

Connecting

THE RESEARCH PROMOTION FOUNDATION PROGRAMMES FOR RESEARCH, TECHNOLOGICAL DEVELOPMENT AND INNOVATION “RESTART 2016 – 2020”	
PILLAR	I. SMART GROWTH
PROGRAMME	RESEARCH INFRASTRUCTURES/SMALL SCALE INFRASTRUCTURES
RPF PROPOSAL NUMBER	SMALL SCALE INFRASTRUCTURES/1222/0062
PROJECT FULL TITLE	Research and Innovation Knowledge Centre for Engineering in Heritage
PROJECT ACRONYM	CONNECTING
Work Package	WP2 – Dissemination and Exploitation Activities
Deliverable	D11 – 2 Paper submission to international journals



Co-funded by
the European Union



Republic of Cyprus



RESEARCH
& INNOVATION
FOUNDATION



Cyprus
University of
Technology



getting you to solve better



University
of Cyprus



DEPARTMENT
OF ANTIQUITIES
CYPRUS

ACKNOWLEDGEMENT: This project has received funding from the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation of Cyprus under contract SMALL SCALE INFRASTRUCTURES/1222/0062. © Copyright by the CONNECTING consortium, 2023-2026. More info regarding the project you can find here: <https://connecting-infra.eu/>

Dissemination Level		
PU	Public	✓
CO	Confidential, only for members of the consortium (including the Agency Services)	

Document Sign-off		
Status of the document	Organization	Date
Drafted	CUT	20/04/2026
Reviewed	CONNECTING CONSORTIUM	20/05/2026
Approved	CONNECTING CONSORTIUM	25/05/2026
Submitted	CUT	31/05/2026

[illegible]

Review

Unequal Horizons: Global North–South Disparities in Archaeological Earth Observation (2000–2025)

Athos Agapiou 

Digital Cultural Heritage Unit, Department of Civil Engineering and Geomatics, Faculty of Engineering and Technology, Cyprus University of Technology, Saripolou 2-8, Limassol 3036, Cyprus; athos.agapiou@cut.ac.cy

Highlights

What are the main findings?

- Archaeological Earth observation research expanded nearly 1661% from 2000 to 2025, yet 78.9% of publications originate from Global North institutions, despite the Global South hosting the majority of UNESCO World Heritage Sites.
- A small group of countries (Italy, USA, UK) account for almost half of global out-put, while regions such as Sub-Saharan Africa remain critically un-der-represented (<1% of publications versus 9.4% of sites).

What is the implication of the main findings?

- This study demonstrates persistent structural inequalities in access to satellite data, computing infrastructure, and expertise, with implications for cultural her-itage protection and sustainable development.
- Redressing this imbalance requires coordinated capacity building, equitable data access, and inclusive frameworks that integrate Global South perspectives into Earth observation research.

Abstract

This systematic review analyzes 4359 archaeologically relevant publications spanning 25 years to examine global disparities in archaeological remote sensing research between Global North and Global South participation. This study reveals deep inequalities among these regions, with 72.1% of research output originating from Global North-only institutions, despite these regions hosting less than half of UNESCO World Heritage Sites. The temporal analysis demonstrates exponential growth, with 47.2% of all research published in the last five years, indicating rapid technological advancement concentrated in well-resourced institutions. Sub-Saharan Africa produces only 0.6% of research output while hosting 9.4% of World Heritage Sites, highlighting a technology gap in heritage protection. The findings suggest an urgent need for coordinated interventions to address structural inequalities and promote technological fairness in global heritage preservation. The research employed bibliometric analysis of Scopus database records from four complementary search strategies, revealing that just three countries—Italy (20.3%), the United States (16.7%), and the United Kingdom (10.0%)—account for nearly half of all archaeological remote sensing research and applications worldwide. This study documents patterns that have profound implications for cultural heritage preservation and sustainable development in an increasingly digital world where advanced Earth observation technologies have become essential for effective heritage protection and archaeological research.



Academic Editors: Fulong Chen and Massimiliano Pepe

Received: 30 August 2025

Revised: 3 October 2025

Accepted: 5 October 2025

Published: 6 October 2025

Citation: Agapiou, A. Unequal Horizons: Global North–South Disparities in Archaeological Earth Observation (2000–2025). *Remote Sens.* **2025**, *17*, 3371. <https://doi.org/10.3390/rs17193371>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: earth observation; remote sensing archaeology; satellite sensors; global survey; literature review; Scopus; UNESCO World Heritage Sites

1. Introduction

The integration of satellite-based Earth observation technologies into archaeological research has fundamentally transformed the discipline over the past quarter of a century, offering new capabilities for site discovery, landscape analysis, and cultural heritage monitoring across vast spatial and temporal scales [1–3]. This technological revolution emerged from pioneering work in the 1970s and 1980s using early satellite platforms such as Landsat [4,5] and SPOT [6,7] imagery, evolving into a sophisticated interdisciplinary domain that combines remote sensing technologies, geospatial analysis, artificial intelligence, and archaeological interpretation to advance our understanding of human settlement patterns, landscape modification, and cultural heritage preservation [8–12].

The scientific field has been driven by several factors including the increasing availability and declining costs of high-resolution satellite imagery, advances in computational processing capabilities, the development of sophisticated analytical software packages, and growing recognition within the archaeological community of remote sensing's potential for addressing fundamental research questions about past human behavior and cultural change [13–15]. The democratization of Geographic Information System (GIS) technology and the emergence of cloud-based processing platforms have further accelerated adoption by reducing technical barriers and computational requirements that previously limited access to advanced analytical capabilities [16–18].

However, despite the global nature of archaeological heritage and the wide applicability of satellite technologies, significant questions remain about the equitable distribution of these technological capabilities and their benefits across different regions of the world [19,20]. The distinction in archaeological remote sensing research represents more than a simple geographical categorization [21]; it reflects deeper structural inequalities in access to advanced technologies, funding resources, institutional capacity, research infrastructure, and the technological required for independent scientific innovation [19,21,22].

In this study, the term Global South is employed in line with Mahler's [23] definition, referring not to a strictly geographic designation but to a socio-political construct encompassing regions historically marginalized in global economic, scientific, and technological systems (e.g., much of Africa, Latin America, the Middle East, and parts of Asia). This conceptualization allows us to capture disparities in access to resources, research capacity, and technological infrastructures when comparing the Global South with the more industrialized and research-intensive Global North. The Global North, encompassing primarily developed nations in North America (United States and Canada), Europe (including all European Union member states, United Kingdom, Norway, Switzerland, Iceland, and other Western European nations), and developed Asia-Pacific regions (Japan, South Korea, Australia, and New Zealand), has historically dominated not only the development and deployment of Earth observation satellites but also the creation of analytical frameworks, artificial intelligence models, and institutional knowledge that define the field's methodological standards and research priorities [24,25]. This is visible across multiple dimensions including the ownership and operation of satellite systems by agencies such as NASA, ESA, JAXA, and commercial providers like DigitalGlobe, Planet Labs, and Maxar Technologies, the development of processing software and analytical algorithms by companies like Google, Microsoft, ESRI, and specialized remote sensing firms, the availability of high-performance computing infrastructure in universities and research institutions with

substantial computational resources, access to commercial high-resolution imagery through institutional licenses and research agreements that provide preferential pricing and data access, and the concentration of specialized expertise in academic programs, professional networks, and research centers with decades of accumulated experience [26–28].

Conversely, the Global South, representing developing nations primarily in Africa (including all Sub-Saharan African countries such as Kenya, Tanzania, South Africa, Ghana, Nigeria, Senegal, and North African countries like Egypt, Morocco, Tunisia, Algeria), Latin America and the Caribbean (including Brazil, Mexico, Argentina, Peru, Colombia, Chile, and Caribbean island nations), and developing Asia (including major populous nations like China, India, Indonesia, Pakistan, Bangladesh, Thailand, Vietnam, Philippines, and Middle Eastern countries such as Iran, Iraq, Turkey, Syria, Jordan, and Lebanon) often hosts some of the world's most significant archaeological sites and cultural heritage resources [29,30]. These include numerous UNESCO World Heritage Sites of global importance such as Angkor Wat in Cambodia, Machu Picchu in Peru, the Pyramids of Giza in Egypt, Great Zimbabwe in Zimbabwe, Petra in Jordan, Borobudur in Indonesia, and countless other irreplaceable cultural landscapes and archaeological complexes that represent humanity's shared heritage. As of 2024, there are 1223 UNESCO World Heritage Sites globally, with 952 cultural, 231 natural, and 40 mixed properties distributed across 168 countries [31]. Yet these regions face substantial and various barriers to accessing and effectively utilizing the advanced Earth observation technologies that have become essential for contemporary archaeological research and heritage protection in the digital age [32–35].

These barriers are also linked with technological, economic, educational, and institutional dimensions, creating a complex web of interconnected constraints that collectively limit meaningful participation in cutting-edge archaeological research and heritage monitoring activities. The technological barriers are particularly pronounced and include severely limited access to high-speed internet connectivity required for downloading and processing large satellite datasets that can range from gigabytes to terabytes in size, insufficient computational infrastructure for processing high-resolution imagery and running sophisticated analytical algorithms that require substantial processing power and memory resources, a systematic lack of access to commercial satellite imagery due to prohibitive cost constraints and restrictive licensing agreements that favor institutional buyers in developed countries, and the critically limited availability of specialized software packages and analytical tools that require expensive licenses and ongoing technical support [10,15,25,36].

Therefore, economic constraints extend far beyond the direct costs of satellite imagery and software licenses to encompass the broader financial ecosystem required for supporting long-term research programs, including substantial investments in hiring and training technical staff with specialized skills in remote sensing and archaeological interpretation, maintaining and upgrading computational infrastructure to keep pace with rapidly evolving technological requirements, participating in international conferences and collaborative networks that facilitate knowledge exchange and professional development, and sustaining research programs over multi-year periods required for meaningful archaeological investigation and heritage monitoring initiatives [37–41].

