, J. R., & Cordova-Esparza, D. M. (2024). A comprehensive reviw of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 1-36. doi:10.3390/make5040083

Theunissen, E. M. (1999). *Midden-Bronstijdsamenlevingen in het zuiden van de Lage Landen: Een evaluatie van het begrip 'Hilversum-cultuur'.* [Doctoral dissertation, Leiden University].

Theuws, F. (2013). Vroegmiddeleeuwse huisplattegronden uit Zuid-Nederland en hun weergave. In A. Lange, E. Theunissen, J. Deeben, J. van Doesburg, & T. de Groot (Eds.), *Huisplattegronden in Nederland: Archeologische sporen van het huis* (pp. 313-340). Amersfoort: Barkhuis & Rijksdienst voor het Cultureel Erfgoed.

Toumzet, J., Vautier, F., Roussel, E., & Dousteyssier, B. (2017). Automatic detection of complex archaeologcial grazing strcutures using airborne laser scanning data. *Journal of Archaeological Sciences, 12*(1), 569-579.

Trebsche, P. (2009). Does form follow function? Towards a methodical interpretation of archaeological building features. *World Archaeology, 41*(3), 505-519. doi:10.1080/00438240903112534

Ultralytics. (2023). YOLOv8 Python package. Retrieved from https://github.com/ultralytics/ultralytics

Vadineanu, S., Kalayci, T., Pelt, D. M., & Batenburg, K. J. (2024). Convolutional Neural Networks and Their Activations: An Exploratory Case Study on Mounted Settlements. *Journal of Computer Applications in Archaeology, 7*(1), 262-282. doi:10.5334/jcaa.163

van den Broeke, P. W. (2002). Een vurig afscheid? Aanwijzingen voor verlatingsrituelen in ijzertijdnederzettingen. In H. Fokkens, & R. Jansen, *2000 jaar bewoningsdynamiek. Brons- en IJzertijd bewoning in het Maas-Demer-Scheldegebied* (pp. 45-61). Leiden: Leiden University Press.

van der Meer, W. (2014). *Archeobotanisch onderzoek van spiekers en kuilen op de vindplaats Nederweert-Hoebenakker (BRONSL-IJZV).* BIAX Consult.

VanValkenburgh, P., & Dufton, J. (2020). Big archaeology: horizons and blindspots. *Journal of Field Archaeology, 45*(1), 1-7. doi:10.1080/00934690.2020.1714307

Verschoof-van der Vaart, W. (2022). *Learning to look at LiDAR: Combining CNN-based object detection and GIS for archaeological prospection in remotely sensed data*. [Doctoral dissertation, Leiden University]. doi:10.5334/jcaa.32

Verschoof-van der Vaart, W. B., & Landauer, J. (2021). Using CarcassonNet to automatically detect and trace hollow roads in LiDAR data from the Netherlands. *Journal of Cultural Heritage, 47*(1), 143-154. doi:10.1016/j.culher.2020.10.009

Verschoof-van der Vaart, W. B., & Olivier, M. (2021). Implementing state-of-the-art Deep Learning approaches for archaeological object detection in remotely-sensed data: The results of cross-domain collaboration. *Journal of Computer Applications in Archaeology, 4*(1), 274-289. doi:10.5334/jcaa.78

Verschoof-van der Vaart, W. B., Lambers, K., Kowalczyk, W., & Bourgeois, Q. (2020). Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands. *ISPRS Int. J. Geo-Inf, 9*(5), 293. doi:10.3390/ijgi9050293

Verschoof-van der Vaart, W. B., Lambers, K., Kowalczyk, W., & 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), 293-294. doi:10.3390/ijgi9050293

Villes, A. (1985). Les greniers de l'habitat protohistorique en France septentrionale. *Les techniques de conservation des grains à long terme, 3*(2), 409-434.

Vokhmintcev, A., Khristodulo, O., Melnikov, A., & Romanov, M. (2023). Application of Dynamic Graph CNN* and FICP for Detection and Research Archaeology Sites. *International Conference on Analysis of Images, Social Networks and Texts* (pp. 294-308). Springer Nature Switzerland AG. doi:10.1007/978-3-031-54534-4_21

Weiss, E., Islev, M. E., & Hartmann, A. (2006). Autonomous cultivation before domestication. *Science, 312*(5780), 1608-1610.

Wesson, C. B., & Cottier, J. W. (2014). Big sites, big questions, big data, big problems: scales of investigation and changing perceptions of archaeological practice in the Southeastern United States. *Bulletin of the History of Archaeology, 24*(16), 1-11. doi:10.5334/bha.2416

Wheatley, D. (2004). Making space for an archaeology of place. *Internet Archaeology, 15*(1), 1-22.

Wheatley, D., & Gillings, M. (2003). *Spatial technology and archaeology.* New York: Taylor & Francis.

Whitely, D. S. (1998). *Reader in archaeological theory: post-processual and cognitive approaches.* New York: Routledge.

Winter, M. (2021). Beyond tomb and relic: Anthropological and pedagogical approaches to archaeogaming. *Near Eastern Archaeology, 84*(1), 12-21.

Zoph , B., Cubuk, E. D., Lin, T. Y., Shlens, J., & Le, Q. V. (2020). Learning data augmentation strategies for object detection. *Computer Vision–ECCV 2020: 16th European Conference* (pp. 566-583). Glasgow: Springer International Publishing.

Zubrow, E. B. (2005). Prehistoric space: an archaeological perspective. *Journal of World Anthropology, 2*(1), 1-42.

Zubrow, E. B. (2006). Digital archaeology: a historical context. In P. Daly, & T. L. Evans, *Digital archaeology: bridging method and theory* (pp. 10-31). New York: Routledge.

Appendices

Appendix 1: the specific code for the parameters of the three models

Model_1

```python
from ultralytics import YOLO

# Load the model
model = YOLO('yolov8s.pt')

# The parameters for model_1
model.train(
    data='dataset.yaml',
    epochs=50,
    batch=16,
    lr0=0.01,
    iou=0.4,
    conf=0.2,
    weight_decay=0.0001
)

# Save the model after training
model.save('model_1.pt')
```

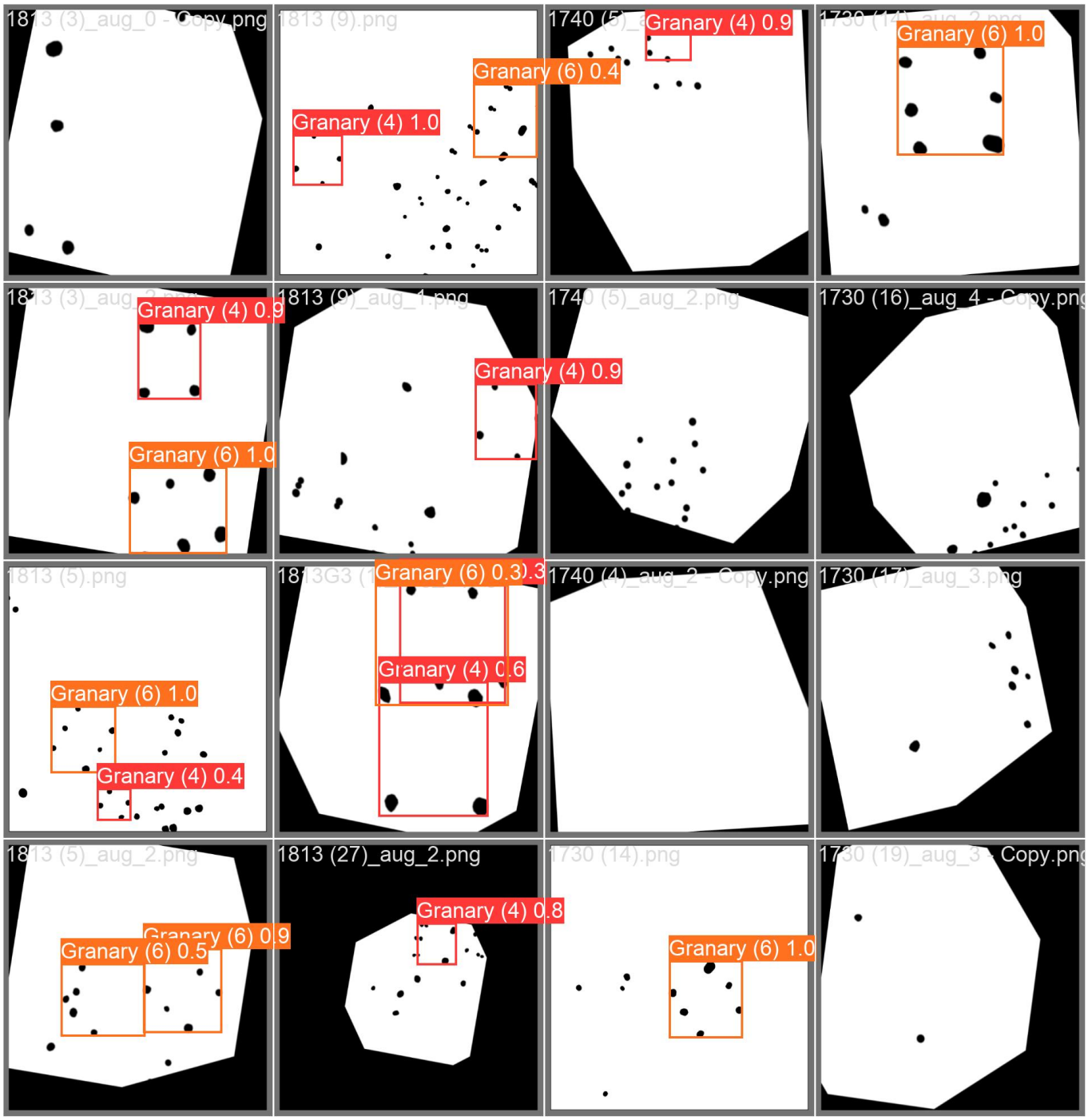
Model_2

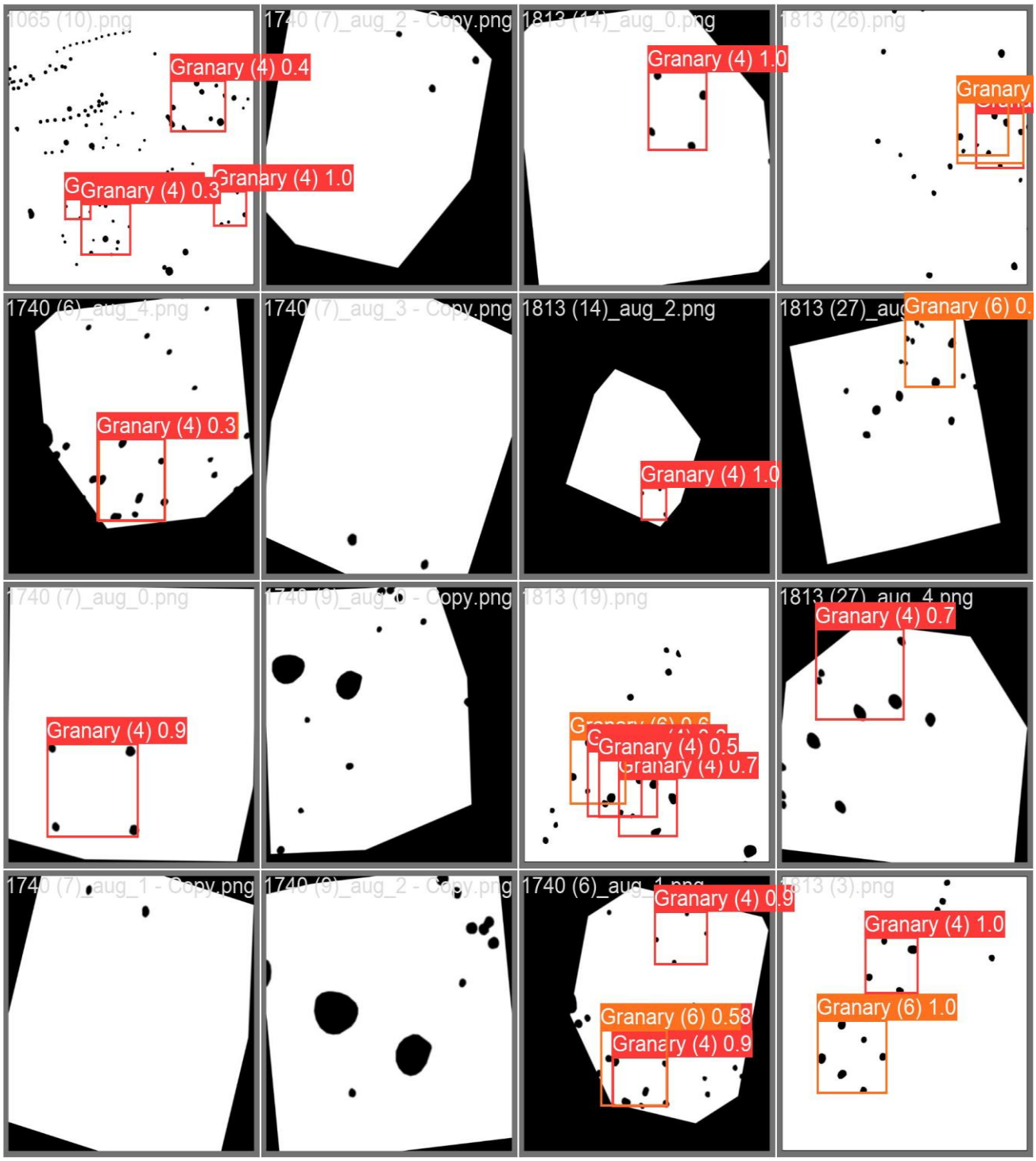
```python
from ultralytics import YOLO

# Load the model
model = YOLO('yolov8s.pt')

# The parameters for model_2
model.train(
    data='dataset.yaml',
    epochs=100,
    batch=32,
    lr0=0.005,
    iou=0.5,
    conf=0.3,
    weight_decay=0.0005
)

# Save the model after training
model.save('model_2.pt')
```