Educational and institutional barriers include the limited availability of specialized training programs in archaeological remote sensing, the shortage of faculty with expertise in Earth observation technologies, the lack of established curricula integrating remote sensing with traditional archaeological methods, and insufficient institutional support for developing interdisciplinary research capabilities [42]. These educational gaps are particularly problematic because effective use of Earth observation technologies requires not only technical skills in image processing and analysis but also a deep understanding of

archaeological theory, site formation processes, and landscape evolution that can only be developed through comprehensive interdisciplinary training programs.

Moreover, the recent developments of artificial intelligence and machine learning in archaeological remote sensing have introduced new and increasingly complex dimensions to these existing disparities, potentially either serving as a democratizing force that could provide powerful analytical tools to under-resourced institutions, or alternatively functioning to further concentrate advanced capabilities in already well-resourced institutions with the technical expertise and computational infrastructure necessary to develop and deploy these sophisticated systems [43–47]. Machine learning algorithms, particularly deep learning models and convolutional neural networks used for automated feature detection, classification, and pattern recognition in satellite imagery, are fundamentally dependent on the representativeness, quality, and cultural appropriateness of their training datasets, which must encompass diverse archaeological contexts, environmental conditions, and cultural landscapes to function effectively across different geographic and cultural settings [48–50].

The development and deployment of AI systems [51] for archaeological applications [52] has been overwhelmingly dominated by institutions with substantial computational resources, advanced technical expertise, and access to high-performance computing infrastructure, primarily located in technologically advanced countries with well-established computer science departments, remote sensing research programs, and collaborative networks linking academic institutions with technology companies and government agencies [53,54]. This concentration reflects the substantial barriers to entry for meaningful AI development and deployment, including stringent requirements for GPU-accelerated computing systems capable of training complex neural networks, access to large-scale, high-quality training datasets that require extensive curation and validation, specialized expertise in machine learning, computer vision, and archaeological interpretation that combines technical skills with domain knowledge, and sustained funding for long-term research and development activities that can extend over multiple years before producing practical applications. The result is a technological landscape where the most powerful analytical tools for archaeological remote sensing are developed by and for research contexts that may not be representative of global archaeological diversity.

This review was designed to systematically identify, analyze, and synthesize the available literature on Earth observation applications in archaeology, with a particular focus on identifying geographical disparities and technological access patterns between these two geographical contexts.

2. Methodology

The methodology incorporated both quantitative bibliometric analysis and qualitative content synthesis to provide a comprehensive assessment of the field's development, geographic distribution, and technological evolution over the 25-year study period from 2000 to 2025 [55]. The year 2000 was selected as the starting point, as it marks the advent of high-resolution commercial satellite sensors (e.g., IKONOS), which enabled a new era of archaeological and heritage-related Earth observation research [9]. This mixed-methods approach was selected to capture broad patterns in research output and publication trends as well as nuanced insights into technological applications, barriers to access, and collaborative relationships that might not be apparent through purely quantitative analysis [56].

Scopus was selected as the sole source for bibliometric data because it provides the most comprehensive coverage of peer-reviewed journals in the fields of Earth observation, remote sensing, and archaeology, while also offering standardized metadata (including author affiliations, keywords, and DOIs) necessary for reliable comparative analysis [9,57].

The database provides robust search capabilities with advanced Boolean operators, field-specific searching, and sophisticated filtering options that make it particularly suitable for complex interdisciplinary queries spanning multiple academic domains and publication types [58]. In contrast, Web of Science offers narrower journal coverage in the humanities and applied sciences, and Google Scholar—while broader—lacks the metadata quality, transparency, and deduplication mechanisms required for rigorous bibliometric work.

Four complementary search queries were developed to capture different aspects of the research domain, following established systematic review protocols for search strategy development and validation [59,60]. This process was based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology review [61], ensuring in this way the transparency, reproducibility, and completeness of the survey. This survey employed a keyword-based approach rather than field-of-research classifications, since the remote sensing literature is fragmented, as noted in previous reviews [3,9]. The use of Earth observation technologies in archaeology is evident across multiple disciplines and areas of expertise, reflecting its multidisciplinary nature.

The first query focused on core remote sensing applications in archaeology, combining terms related to satellite and aerial remote sensing technologies with archaeological and cultural heritage applications. The second query specifically targeted artificial intelligence and machine learning applications in archaeological remote sensing, reflecting the growing importance of these technologies in the field. The third query explicitly searched for studies addressing geographic disparities, digital divides, or bias in remote sensing applications. The fourth query captured broader applications in archaeological survey and heritage management contexts (see Table 1).

Table 1. Search strategy performance and literature coverage.

Query Focus	Records	Percentage	Coverage	Research Emphasis
Core Remote Sensing ^{Q1}	4573	92.2%	Comprehensive	Traditional methods
AI/ML Applications ^{Q2}	272	5.5%	Advanced techniques	Computational methods
Geographic Disparities ^{Q3}	44	0.9%	Critical perspectives	Social analysis
Heritage Management ^{Q4}	72	1.4%	Practical applications	Implementation focus
Combined Total	4961	100%		

^{Q1} TITLE-ABS-KEY (("remote sensing" OR "satellite imagery" OR "earth observation" OR "space archaeology" OR "aerial photography" OR "LIDAR" OR "hyperspectral") AND (archaeology OR archaeological OR "cultural heritage" OR "heritage monitoring" OR "site detection" OR "landscape archaeology")). ^{Q2} TITLE-ABS-KEY (("artificial intelligence" OR "machine learning" OR "deep learning" OR "convolutional neural network" OR "automated detection" OR "computer vision") AND ("remote sensing" OR "satellite imagery" OR "earth observation") AND (archaeology OR archaeological OR "cultural heritage")). ^{Q3} TITLE-ABS-KEY (("remote sensing" OR "satellite archaeology" OR "earth observation") AND (archaeology OR archaeological) AND ("global south" OR "developing countries" OR "global north" OR "developed countries" OR "digital divide" OR bias OR disparity OR inequality)). ^{Q4} TITLE-ABS-KEY (("satellite remote sensing" OR "space-based" OR "earth observation") AND ("archaeological survey" OR "archaeological prospection" OR "heritage management" OR "cultural landscape")).

The search was conducted in August 2025, using the Scopus Advanced Search interface. Query 1, focusing on core remote sensing archaeology applications, retrieved 5124 records, representing the largest subset of the literature. Query 2, targeting AI and machine learning applications, identified 317 records, reflecting the more specialized nature of these technological approaches. Query 3, explicitly addressing geographic and equity issues, yielded only 48 records, suggesting limited scholarly attention to these concerns within the published literature. Query 4, capturing broader archaeological remote sensing applications, identified 85 additional records, bringing the total initial retrieval to 5574 records across all four search strategies.

As the search was limited to the period 2000–2025 for papers published as journal research articles, conference papers, and review papers, further papers from the initial query have been excluded (4573 papers from Query 1; 272 papers from Query 2; 44 papers from Query 3; and 72 papers from Query 4). The remaining papers totaled 4961 records.

Given the overlapping nature of the search queries, systematic deduplication was essential to ensure accurate analysis. The deduplication process followed established protocols using Electronic Identifier (EID) matching as the primary method, as Scopus assigns unique EIDs to each indexed publication regardless of variations in citation format. The identification of 371 duplicate records across the four queries, representing a 7.5% duplication rate, indicated that the four search strategies successfully captured distinct but complementary aspects of archaeological remote sensing research. From the remaining papers, 231 papers were manually excluded as their topic was not aligned with the scope of this study. Screening was performed by combining the “Subject Area” filter in Scopus to remove non-relevant material with a manual review of titles and abstracts conducted by the author. Therefore, the final dataset comprising 4359 archaeologically relevant publications was further screened and processed by the author. The dataset encompasses research spanning from traditional satellite imagery interpretation to cutting-edge artificial intelligence applications, confirming the comprehensive nature of the search strategy in capturing genuine archaeological applications of Earth observation technologies across diverse geographic and temporal contexts.

The above dataset was also compared and discussed against the overall growth of scientific papers in the domains of remote sensing and earth observation. Therefore, a similar approach (query) was followed again in the Scopus engine, using the terminology “satellite remote sensing” or “earth observation”, limiting our findings to the period between 2000 and 2025 and only for papers published as journal research articles, conference papers, or review papers. In total, 173,474 papers were published within this period.

Study locations and author institutional affiliations were classified according to established United Nations geographic groupings supplemented by World Bank income classifications to capture both the geographic and economic dimensions of global inequality. As mentioned earlier, the Global North classification includes North America (United States and Canada), Europe (including European Union member states, the United Kingdom, Norway, Switzerland, and Iceland), and developed Asia-Pacific regions (Japan, South Korea, Australia, and New Zealand). The Global South classification encompasses Africa, Latin America and the Caribbean, and developing Asia (including China, India, Southeast Asia, Central Asia, and the Middle East). Moreover, information related to the UNESCO World Heritage Sites has been retrieved, as shown in Figure 1.

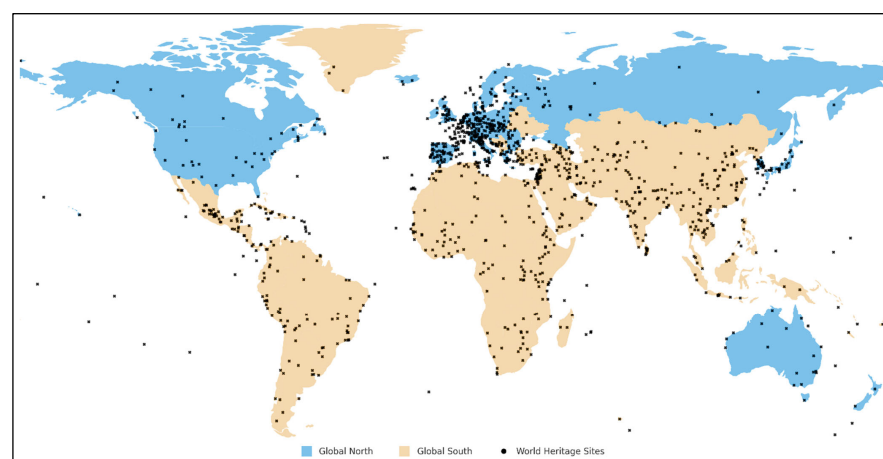


Figure 1. Global North/South classification per this study.

3. Results

3.1. Temporal Distribution and Research Evolution Patterns

Analysis of publication years within the archaeologically relevant dataset reveals significant temporal trends that demonstrate the explosive growth of archaeological remote sensing research, particularly in the most recent decade. The temporal distribution demonstrates unprecedented acceleration in research output, with over 47% of all archaeological remote sensing research published in just the last five years of the study period. Table 2 compares the temporal distribution and growth of publications in archaeological Earth observation (EO) with those in Earth observation overall between 2000 and 2025 for the top 10 countries. Archaeological EO shows a sharp acceleration, from just 117 publications (2.7%) in 2000–2004 to 2059 (47.2%) in 2020–2025, marking a +3626% growth rate across the period. In contrast, Earth observation overall grew from 13,551 publications (7.8%) to 72,454 (41.8%), corresponding to +1180% growth. Despite its smaller absolute scale, Archaeological EO demonstrates a much stronger relative expansion, especially in the last decade, indicating its increasing importance within the broader EO domain. Figure 2 visualizes the findings as follows: the number of publications related to Earth observation and archaeology (Figure 2a), the percentage of these publications by country (Figure 2b), the total number of Earth observation publications per country (Figure 2c), and their percentage distribution (Figure 2d).

Table 2. Temporal distribution and growth analysis (2000–2025).

Period	Archaeologically Relevant Publications	%	Growth Rate	EO Overall Publications	EO Overall %	EO Overall Growth Rate
2000–2004	117	2.7%	Baseline	13,551	7.8%	Baseline
2005–2009	396	9.1%	+238.5%	20,345	11.7%	+50.1%
2010–2014	658	15.1%	+66.2%	28,117	16.2%	+38.2%
2015–2019	1129	25.9%	+71.6%	39,007	22.5%	+38.7%
2020–2025	2059	47.2%	+82.4%	72,454	41.8%	+85.7%
Total	4359	100%	+3625.6%	173,474	100%	+1180.2%

The early foundational period from 2000 to 2005 produced only 117 publications (2.7% of the total), representing the initial exploration of satellite imagery applications in archaeological contexts when high-resolution commercial satellites were first becoming available. The foundation-building period from 2005 to 2009 witnessed steady growth to 396 publications (9.1% of the total), reflecting growing acceptance of remote sensing methodologies within the archaeological community (Figure 3). As shown in Figure 3, archaeological-related publications showed modest beginnings in 2000–2004, followed by sharp accelerations: +239% in 2005–2009, +66% in 2010–2014, +72% in 2015–2019, and a further +82% in 2020–2025. Earth observation overall follows a steadier but large-scale expansion, +50%, +38%, +39%, and +86%, in the same periods. Although the absolute volume of Earth observation publications is substantially higher, archaeological EO demonstrates a consistently stronger relative growth, especially in the last two decades. This pattern suggests that archaeological Earth observation is increasingly consolidating its role within the broader EO domain, gaining visibility and momentum at a faster pace relative to the general field.

Indeed, the expansion period from 2010 to 2014 produced 658 publications (15.1% of the total). The rapid expansion period from 2015 to 2019 reached substantial output, with 1129 publications (31.8% of the total), reflecting the establishment of archaeological remote sensing as a solid subdiscipline, driven by the launch of the Copernicus Sentinel missions

providing free, high-quality satellite imagery and increased availability of open-access historical satellite archives [16,62].

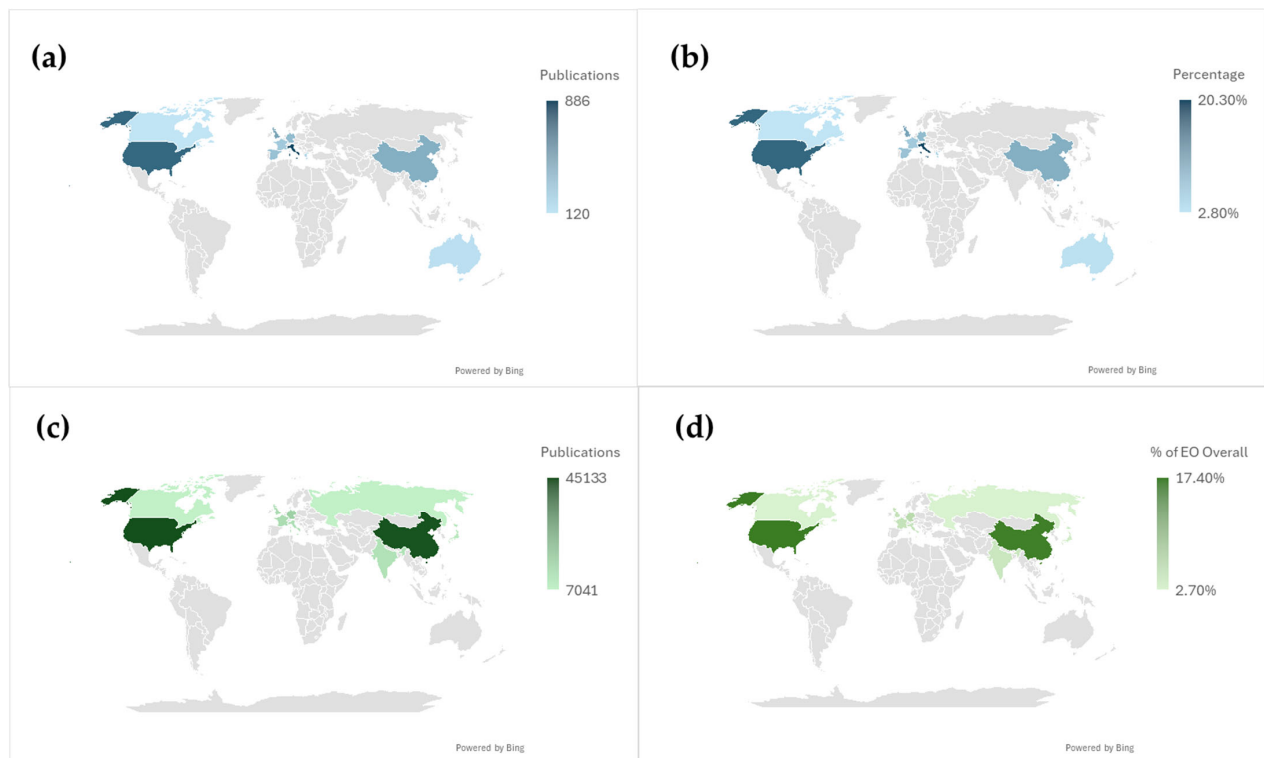


Figure 2. A map visualizing the temporal distribution and growth analysis (2000–2025). Publications related to Earth observation and archaeology: (a) the number of publications per country, (b) the percentage of publications per country, (c) the total number of Earth observation publications per country, and (d) the percentage distribution of Earth observation publications per country.

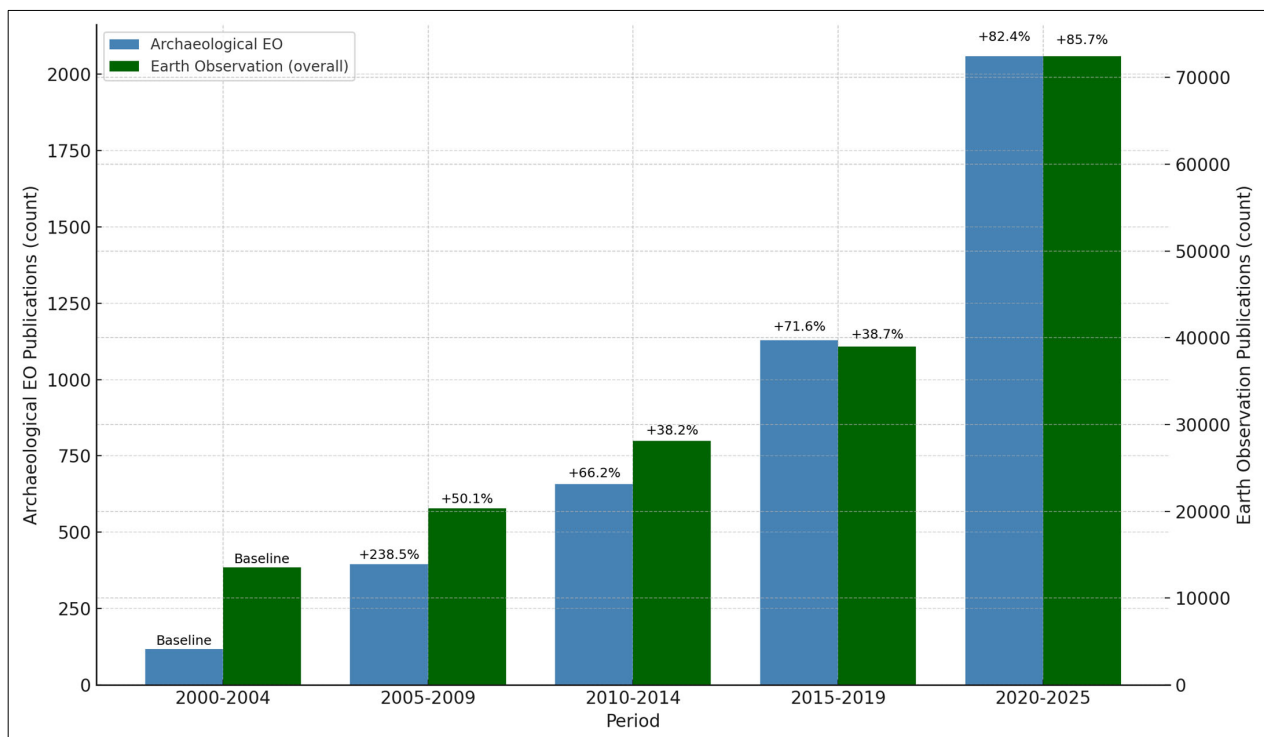


Figure 3. Temporal distribution of scientific publications (2000–2025).