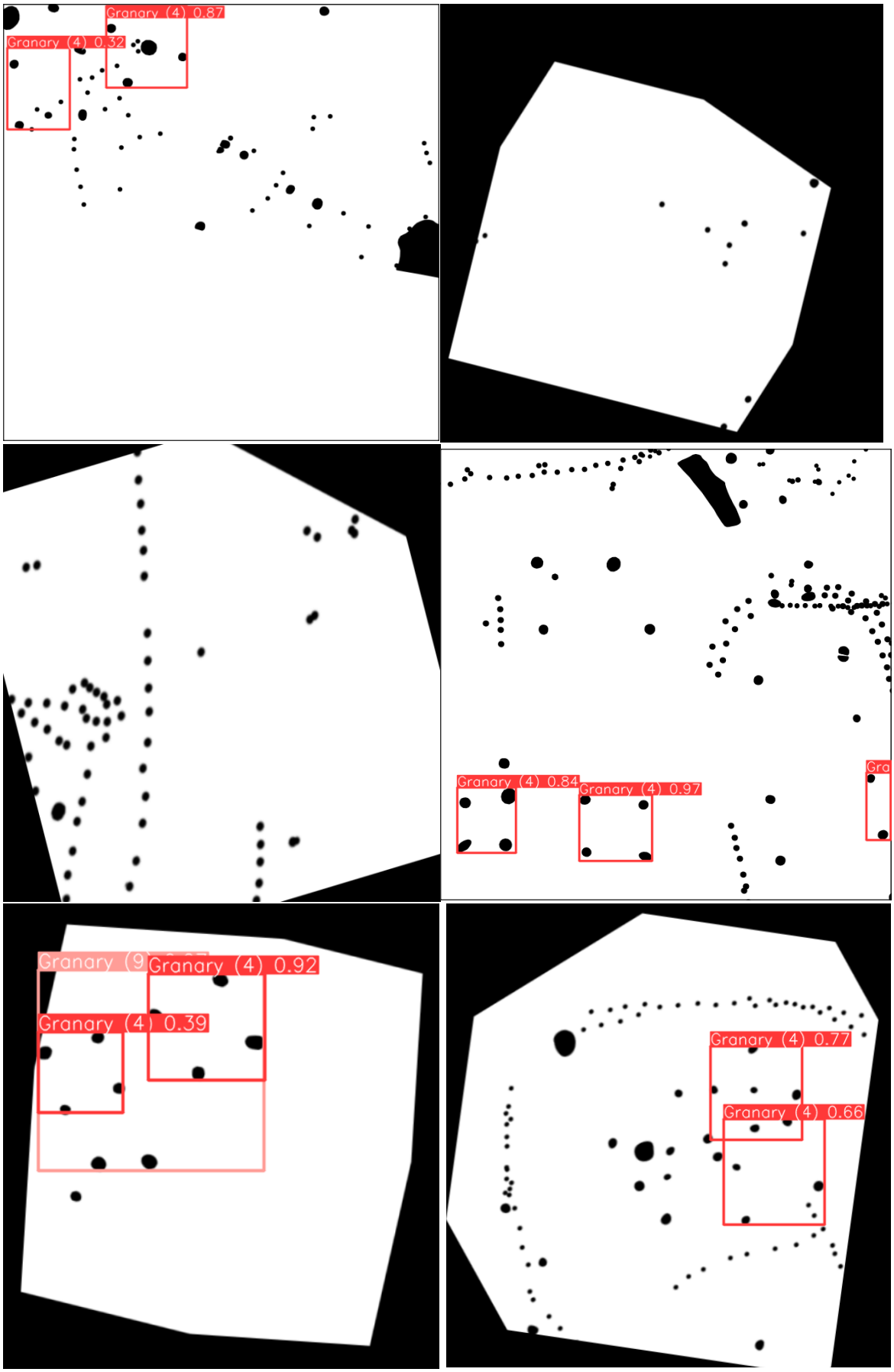
Model_3

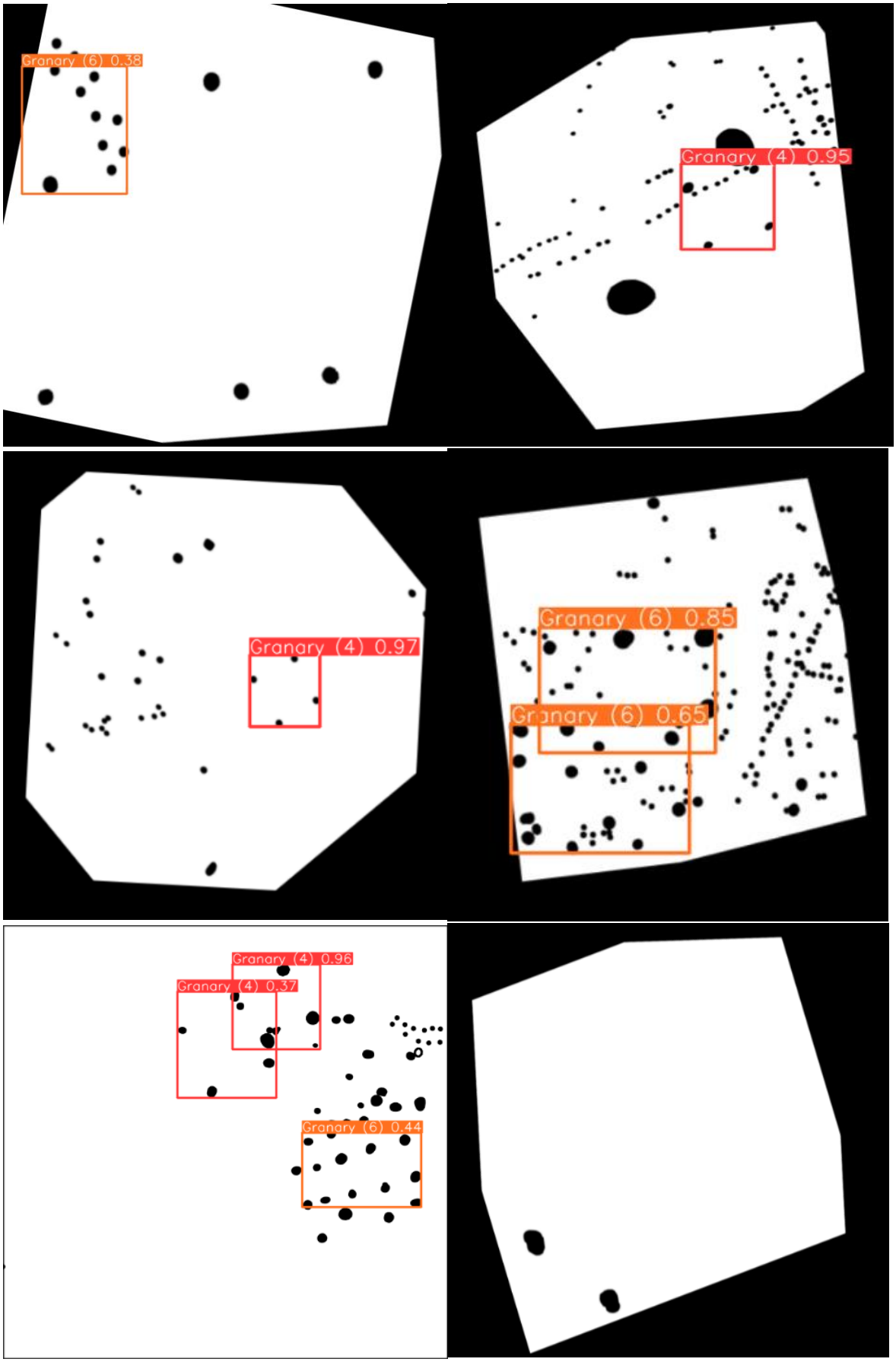
```python
from ultralytics import YOLO

# Load the model
model = YOLO('yolov8s.pt')

# The parameters for model_3
model.train(
    data='dataset.yaml',
    epochs=150,
    batch=32,
    lr0=0.001,
    iou=0.5,
    conf=0.4,
    weight_decay=0.0001
)

# Save the model after training
model.save('model_3.pt')
```

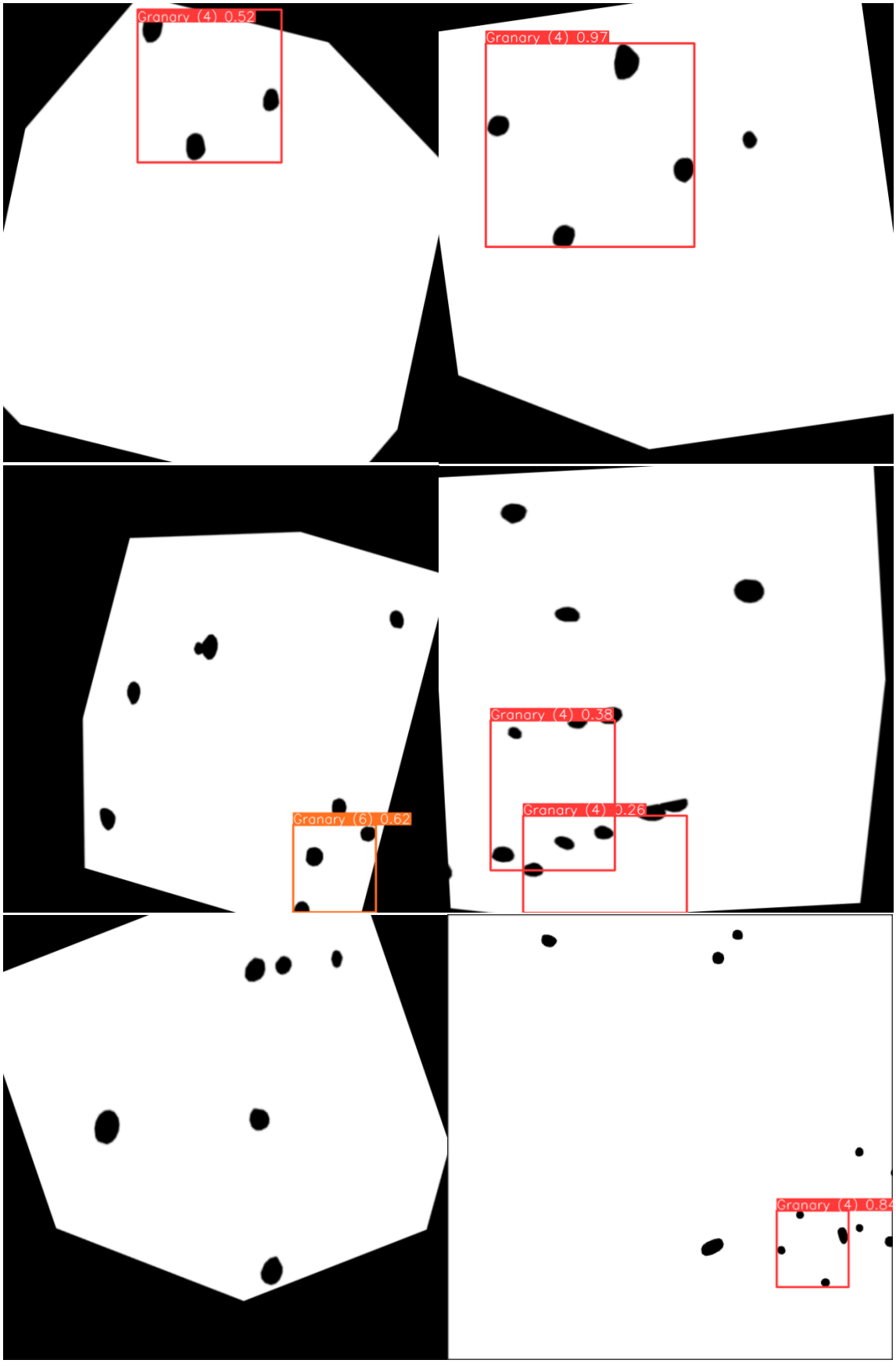
Appendix 2: Example predictions from model_3 on the test set