Most remarkably, the recent acceleration period from 2020 to 2025 produced 2059 publications (47.0% of the total) despite representing only five years, indicating extraordinary growth in research output, with a particular emphasis on computational approaches, machine learning applications, and integration with emerging technologies [44,46].

3.2. Geographic Distribution and Regional Research Patterns

Analysis of author affiliations shows that archaeological remote sensing research is regionally concentrated, with pronounced Global North–Global South gaps reflecting unequal access to technology, infrastructure, and institutional capacity. This geographic pattern provides empirical evidence of concentration and its implications for equity in heritage protection (Table 3).

Table 3. Top contributing countries and research distribution (counts reflect papers with at least one author from country.).

Country	Publications	Percentage	Regional Position	Country	Publications	% of EO Overall	Regional Position
Italy	886	20.3%	Europe leader	United States	45,133	17.4%	North America leader
United States	726	16.7%	North America leader	China	44,247	17.0%	East Asia leader
United Kingdom	438	10.0%	Europe major	Germany	14,598	5.6%	Europe leader
China	380	8.7%	East Asia leader	United Kingdom	11,990	4.6%	Europe major
Germany	319	7.3%	Europe major	France	11,763	4.5%	Europe major
Spain	289	6.6%	Iberian leader	Italy	11,077	4.3%	Europe major
France	250	5.7%	Europe major	India	10,537	4.1%	South Asia leader
Greece	187	4.3%	Mediterranean focus	Japan	9031	3.5%	East Asia major
Australia	142	3.3%	Oceania leader	Canada	7073	2.7%	North America secondary
Canada	120	2.8%	North America secondary	Russian Federation	7041	2.7%	Eastern Europe/Eurasia

Based on institutional affiliation analysis of the dataset ($n = 4359$), Italy is the largest contributor with 886 publications (20.3%), followed by the United States with 726 (16.7%), the United Kingdom with 438 (10.0%), and China with 380 (8.7%). These patterns are consistent with differences in national investment in Earth observation infrastructure and strong university programs [63]. Importantly, Greece (4.3%), though a smaller research system, holds a significant Mediterranean position, demonstrating its disproportionate influence compared to its size. By contrast, in the broader EO landscape, China (17.0%) nearly matches the United States (17.4%), illustrating its rise as a dominant global hub. Europe remains strong through Germany, the UK, France, and Italy, but with smaller shares than in archaeological EO. This comparison highlights how countries like Italy and Greece are over-represented in Archaeological EO, while China’s global weight is more visible in overall EO.

European countries collectively show strong output in archaeological remote sensing. Germany contributes 319 publications (7.3%), Spain 289 (6.6%), France 250 (5.7%), and Greece 187 (4.3%). This distribution aligns with coordinated EU funding, shared ESA–Copernicus infrastructure, and the region’s dense archaeological resource base requiring advanced monitoring [62,64–66].

3.3. Regional Analysis and Global North–South Disparities

Using first-listed author affiliation to assign each paper to one region, the Global North accounts for 3438 publications (78.9%) and the Global South for 789 (18.1%), with 132

(3.0%) unassigned (in cases of international organization publications)—an overall ~4.4:1 Global North–Global South ratio. Within the first region, Europe dominates with 2663 (61.1%), followed by North America with 607 (13.9%) and developed Asia-Pacific with 168 (3.9%). On the Global South side, China leads with 339 (7.8%), then the Middle East/North Africa with 204 (4.7%), South Asia with 103 (2.4%), Latin America with 92 (2.1%), Sub-Saharan Africa with 26 (0.6%), and Other East Asia with 25 (0.6%). Table 4 summarizes this information along with the number of UNESCO World Heritage Sites per region.

Table 4. Regional distribution and global equity analysis.

Region	Publications	Percentage	World Heritage Sites	Research Intensity
Global North				
North America	607	13.9%	45 (3.6%)	Very High
Europe	2663	61.1%	492 (39.4%)	High
Developed Asia-Pacific	168	3.9%	66 (5.3%)	Low
Global North Total	3438	78.9%	603 (48.3%)	
Global South				
China	339	7.8%	60 (4.8%)	High
Other East Asia	25	0.6%	97 (7.8%)	Very Low
South Asia	103	2.4%	66 (5.3%)	Very Low
Middle East/North Africa	204	4.7%	151 (12.1%)	Very Low
Sub-Saharan Africa	26	0.6%	117 (9.4%)	Very Low
Latin America	92	2.1%	154 (12.3%)	Very Low
Global South Total	789	18.1%	645 (51.7%)	

North America, led by the United States and Canada, accounts for 607 publications (13.9%), reflecting substantial government and institutional investment in satellite technology and archaeological research. Europe is the dominant contributor, with 2663 publications (61.1%), benefiting from coordinated multinational programs and shared ESA–Copernicus infrastructure. Taken together, the Global North produces 3438 publications (78.9%), while the South contributes 789 (18.1%). This imbalance is visible when viewed against heritage distribution: Global South regions host 645 UNESCO World Heritage Sites (51.7%) yet generate under one-fifth of the research output.

Within the South, China contributes 339 publications (7.8%) and hosts 60 World Heritage Sites (4.8%), indicating relatively high research intensity compared with its site share, though most work targets domestic applications. For instance, during the period 2000–2025, 78.9% of the total publications were generated from the North region compared to only 18.1% of the South areas (Figure 3). Other developing regions are markedly under-represented relative to heritage significance: Sub-Saharan Africa hosts 117 World Heritage Sites (9.4%) but accounts for only 26 publications (0.6%); Latin America and the Caribbean host 154 World Heritage Sites (12.3%) but contribute 92 publications (2.1%); and the Middle East/North Africa hosts 151 World Heritage Sites (12.1%) but contributes 204 publications (4.7%). These gaps are especially concerning in areas facing acute protection challenges, where research capacity lags conservation need.

Figure 4 presents a comparative analysis of the distribution of publications between the two regions in both archaeological Earth observation (EO) and Earth observation overall for the period 2000–2025. Within archaeological EO, the Global North produces 3438 publications (81.3%), whereas the Global South contributes 789 publications (18.7%). When contrasted with the broader EO domain, however, a different pattern emerges: the North accounts for 88,387 publications (65.3%), while the South contributes a comparatively larger share of 46,950 publications (34.7%).

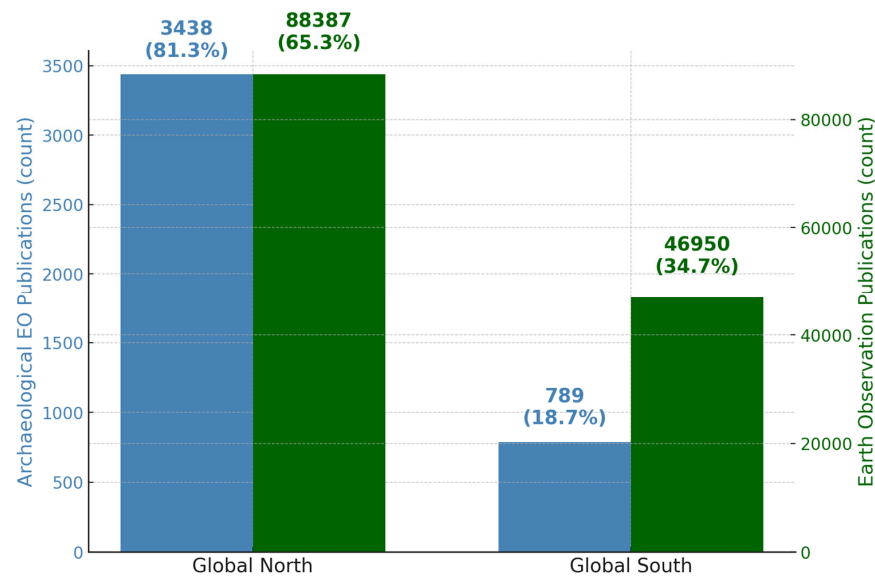


Figure 4. Global North vs. Global South publications during the period 2000–2025.

This situates archaeological Earth observation within its disciplinary context and reveals that the under-representation of the South is not merely a reflection of global publication asymmetries but is amplified in the archaeological subfield. Whereas the broader EO domain demonstrates a gradual narrowing of the North–South gap, archaeological EO exhibits a more entrenched imbalance. This finding underscores that the issue is not simply the expected predominance of the North in scientific production but a disproportionate exclusion of the Global South from a field directly relevant to its cultural and archaeological heritage.

The Research Intensity Index (RII), defined as the ratio of a region’s share of publications to its share of World Heritage Sites, can be used as a metric as it indicates over- or under-representation in research. We can therefore interpret the RII as follows: Very High ($\text{RII} \geq 2.5$) signals strong over-representation, with much more research than site share; High ($1.5 \leq \text{RII} < 2.5$) indicates clear over-representation; Moderate ($0.8 \leq \text{RII} < 1.5$) reflects research roughly proportional to site share; Low ($0.5 \leq \text{RII} < 0.8$) denotes under-representation; and Very Low ($\text{RII} < 0.5$) denotes strong under-representation. The findings of the RII metric for the period 2000–2025 per region are depicted in Figure 5. Again, the differences between these areas are obvious (except for China, as discussed below).

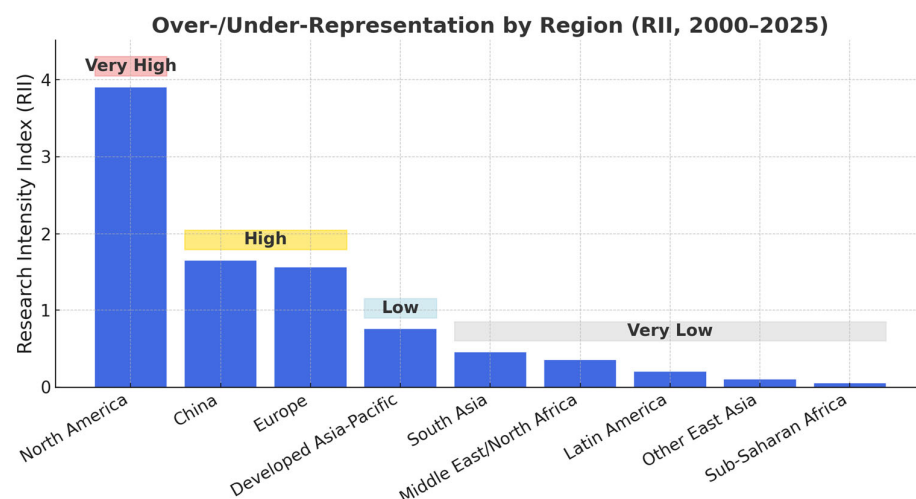


Figure 5. Research Intensity Index by region.

3.4. Technological and Methodological Evolution Patterns

Furthermore, analysis of research content shows that most studies rely on open multispectral missions—notably Landsat (TM/ETM+/OLI/TIRS, 30 m) and Sentinel-2 MSI (10–20 m)—for diachronic change detection, spectral indices, and cost-effective baselines [62,63,67]. Very-high-resolution (VHR) commercial sensors (e.g., WorldView-2/3, GeoEye-1, Pléiades/Pléiades-Neo, SPOT-6/7, PlanetScope/SkySat, Gaofen) support feature-scale site mapping, looting assessment, and excavation planning; WorldView-3 SWIR is frequently used for material discrimination [68,69]. SAR usage has risen markedly since ~2015, with Sentinel-1 (C-band) for all-weather monitoring, ALOS-1/2 PALSAR (L-band) for vegetation/moisture penetration and ground condition mapping, and TerraSAR-X/COSMO-SkyMed (X-band) for fine-texture change and deformation analysis [70–73].

Thermal/SWIR data—Landsat TIRS, ASTER VNIR/SWIR/TIR, and WorldView-3 SWIR—support material classification, paleohydrology, and burn scar/soil mark detection [68,74]. Spaceborne hyperspectral has become a promising frontier, with PRISMA and EnMAP enabling pigment/mineral/soil spectral unmixing and subtle landscape process detection [75–77]. Topographic products such as SRTM and TanDEM-X DEMs underpin landscape modeling, viewsheds, and change detection [78,79]. Historic CORONA (KH-4) imagery remains invaluable for sub-meter baselines in site discovery and long-term landscape change analysis [80,81]. Across the corpus, satellite scenes appear in >85% of publications and GISs remain ubiquitous (~60–70%), with growing multi-sensor fusion (optical + SAR + DEM) feeding contemporary machine learning workflows.

4. Discussion

4.1. Advancing Fair Access to Heritage Protection Technologies

The record shows a digital divide in heritage protection, with high-end capabilities clustered in certain regions and limited availability where resources are most at risk. The ~4.4:1 ratio between Global North and Global South research output (78.9% vs. 18.1%), despite Global South regions hosting 51.7% of UNESCO World Heritage Sites, represents a fundamental failure and gap of current approaches to technological development and dissemination. While this observation is obviously for the entire period of this study (i.e., between 2000 and 2025), as shown in Figure 6, a steady increase in papers from the Global South is reported, which should be primarily linked with increased recording of relevant publications from China (see Section 4.3).

Sub-Saharan Africa's representation of only 0.6% of research output while hosting 9.4% of UNESCO World Heritage Sites exemplifies this, indicating massive under-representation with profound implications for heritage protection capacity. The region faces severe threats to archaeological heritage, including rapid urbanization, climate change impacts, armed conflict, and extractive industry development, yet lacks access to remote sensing technologies essential for modern heritage monitoring [82,83].

This capacity gap has consequences that go well beyond academia, implicating the long-term protection and management of cultural heritage. When primary tools for documenting and monitoring archaeological heritage are developed by well-resourced institutions for their contexts, there is a risk that diverse approaches to understanding cultural landscapes will be replaced by standardized frameworks inappropriate for all cultural contexts.

The concentration of artificial intelligence development in North institutions creates additional risks of algorithmic “colonialism”, where automated heritage detection systems encode cultural assumptions inappropriate for diverse global heritage contexts. When AI systems are trained primarily on specific landscapes and well-documented sites,

their application to other environments may systematically under-represent important cultural resources.

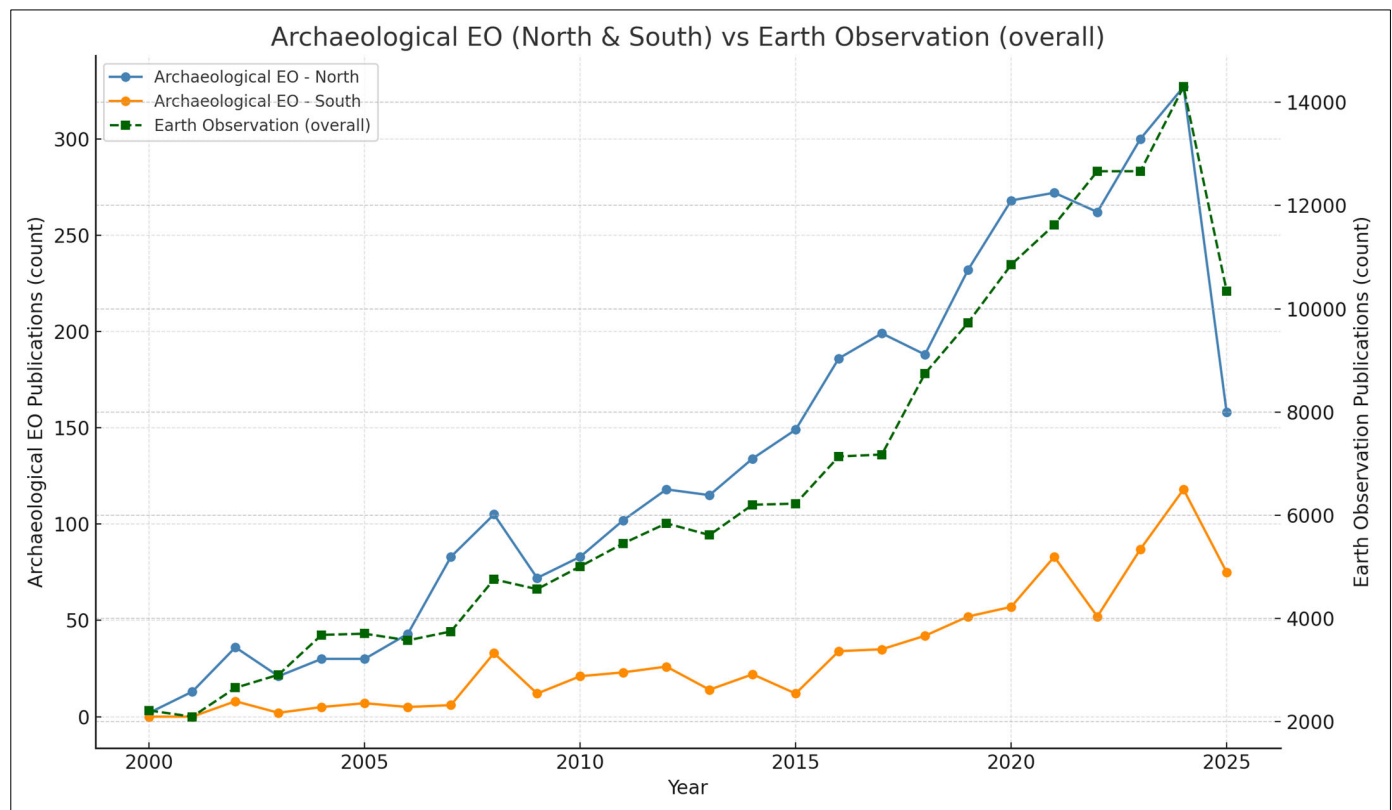


Figure 6. Publications per year (2000–2025) for Global South against Global North. Earth observation-related publications are also indicated with green (secondary Y-axis).

Funding analysis from the available references indicates diverse support mechanisms including national research councils (such as the National Social Science Fund of China), though many publications lack explicit funding acknowledgments, potentially indicating either unfunded research or different reporting practices across regions and institutions. The presence of both government and private funding sources suggests multiple pathways for supporting archaeological remote sensing research, though the distribution of funding appears uneven across geographic regions.

Figure 7 illustrates the regional distribution of publications in archaeology using Earth observation research from 2000 to 2025, expressed as the share of total output per year. Europe consistently dominates the field, accounting for the largest share throughout the period, although its relative proportion shows fluctuations over time. North America also maintains a significant but smaller share, while Asia demonstrates a steady increase, particularly after 2015, reflecting the region’s growing research capacity and investment in space-based applications. Latin America, the Middle East, Oceania, and Africa remain under-represented, with Africa showing only a marginal contribution despite its extensive cultural heritage and research needs. The “Other” category captures a persistent portion of publications, indicating contributions from international organizations. Overall, the chart highlights enduring regional disparities, with the Global North maintaining dominance but Asia emerging as a notable contributor in recent years.

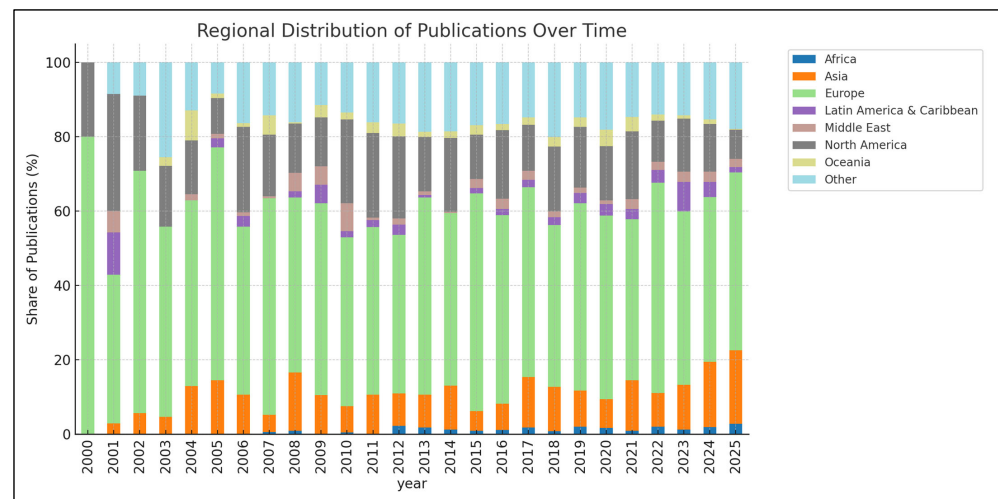


Figure 7. Publications per year (2000–2025) per region.