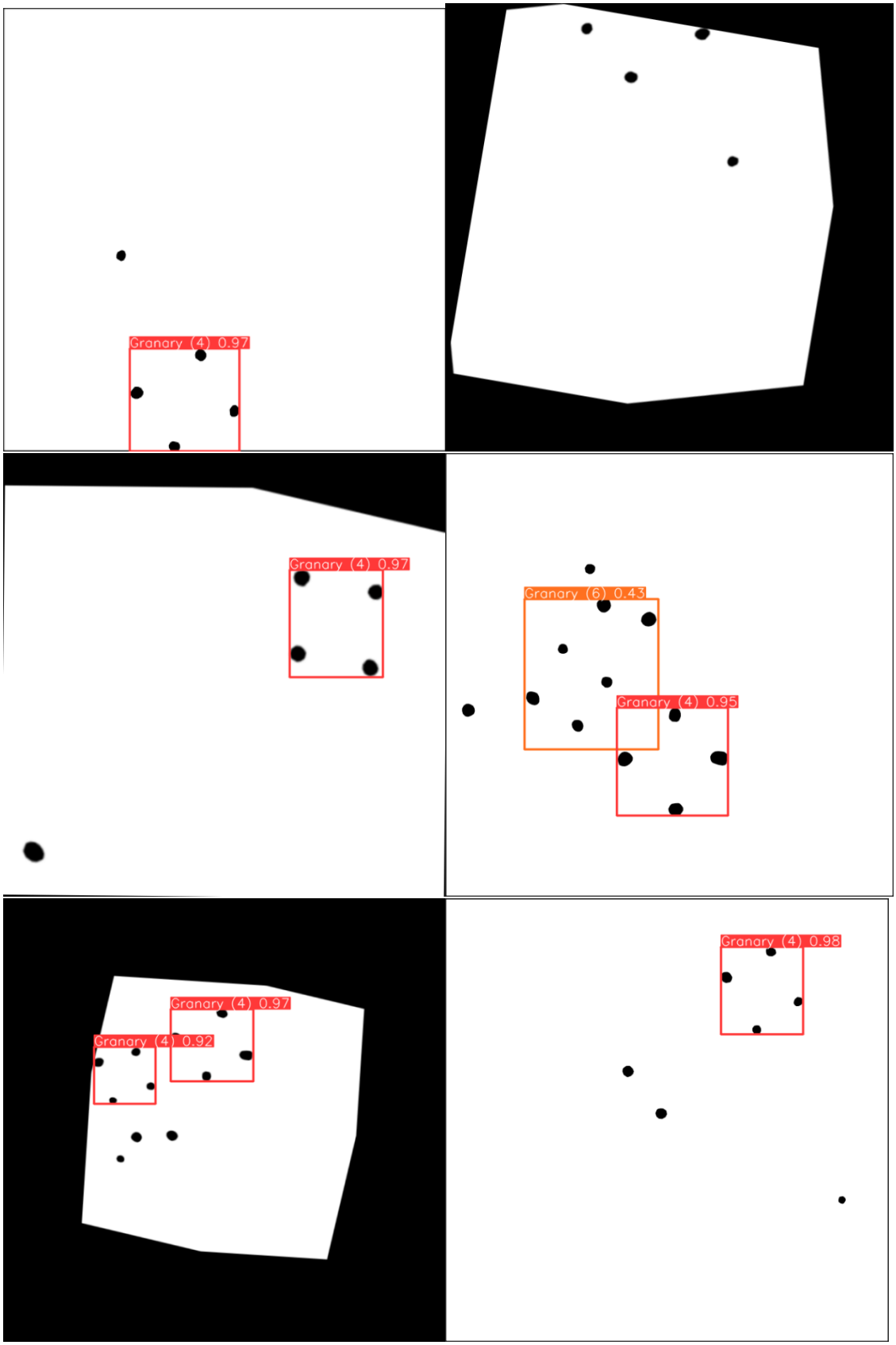


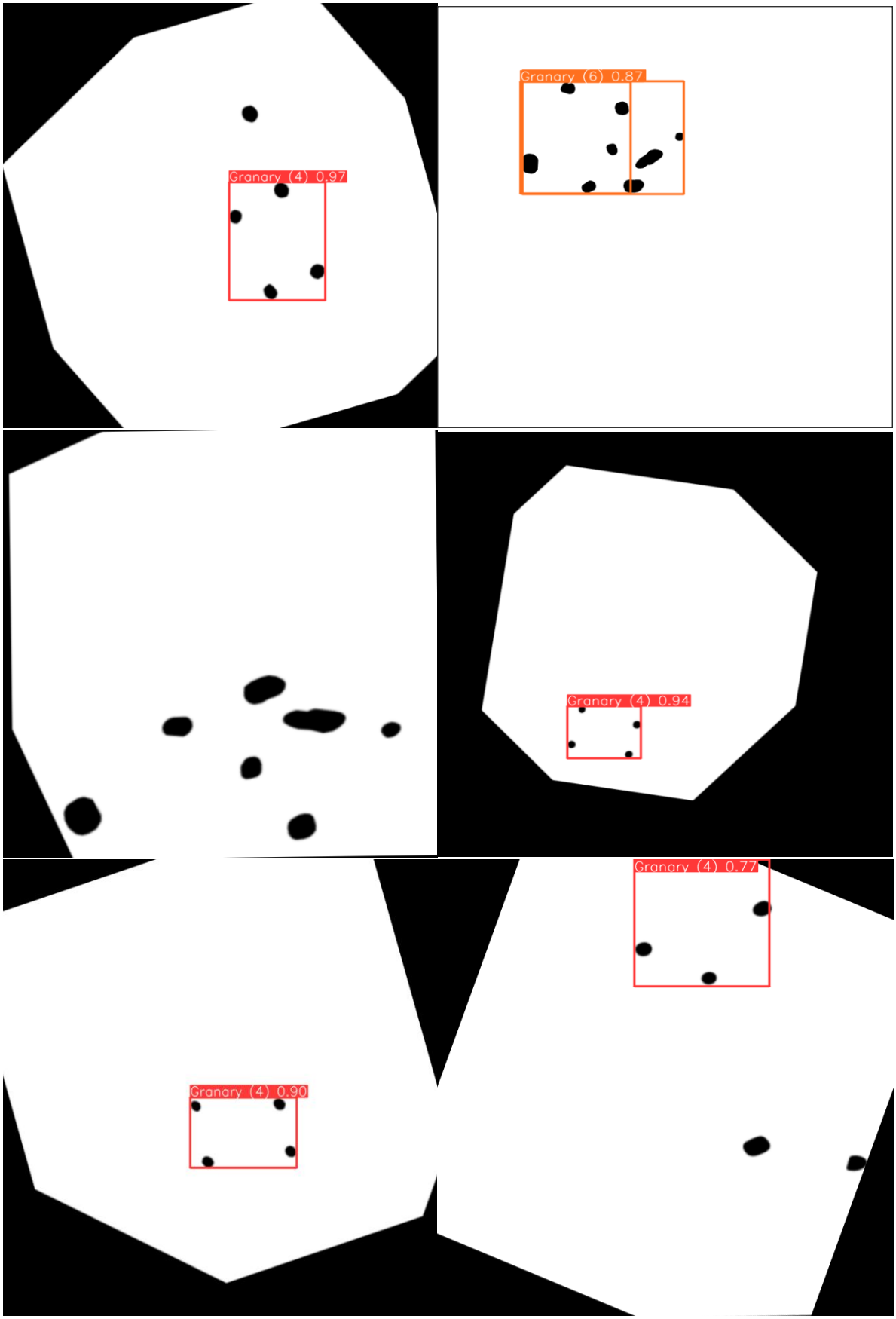

Granary (4) 0.4
Granary (4) 0.3
Granary (4) 1.0
Granary (4) 1.0
Granary
Granary (4) 0.3
Granary (4) 1.0
Granary (6) 0.
Granary (4) 0.9
Granary (4) 0.5
Granary (4) 0.7
Granary (4) 0.7
Granary (4) 0.9
Granary (6) 0.58
Granary (4) 0.9
Granary (4) 1.0
Granary (6) 1.0

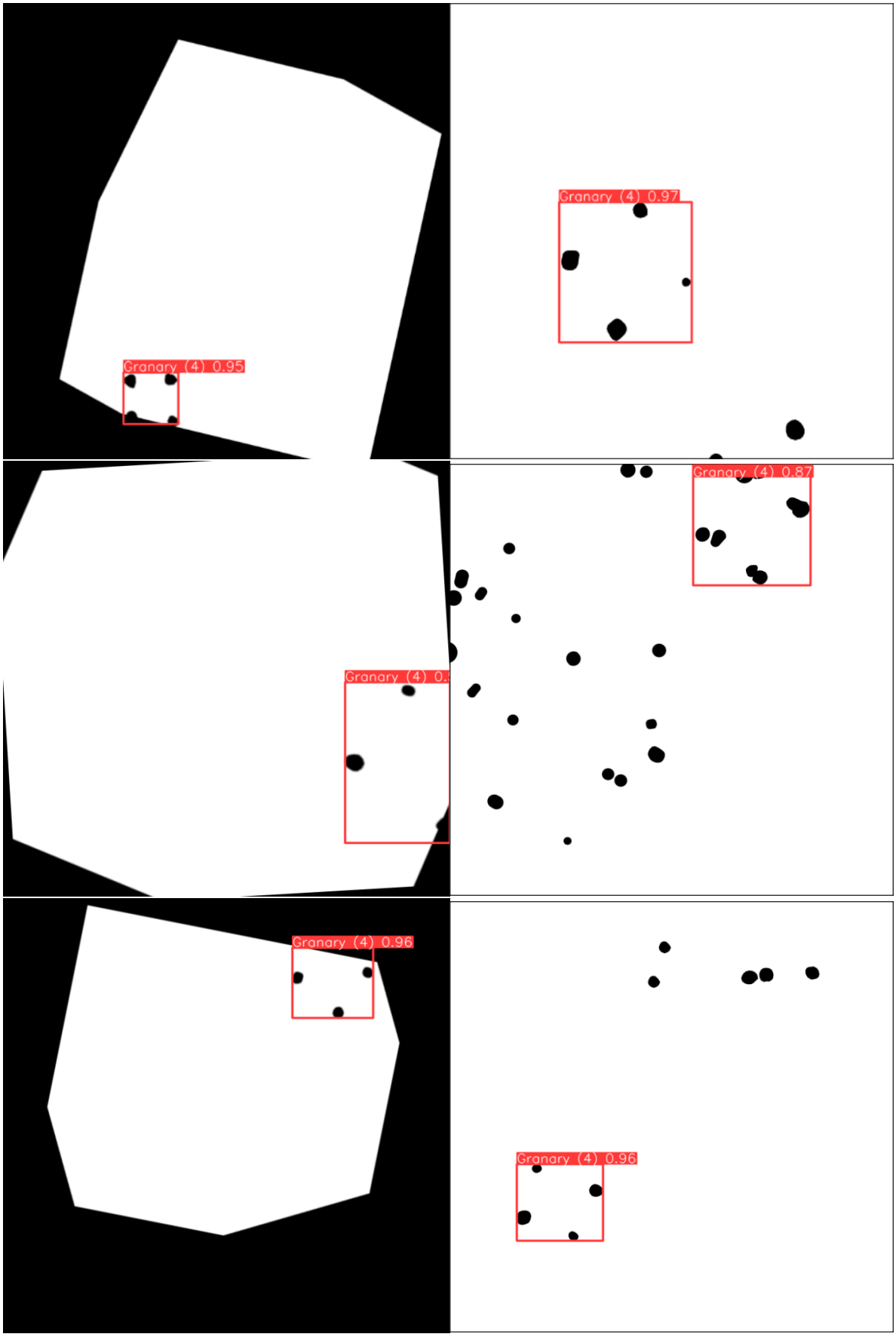

Granary (4) 0.87
Granary (4) 0.32
Granary (4) 0.84
Granary (4) 0.97
Granary (4) 0.92
Granary (4) 0.39
Granary (4) 0.77
Granary (4) 0.66

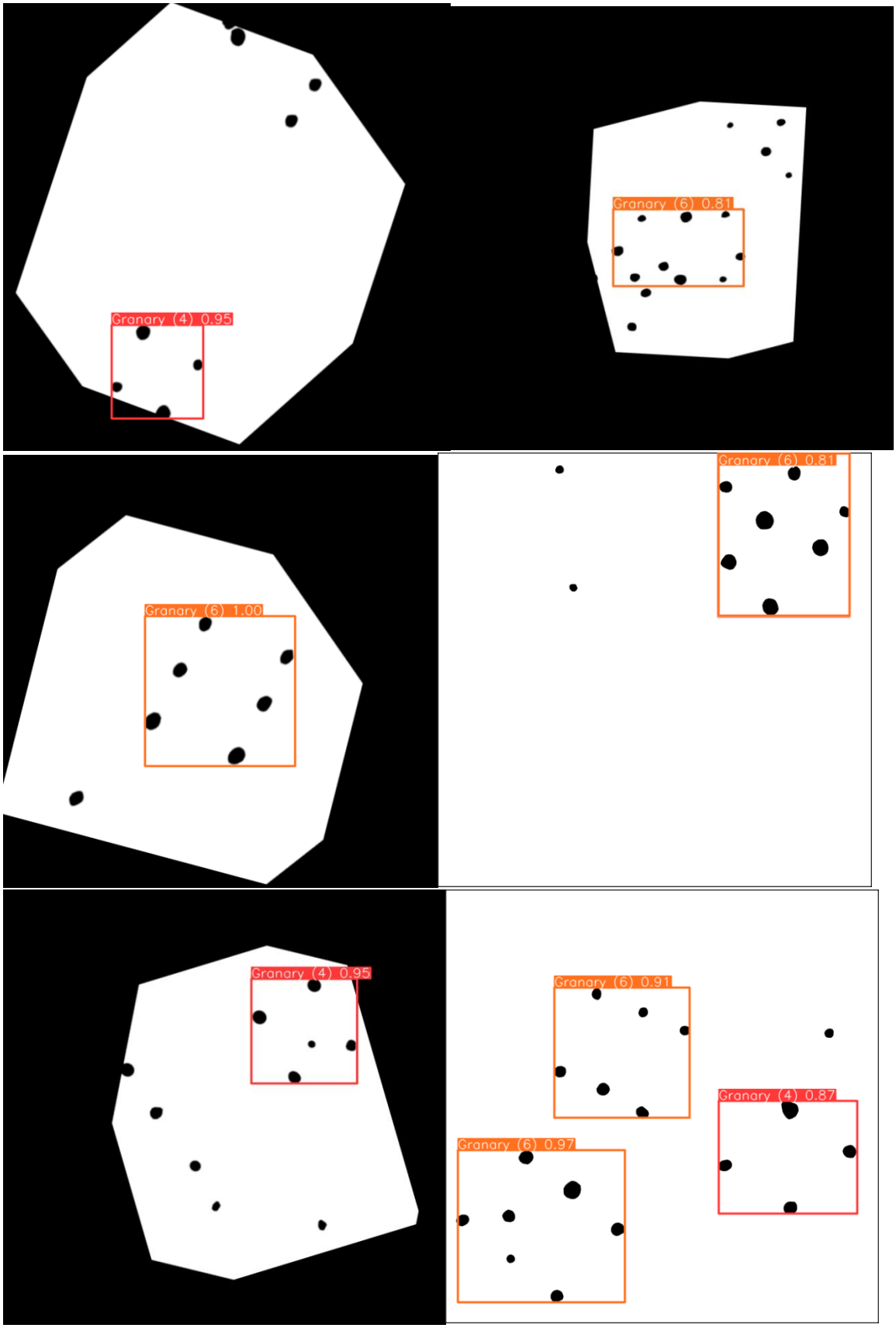

Granary (6) 0.38
Granary (4) 0.95
Granary (4) 0.97
Granary (6) 0.85
Granary (6) 0.65
Granary (4) 0.96
Granary (4) 0.37
Granary (6) 0.44