4.2. The European Model and the Acceleration Paradox

As mentioned earlier, European institutions account for 61.1% of global archaeological remote sensing output, illustrating how coordinated regional approaches can lift capacity beyond what individual countries might achieve alone. Key drivers include shared technological infrastructure through ESA’s Copernicus program, coordinated EU funding frameworks, and complementary national strengths in archaeological heritage and remote-sensing technology. Italy is the largest single-country contributor at 20.3%, reflecting both its exceptional archaeological heritage and deep integration in European collaboration networks and technology investment.

Countries such as Greece (4.3%) show how strong heritage assets, combined with access to shared infrastructure, can sustain notable research productivity even in smaller economies. Replicating the European model elsewhere will require tailoring to local historical, institutional, and economic contexts; for the Global South, this may mean prioritizing South–South collaboration networks, regional centers of excellence that build capability without dependency, and staged technology transfer programs that progressively develop local capacity [84].

With 47.2% of all papers published in 2020–2025 (2059 of 4359; +82.4% vs. 2015–2019), the field’s acceleration creates a paradox: sophistication expands possibilities while erecting new barriers to participation. It should be mentioned that comparable regional initiatives are also emerging in the Global South. In Latin America, programs such as SERVIR Amazonia [85] promote EO capacity building, while in Africa, the GMES & Africa initiative [86] fosters regional EO services and training. Although smaller in scope and less heritage-focused than Europe’s, these efforts highlight the potential of regional models beyond the European context.

The rapid uptake of AI (see [48,87–89]) promises major gains in site discovery and monitoring, yet it depends on substantial computing resources, specialist expertise, and suitable training datasets that many communities lack. This acceleration is especially challenging for Global South institutions still building basic remote sensing capacity as the field shifts toward compute-intensive methods and advanced infrastructure. The growth pattern also raises questions about sustainability and whether the pace of development fosters unrealistic expectations for purely technological solutions to heritage protection.

4.3. The Chinese Case Study

China’s position as the fourth-largest contributor to archaeological remote sensing research—accounting for 8.7% of global output—presents a complex case that challenges

simple Global North–South categorizations while raising questions about technological sovereignty and research autonomy in archaeological applications [13,90]. China’s substantial output reflects major state investment in satellite technology and archaeological research infrastructure, demonstrating how national commitment to technological development can rapidly build research capacity in previously under-represented regions.

At the same time, Chinese publishing patterns reveal limitations that distinguish its position from traditional Global North leaders. Much of the work centers on domestic heritage sites and applications rather than international collaboration or methods with broad global applicability. Nearly one-fifth of the relevant records from China are Chinese-language publications ($n = 66$). This pattern suggests that, despite significant technological capacity, research remains relatively insulated from international networks and may exert limited influence on global methodological development and collaborative frameworks. Nevertheless, this appears to be changing, particularly over the last decade.

The Chinese case illustrates both the potential and limits of nationally focused capacity building in archaeological remote sensing. While substantial public investment can rapidly expand technical capabilities and publication output, weak ties to international collaboration may constrain global impact and the transferability of findings. Technological capacity in this field requires not only access to satellites and computational resources but also integration with international research communities and collaborative infrastructures. For other Global South countries seeking independent capacity, the Chinese experience suggests that sustained international collaboration and knowledge sharing are essential complements to investment, maximizing both global impact and practical application.

Nevertheless, steps forward have been made, mostly driven by China. Figure 8 compares the share of publications between the two geographic areas from 2000 to 2025. The Global North initially dominates the field with nearly 100% of publications in 2000, but its share shows a gradual decline over time, falling to around 60% by 2025. In contrast, the Global South starts from a marginal contribution, but its share steadily increases, particularly after 2010, reaching close to 40% by 2025. This trend highlights a narrowing gap between the two regions, suggesting growing research participation and capacity in the Global South—mainly due to China—though the North continues to hold the majority share. The overall pattern indicates a slow but significant shift toward greater global inclusivity in Earth observation-related archaeological research. Notably, the proportions shown in Figure 8, especially for 2025, align closely with the standard pattern documented in Earth observation research (see Figure 4).

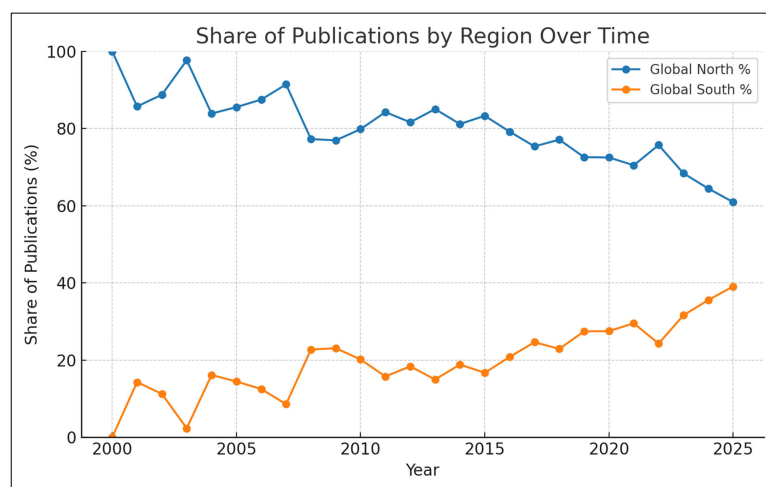


Figure 8. Share (%) of publications per year (2000–2025) per region.

4.4. Pathways Toward Technological Justice

International funding agencies need to restructure their approaches to prioritize comprehensive capacity building over short-term technology transfer, with explicit requirements for Global South leadership and sustainability. This includes supporting regional collaborative networks enabling peer-to-peer knowledge sharing, investing in computational infrastructure and educational programs building local expertise, and funding culturally appropriate technology development. The “Endangered Archaeology in the Middle East and North Africa” (EAMENA) initiative [91] as well as other similar projects are some examples of this approach.

Moreover, efforts should be made in the domain of AI systems with globally representative training datasets, creating simplified analytical tools and providing essential capabilities without requiring advanced infrastructure, establishing open-source development communities, and enabling meaningful Global South participation.

Educational institutions should recognize technology transfer as legitimate scholarly activity deserving professional recognition equivalent to traditional research. This includes developing appropriate evaluation metrics extending beyond publication counts, creating interdisciplinary curricula integrating technological training with cultural sensitivity, and establishing international partnership programs prioritizing capacity building.

5. Conclusions

This systematic review of 4359 unique publications spanning 25 years documents both extraordinary growth and profound geographic disparities in archaeological remote sensing. Quantitatively, 47.2% of all output appears in 2020–2025, while 78.9% originates from North institutions—even though the Global South hosts 51.7% of UNESCO World Heritage Sites (the Global North hosts 48.3%). At the same time, the overall trend of the scientific field of Earth observation was studied for comparison purposes.

The temporal analysis shows near-exponential growth: output rose by $\approx 1661\%$ from 2000–2004 to 2020–2025. This acceleration signals that remote sensing has become central to archaeological practice, driven by advances in satellites, computing, and AI. Yet growth is concentrated geographically, reinforcing rather than easing global inequalities.

Just three countries—Italy (20.3%), the United States (16.7%), and the United Kingdom (10.0%)—account for $\approx 47\%$ of global output, while regions with the greatest protection challenges remain under-represented. Sub-Saharan Africa contributes only 0.6% of research despite hosting 9.4% of WHSs, a stark capacity gap with direct implications for heritage protection. This gap between the two regions is also wider than that observed in the general Earth observation field (Figure 3), although it has narrowed in recent years (Figure 6).

The crisis extends beyond publication counts to cultural survival, equity, and excellence in a digital era where Earth observation technologies are essential for safeguarding heritage [92–98]. When tools are developed mainly by well-resourced institutions for their own contexts, diverse ways of reading cultural landscapes risk being displaced by standardized frameworks ill suited elsewhere. The concentration of AI development in Global North institutions adds risks of algorithmic “colonialism”: models trained largely on well-documented sites can underperform in other environments. The concentration of artificial intelligence development in Global North institutions, evident in sophisticated applications like the vision transformer research achieving 98.1% settlement recognition accuracy [99], creates additional risks of algorithmic colonialism where automated heritage detection systems encode cultural assumptions and analytical frameworks that may be inappropriate for diverse global heritage contexts.

Europe’s success illustrates what coordinated regional approaches can achieve—shared infrastructure and funding enabling productivity beyond any single country. The lack of

analogous frameworks across much of the Global South underscores the need for institutional interventions that tackle structural barriers to cooperation and capacity building.

Acceleration creates both opportunity and risk. Without deliberate action to broaden access to data, computer, and training resources, rising technical sophistication may harden global stratification rather than reduce it. Heritage sites are under threat globally due to the same reasons: climate change, urbanization, and conflict intensify threats to sites worldwide. Concentrated capabilities leave many at-risk places without effective monitoring. The evidence base here supports concrete strategies for a more equitable technological pathway—recognize heritage protection as a global public good, move beyond market-only logics, and invest in coordinated measures that address structural inequalities—so the demonstrable potential of archaeological remote sensing benefits all communities, not just the best resourced.

Of course, we need to acknowledge that the present analysis is bounded by several methodological and conceptual constraints. First, it adopts the Global North–Global South dichotomy as the principal analytical lens. While this binary provides a widely recognized heuristic for examining structural asymmetries, it inevitably obscures other salient dimensions of research capacity, including the possession of national space programs, environmental constraints such as persistent cloud cover in tropical and equatorial regions, the uneven global distribution of cultural heritage resources, and the heterogeneity of national research and heritage management policies. China, in particular, presents a case that challenges the simplicity of the North–South binary. While classified within the Global South, China simultaneously operates as a global leader in space technology and Earth observation research. Its dual position highlights the limitations of rigid categorizations and underscores the need for frameworks that account for both structural inequalities and emerging scientific powers.