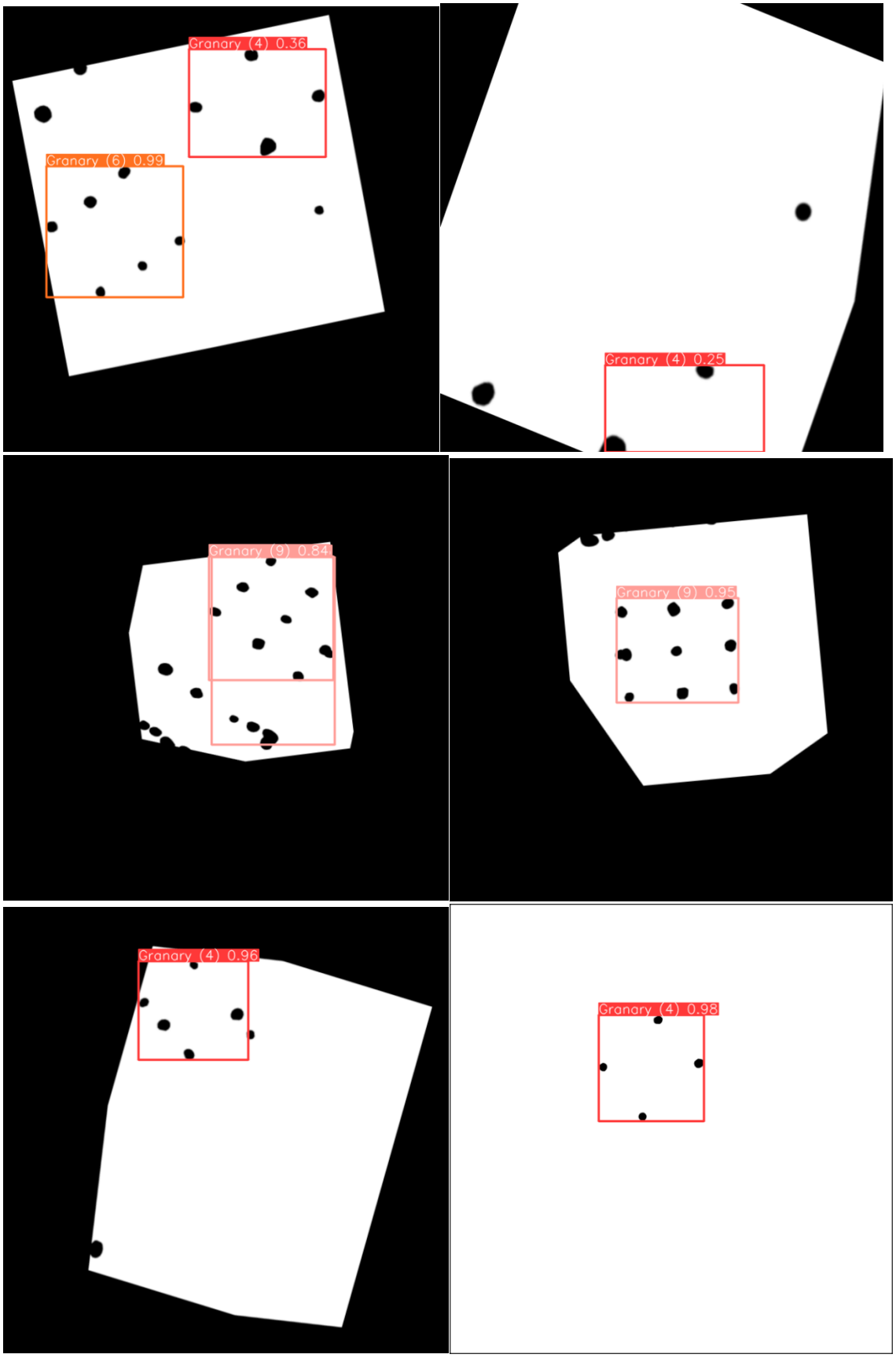

Granary (4) 0.52
Granary (4) 0.97
Granary (4) 0.38
Granary (6) 0.62
Granary (4) 0.26
Granary (4) 0.84

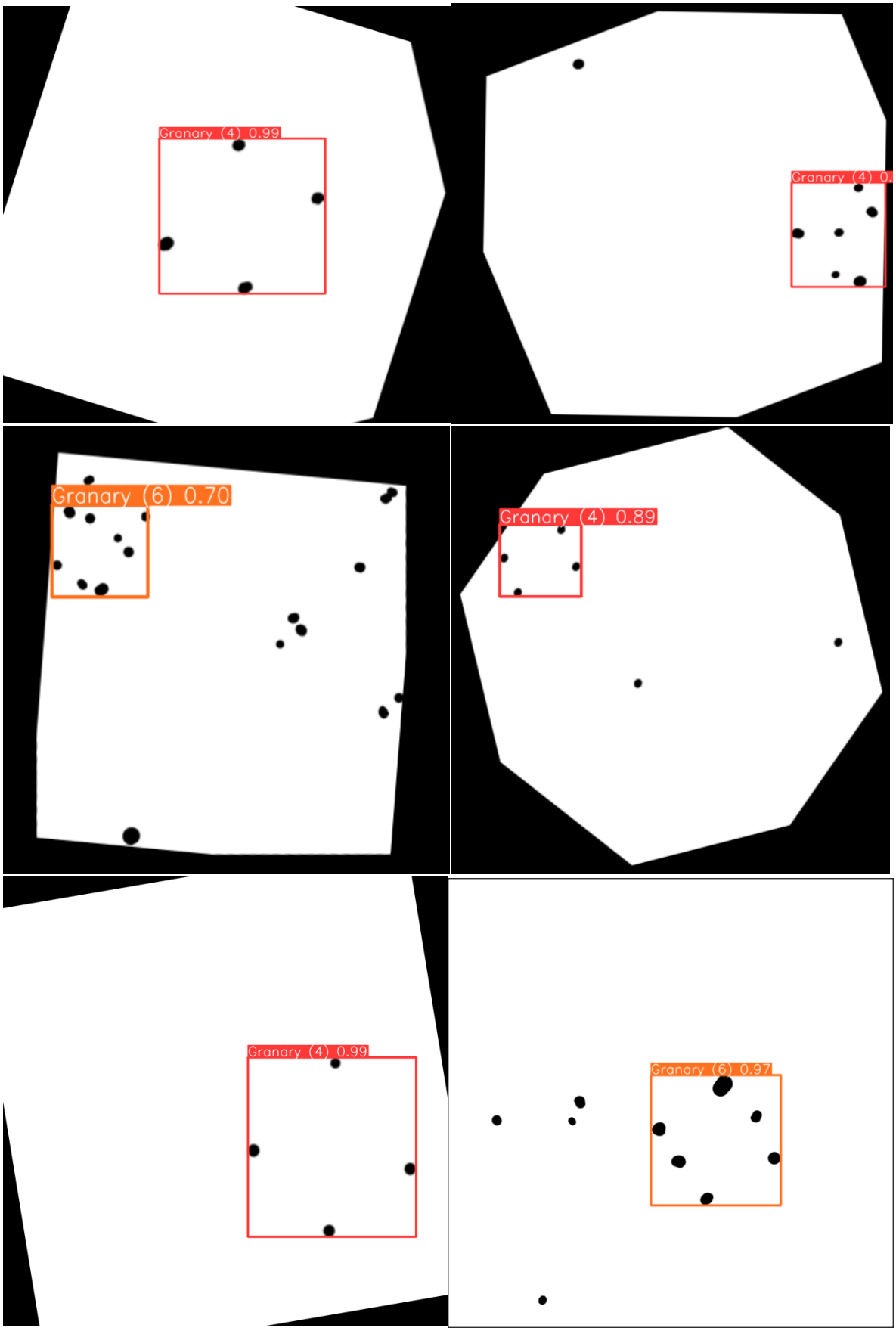

Granary (4) 0.97
Granary (4) 0.97
Granary (6) 0.43
Granary (4) 0.95
Granary (4) 0.97
Granary (4) 0.92
Granary (4) 0.98

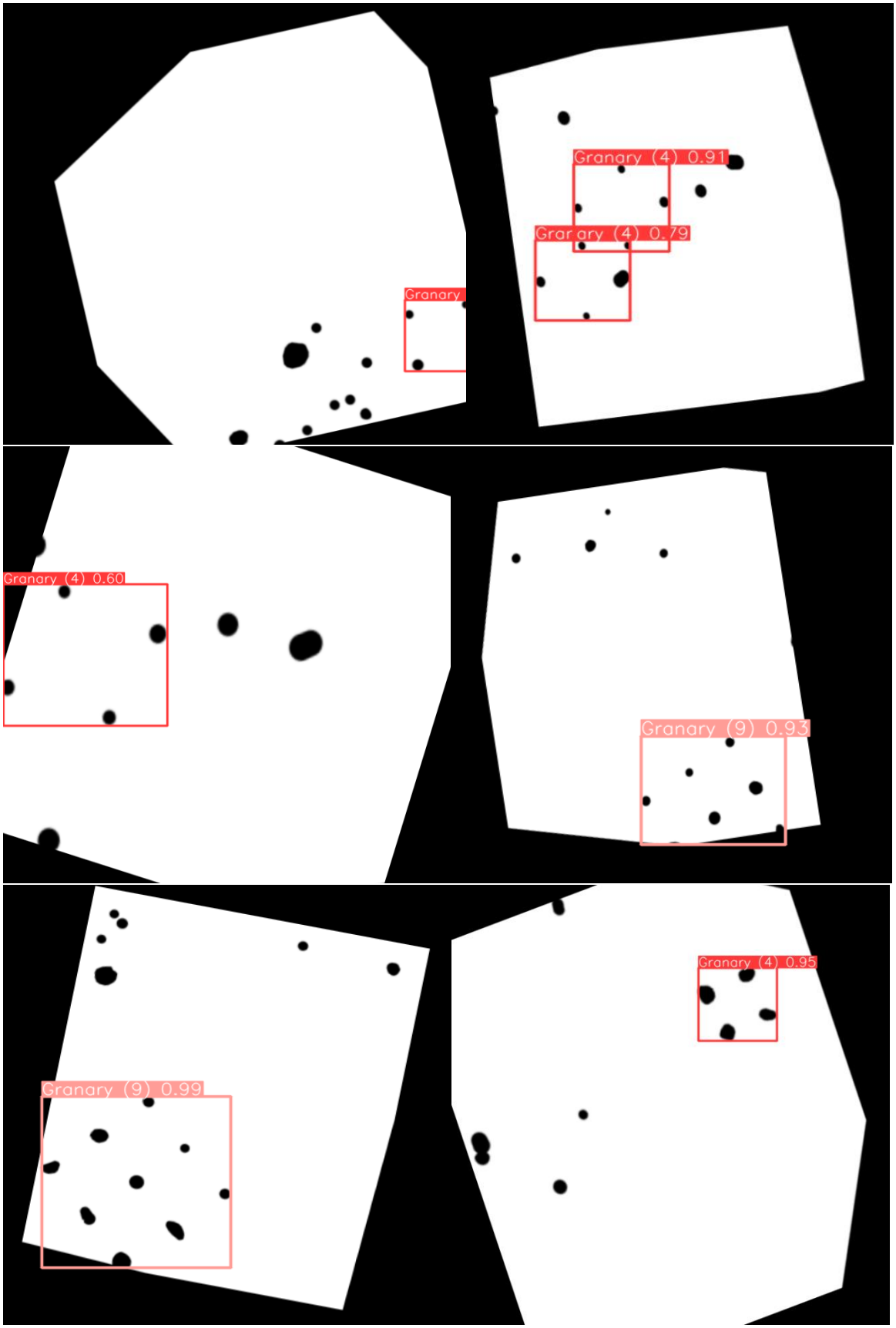

Granary (4) 0.97
Granary (6) 0.87
Granary (4) 0.94
Granary (4) 0.77
Granary (4) 0.90