Second, the exclusive reliance on Scopus as a bibliometric source introduces potential coverage biases. Although Scopus offers extensive metadata and robust analytical functionality, its indexing practices tend to favor journals from established publishers and high-income regions, and may therefore under-represent scholarly outputs from the Global South, particularly those disseminated through regional or local outlets. These limitations suggest that the disparities documented here should be regarded as conservative estimates, and they highlight the importance of future work that incorporates multiple bibliometric databases and alternative classificatory frameworks.

Funding: This paper was supported by the ENGINEER and CONNECTING project. The CONNECTING project has received funding from the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation of Cyprus under contract SMALL SCALE INFRASTRUCTURES/1222/0062. The ENGINEER project has received funding from the European Union’s Horizon Europe Framework Programme (HORIZON-WIDERA-2021-ACCESS-03, Twinning Call) under the grant agreement no. 101079377 and the UKRI under project number 10050486.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The author acknowledges the use of language editing tools (Grammarly (2025)) and AI-based assistants (ChatGPT (version GPT-5) by OpenAI, Claude (version 4.5) by Anthropic) to improve the readability and grammar of the manuscript. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest. The views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or the UKRI. Neither the European Union nor the UKRI can be held responsible for them.

References

1. Lasaponara, R.; Masini, N. Satellite remote sensing in archaeology: Past, present and future perspectives. *J. Archaeol. Sci.* **2011**, *38*, 1995–2002. [CrossRef]
2. Zingaro, M.; Scicchitano, G.; Capolongo, D. The Innovative growth of space archaeology: A brief overview of concepts and approaches in detection, monitoring, and promotion of the archaeological heritage. *Remote Sens.* **2023**, *15*, 3049. [CrossRef]
3. Tapete, D. Remote sensing and geosciences for archaeology. *Geosciences* **2018**, *8*, 41. [CrossRef]
4. Custer, J.F.; Eveleigh, T.; Klemas, V.; Wells, I. Application of Landsat data and synoptic remote sensing to predictive models for prehistoric archaeological sites: An example from the Delaware coastal plain. *Am. Antiq.* **1986**, *51*, 572–588. [CrossRef]
5. Pope, K.O.; Dahlin, B.H. Ancient Maya wetland agriculture: New insights from ecological and remote sensing research. *J. Field Archaeol.* **1989**, *16*, 87–106. [CrossRef]
6. Sever, T.L. *Remote Sensing Applications in Archeological Research: Tracing Prehistoric Human Impact Upon the Environment*; University of Colorado at Boulder: Boulder, CO, USA, 1990.
7. Sever, T.L.; Irwin, D.E. Landscape archaeology: Remote-sensing investigation of the ancient Maya in the Peten rainforest of northern Guatemala. *Anc. Mesoam.* **2003**, *14*, 113–122. [CrossRef]
8. Campana, S.R. Remote sensing in archaeology. In *Encyclopedia of Geoarchaeology*; Springer International Publishing: Cham, Switzerland, 2023; pp. 1–25.
9. Agapiou, A.; Lysandrou, V. Remote sensing archaeology: Tracking and mapping evolution in European scientific literature from 1999 to 2015. *J. Archaeol. Sci. Rep.* **2015**, *4*, 192–200. [CrossRef]
10. Opitz, R.; Herrmann, J. Recent trends and long-standing problems in archaeological remote sensing. *J. Comput. Appl. Archaeol.* **2018**, *1*, 19–41. [CrossRef]
11. Evans, D.; Travaglia, A. Uncovering Angkor: Integrated remote sensing applications in the archaeology of early Cambodia. In *Satellite Remote Sensing*; Springer: Dordrecht, The Netherlands, 2012; pp. 197–230.
12. Lasaponara, R.; Masini, N. Remote sensing in archaeology: From visual data interpretation to digital data manipulation. In *Satellite Remote Sensing*; Springer: Dordrecht, The Netherlands, 2012; pp. 3–16.
13. Luo, L.; Wang, X.; Guo, H.; Lasaponara, R.; Zong, X.; Masini, N.; Wang, G.; Shi, P.; Khatteli, H.; Chen, F.; et al. Airborne and spaceborne remote sensing for archaeological and cultural heritage applications: A review of the century (1907–2017). *Remote Sens. Environ.* **2019**, *232*, 111280. [CrossRef]
14. Stewart, C.; Lasaponara, R.; Schiavon, G. ALOS PALSAR analysis of the archaeological site of Pelusium in the Eastern Nile Delta (Egypt). *Archaeol. Prospect.* **2016**, *23*, 109–116.
15. Agapiou, A. Remote sensing heritage in a petabyte-scale: Satellite data and heritage Earth Engine© applications. *Int. J. Digit. Earth* **2017**, *10*, 85–102. [CrossRef]
16. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
17. Fu, D.; Xiao, H.; Su, F.; Zhou, C.; Dong, J.; Zeng, Y.; Yan, K.; Li, S.; Wu, J.; Wu, W.; et al. Remote sensing cloud computing platform development and Earth science application. *Natl. Remote Sens. Bull.* **2021**, *25*, 220–230. [CrossRef]
18. Luo, L.; Wang, X.; Guo, H.; Jia, X.; Fan, A. Earth observation in archaeology: A brief review. *Int. J. Appl. Earth Obs. Geoinform.* **2023**, *116*, 103169. [CrossRef]
19. Fisher, M.; Fradley, M.; Flohr, P.; Rouhani, B.; Simi, F. Ethical considerations for remote sensing and open data in relation to the endangered archaeology in the Middle East and North Africa project. *Archaeol. Prospect.* **2021**, *28*, 279–292. [CrossRef]
20. Davis, D.S.; Buffa, D.; Rasolondrainy, T.; Creswell, E.; Anyanwu, C.; Ibirogbu, A.; Randolph, C.; Ouarghidi, A.; Phelps, L.N.; Lahiniriko, F.; et al. The aerial panopticon and the ethics of archaeological remote sensing in sacred cultural spaces. *Archaeol. Prospect.* **2021**, *28*, 305–320. [CrossRef]
21. Davis, D.S. Geographic disparity in machine intelligence approaches for archaeological remote sensing research. *Remote Sens.* **2020**, *12*, 921. [CrossRef]
22. Matás, G.J.L.; Lim, J.S. Monumental funerary landscapes of Dhar Tagant (south-eastern Mauritania): Towards ethical satellite remote sensing in the West African Sahel. *Archaeol. Prospect.* **2021**, *28*, 357–378. [CrossRef]
23. Mahler, A.G. What/Where is the Global South? In *Oxford Bibliographies in Literary and Critical Theory*; O'Brien, E., Ed.; Oxford University Press: Oxford, UK, 2017. Available online: <https://www.globalsouthstudies.org/what-is-the-global-south/> (accessed on 30 August 2025).
24. Joyce, K.E.; Nakalembe, C.L.; Gómez, C.; Suresh, G.; Fickas, K.; Halabisky, M.; Kalamandeen, M.; Crowley, M.A. Discovering Inclusivity in Remote Sensing: Leaving No One Behind. *Front. Remote Sens.* **2022**, *3*, 869291. [CrossRef]
25. Thorpe, D. Dependent or not? From a daily practice of Earth observation research in the Global South to promoting adequate developmental spaces in science and technology studies. *Geogr. Helv.* **2023**, *78*, 105–130. [CrossRef]
26. Woodcock, C.E.; Allen, R.; Anderson, M.; Belward, A.; Bindschadler, R.; Cohen, W.; Gao, F.; Goward, S.N.; Helder, D.; Helmer, E.; et al. Free Access to Landsat Imagery. *Science* **2008**, *320*, 1011. [CrossRef] [PubMed]