Granary (4) 0.95
Granary (4) 0.97
Granary (4) 0.87
Granary (4) 0.96
Granary (4) 0.96

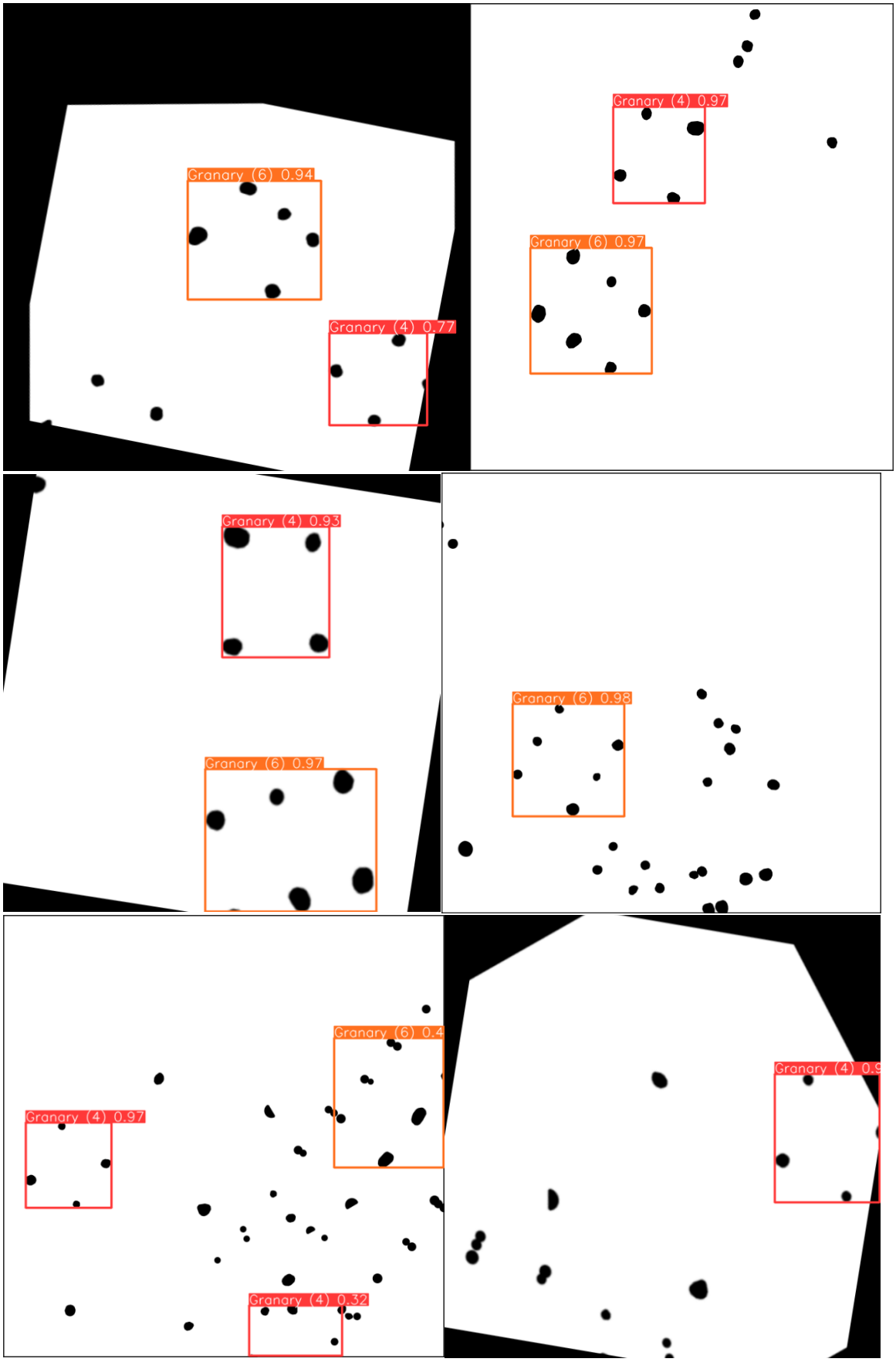

Granary (4) 0.95
Granary (6) 0.81
Granary (6) 1.00
Granary (6) 0.81
Granary (4) 0.95
Granary (6) 0.91
Granary (4) 0.87
Granary (6) 0.97

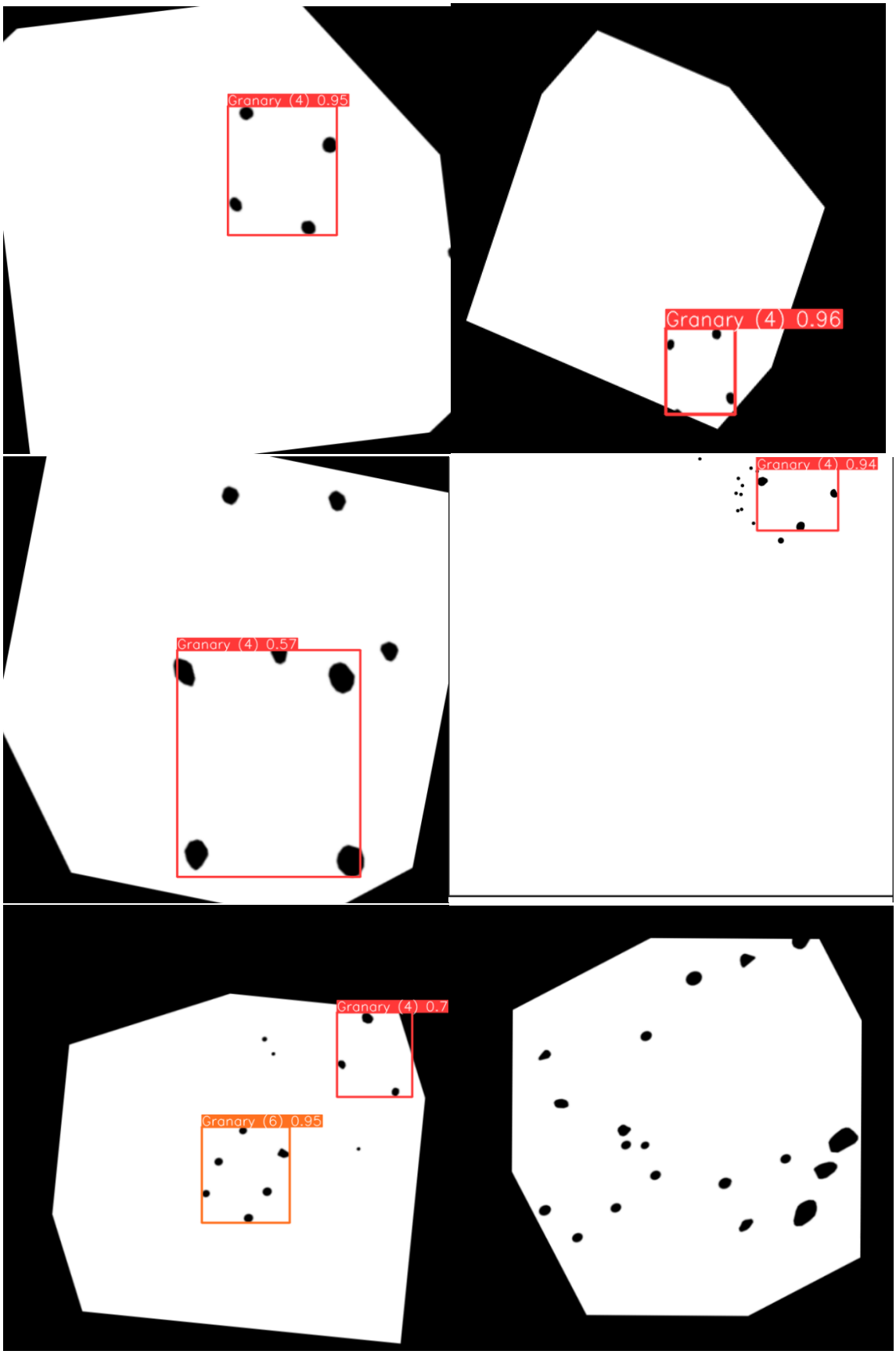

Granary (4) 0.36
Granary (6) 0.99
Granary (4) 0.25
Granary (9) 0.84
Granary (9) 0.95
Granary (4) 0.96
Granary (4) 0.98

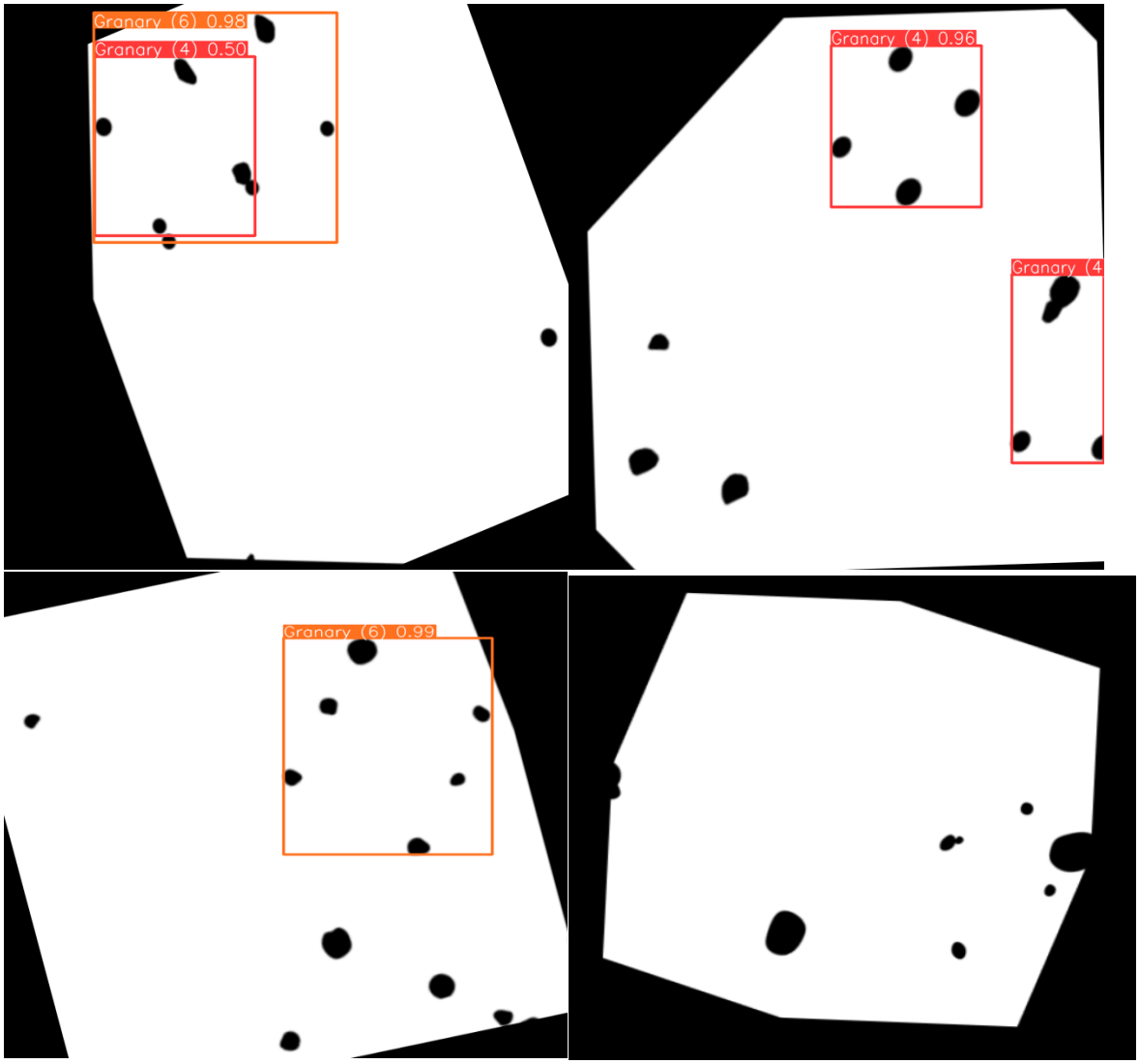

Granary (4) 0.99
Granary (4) 0.
Granary (6) 0.70
Granary (4) 0.89
Granary (4) 0.99
Granary (6) 0.97

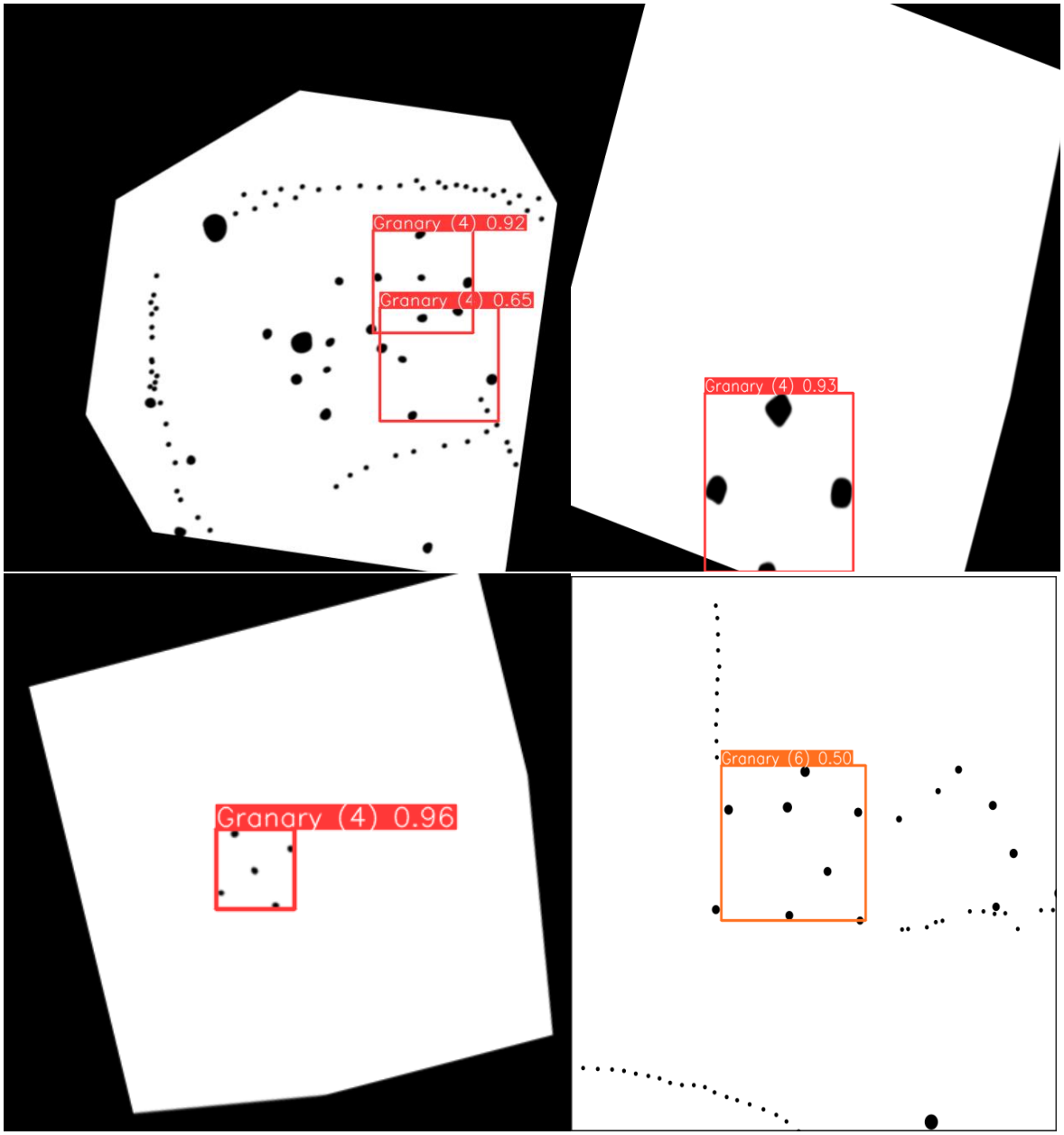

Granary (4) 0.91
Granary (4) 0.79
Granary
Granary (4) 0.60
Granary (9) 0.93
Granary (4) 0.95
Granary (9) 0.99


Granary (4) 0.88
Granary (4) 0.94
Granary (4) 0.78
Granary (4) 0.88
Granary (4) 0.34
Granary (6) 0.55
Granary (4) 0.90


Granary (6) 0.94
Granary (4) 0.77
Granary (4) 0.97
Granary (6) 0.97
Granary (4) 0.93
Granary (6) 0.97
Granary (6) 0.98
Granary (6) 0.4
Granary (4) 0.97
Granary (4) 0.32
Granary (4) 0.9


Granary (4) 0.95
Granary (4) 0.96
Granary (4) 0.57
Granary (4) 0.94
Granary (4) 0.7
Granary (6) 0.95


Granary (6) 0.98
Granary (4) 0.50
Granary (4) 0.96
Granary (4
Granary (6) 0.99


Granary (4) 0.92
Granary (4) 0.65
Granary (4) 0.93
Granary (4) 0.96
Granary (6) 0.50