27. Karnieli, A.; Berk, A.; Bayarjargal, Y. Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey. *Remote. Sens.* **2024**, *17*, 550. [CrossRef]
28. Wulder, M.A.; Coops, N.C. Satellites: Make Earth observations open access. *Nature* **2014**, *513*, 30–31. [CrossRef] [PubMed]
29. Rouhani, B. From Ruins to Records: Digital Strategies and Dilemmas in Cultural Heritage Protection. In *Digital Colonialism and Heritage Documentation*; Routledge: Oxfordshire, UK, 2025; pp. 242–257.
30. Meskell, L. *A Future in Ruins: UNESCO, World Heritage, and the Dream of Peace*; Oxford University Press: Oxford, UK, 2018.
31. UNESCO World Heritage Centre. World Heritage List. 2024. Available online: <https://whc.unesco.org/en/list/> (accessed on 24 August 2024).
32. Casana, J. Emerging approaches to archaeological remote sensing. *Adv. Archaeol. Pract.* **2021**, *9*, 510–523. [CrossRef]
33. Parcak, S. *Archaeology from Space: How the Future Shapes Our Past*; Henry Holt and Co.: New York, NY, USA, 2019.
34. Comer, D.C.; Blakely, R.J. Global heritage in a digital age. *J. Cult. Herit.* **2020**, *41*, 224–231. [CrossRef]
35. Davis, D.S.; Douglass, K. Aerial and spaceborne remote sensing in African archaeology: A review of current research and potential future avenues. *Spat. Approaches Afr. Archaeol.* **2020**, *37*, 9–24. [CrossRef]
36. Tapete, D.; Cigna, F. Trends and perspectives of space-borne SAR remote sensing for archaeological landscape monitoring. *J. Archaeol. Sci. Rep.* **2019**, *25*, 513–525. [CrossRef]
37. Forte, M.; Campana, S. *Digital Methods and Remote Sensing in Archaeology: Archaeology in the Age of Sensing*; Springer: Berlin/Heidelberg, Germany, 2019.
38. UNESCO. *Heritage Emergency Fund: Annual Progress Report 2023*; UNESCO: Paris, France, 2023. Available online: <https://docs.un.org/en/A/78/217> (accessed on 30 August 2025).
39. UNESCO. *Priority Africa: World Heritage Capacity-Building Strategy for Africa*; UNESCO World Heritage Centre: Paris, Italy, 2023. Available online: <https://whc.unesco.org/document/199636> (accessed on 30 August 2025).
40. UNESCO. *Operational Guidelines for the Implementation of the World Heritage Convention*; UNESCO World Heritage Centre: Paris, France, 2022. Available online: <https://icomos.es/wp-content/uploads/2024/04/document-57-39.pdf> (accessed on 30 August 2025).
41. Rouhani, B. Training in Endangered Archaeology: From Digital Documentation to Empowering National Heritage Stakeholders in the Middle East and North Africa. 2018. Available online: <https://publ.icomos.org/publicomos/jlbSai?html=Pag&page=Pml/Not&base=technica&ref=6B9D268B4AD3835AB2B87534065A8AEE> (accessed on 30 August 2025).
42. Menéndez-Marsh, F.; Al-Rawi, M.; Fonte, J.; Dias, R.; Gonçalves, L.J.; Seco, L.G.; Neves, A. Geographic information systems in archaeology: A systematic review. *J. Comput. Appl. Archaeol.* **2023**, *6*, 40–50. [CrossRef]
43. Trier, Ø.D.; Cowley, D.C.; Waldeland, A.U. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeol. Prospect.* **2019**, *26*, 165–175. [CrossRef]
44. Orenge, H.; Garcia-Molsosa, A. A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery. *J. Archaeol. Sci.* **2019**, *112*, 105013. [CrossRef]
45. Bickler, S.H. Machine learning arrives in archaeology, but when do computers see like an archaeologist? *Adv. Archaeol. Pract.* **2021**, *9*, 122–136. [CrossRef]
46. Davis, D.S. Object-based image analysis: A review of developments and future directions of automated feature detection in landscape archaeology. *Archaeol. Prospect.* **2019**, *26*, 155–163. [CrossRef]
47. Soroush, M.; Mehrtash, A.; Khazraee, E.; Ur, J.A. Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq. *Remote Sens.* **2020**, *12*, 500. [CrossRef]
48. Verschoof-van der Vaart, W.B.; Lambers, K. Automated detection of archaeological objects from remotely sensed data using machine learning: A review of progress and challenges. *J. Archaeol. Sci.* **2021**, *133*, 105428.
49. Argyrou, A.; Agapiou, A. A review of artificial intelligence and remote sensing for archaeological research. *Remote Sens.* **2022**, *14*, 6000. [CrossRef]
50. Kadhim, I.; Abed, F.M. A critical review of remote sensing approaches and deep learning techniques in archaeology. *Sensors* **2023**, *23*, 2918. [CrossRef]
51. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
52. Casana, J. Global-scale archaeological prospection using machine learning on satellite imagery. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 8554–8562.
53. Somrak, M.; Džeroski, S.; Kokalj, Ž. Learning to classify structures in ALS-derived visualizations of ancient Maya settlements with CNN. *Remote Sens.* **2020**, *12*, 2215. [CrossRef]
54. Kokalj, Ž.; Žiberna, A.; Briese, C.; Doneus, M. Machine learning-ready remote sensing data for Maya archaeology. *Sci. Data* **2023**, *10*, 24. [CrossRef]
55. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Informetr.* **2017**, *11*, 959–975. [CrossRef]

56. Van Eck, N.J.; Waltman, L. Visualizing bibliometric networks. In *Measuring Scholarly Impact*; Springer International Publishing: Cham, Switzerland, 2014; pp. 285–320.
57. Baas, J.; Schotten, M.; Plume, A.; Côté, G.; Karimi, R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quant. Sci. Stud.* **2020**, *1*, 377–386. [\[CrossRef\]](#)
58. Mongeon, P.; Paul-Hus, A. The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics* **2016**, *106*, 213–228. [\[CrossRef\]](#)
59. Bramer, W.M.; Rethlefsen, M.L.; Kleijnen, J.; Franco, O.H. Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Syst. Rev.* **2017**, *6*, 1–12. [\[CrossRef\]](#)
60. Lefebvre, C.; Glanville, J.; Briscoe, S.; Littlewood, A.; Marshall, C.; Metzendorf, M.I.; Noel-Storr, A.; Rader, T.; Shokraneh, F.; Thomas, J.; et al. Searching for and selecting studies. In *Cochrane Handbook for Systematic Reviews of Interventions*; John Wiley & Sons: Chichester, UK, 2019; pp. 67–107.
61. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#)
62. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [\[CrossRef\]](#)
63. Wulder, M.A.; Loveland, T.R.; Roy, D.P.; Crawford, C.J.; Masek, J.G.; Woodcock, C.E.; Allen, R.G.; Anderson, M.C.; Belward, A.S.; Cohen, W.B.; et al. Current status of Landsat program, science, and applications. *Remote Sens. Environ.* **2019**, *225*, 127–147. [\[CrossRef\]](#)
64. Malenovsky, Z.; Rott, H.; Cihlar, J.; Schaepman, M.E.; García-Santos, G.; Fernandes, R.; Berger, M. Sentinels for science: Potential of Sentinel-1, -2, and -3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sens. Environ.* **2012**, *120*, 91–101. [\[CrossRef\]](#)
65. Abate, N.; Lasaponara, R. Preventive archaeology based on open remote sensing data and tools: The cases of Sant'Arsenio (SA) and Foggia (FG), Italy. *Sustainability* **2019**, *11*, 4145. [\[CrossRef\]](#)
66. Agapiou, A.; Hegyi, A.; Gogăltan, F.; Stavilă, A.; Sava, V.; Sarris, A.; Floca, C.; Dorogostaisky, L. Exploring the largest known Bronze Age earthworks in Europe through medium resolution multispectral satellite images. *Int. J. Appl. Earth Obs. Geoinform.* **2023**, *118*, 103239. [\[CrossRef\]](#)
67. Roy, D.P.; Wulder, M.A.; Loveland, T.R.; Woodcock, C.E.; Allen, R.G.; Anderson, M.C.; Helder, D.; Irons, J.R.; Johnson, D.M.; Kennedy, R.; et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sens. Environ.* **2014**, *145*, 154–172. [\[CrossRef\]](#)
68. DigitalGlobe. WorldView-3 Data Sheet (Instrument & SWIR Specifications). 2014. Available online: https://www.spaceimagingme.com/downloads/sensors/datasheets/DG_WorldView3_DS_2014.pdf (accessed on 30 August 2025).
69. Planet Team. *Planet Application Program Interface: In Space for Life on Earth*; Planet Team: San Francisco, CA, USA, 2017.
70. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [\[CrossRef\]](#)
71. Werninghaus, R.; Buckreuss, S. The TerraSAR-X mission and system design. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 606–614. [\[CrossRef\]](#)
72. Covello, F.; Battazza, F.; Coletta, A.; Lopinto, E.; Fiorentino, C.; Pietranera, L.; Valentini, G.; Zoffoli, S. The COSMO-SkyMed Mission: A SAR Constellation for Earth Observation. IGARSS/IEEE TGRS (Mission Overview). 2010. Available online: <https://hal.science/hal-00615310/document> (accessed on 30 August 2025).
73. Covello, F.; Battazza, F.; Coletta, A.; Zoffoli, S. COSMO-SkyMed mission status. In Proceedings of the EUSAR 2018; 12th European Conference on Synthetic Aperture Radar, Aachen, Germany, 4–7 June 2018.
74. Yamaguchi, Y.; Kahle, A.; Tsu, H.; Kawakami, T.; Pniel, M. Overview of Advanced Spaceborne Thermal Emission and Reflection Radiometer instrument. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1062–1071. [\[CrossRef\]](#)
75. Guarini, R.; Loizzo, R.; Facchinetti, C.; Longo, F.; Ponticelli, B.; Faraci, M.; Dami, M.; Cosi, M.; Amoruso, L.; De Pasquale, V.; et al. Prisma Hyperspectral Mission Products. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 179–182.
76. Shaik, R.U.; Periasamy, S.; Zeng, W. Potential Assessment of PRISMA Hyperspectral Imagery for Remote Sensing Applications. *Remote Sens.* **2023**, *15*, 1378. [\[CrossRef\]](#)
77. Guanter, L.; Kaufmann, H.; Segl, K.; Foerster, S.; Rogass, C.; Chabrillat, S.; Kuester, T.; Hollstein, A.; Rossner, G.; Chlebek, C.; et al. The EnMAP Spaceborne Imaging Spectroscopy Mission for Earth Observation. *Remote Sens.* **2015**, *7*, 8830–8857. [\[CrossRef\]](#)
78. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. *Rev. Geophys.* **2007**, *45*, RG2004. [\[CrossRef\]](#)

79. Krieger, G.; Zink, M.; Bachmann, M.; Bräutigam, B.; Schulze, D.; Martone, M.; Rizzoli, P.; Steinbrecher, U.; Antony, J.W.; De Zan, F.; et al. TanDEM-X: A radar interferometer with two formation-flying satellites. *Acta Astronaut.* **2013**, *89*, 83–98. [\[CrossRef\]](#)
80. Casana, J.; Cothren, J. Stereo analysis, DEM extraction and orthorectification of CORONA satellite imagery: Archaeological applications from the Near East. *Antiquity* **2008**, *82*, 732–749. [\[CrossRef\]](#)
81. Agapiou, A. Land cover mapping from colorized CORONA archived greyscale satellite data and feature extraction classification. *Land* **2021**, *10*, 771. [\[CrossRef\]](#)
82. Chirikure, S. Heritage conservation in Africa: The good, the bad, and the challenges. *S. Afr. J. Sci.* **2013**, *109*, 1–3. [\[CrossRef\]](#)
83. Brooks, N.; Clarke, J.; Ngaruiya, G.W.; Wangui, E.E. African heritage in a changing climate. *Azania Archaeol. Res. Afr.* **2020**, *55*, 297–328. [\[CrossRef\]](#)
84. Sesay, A.; Olusola, O.K.; Omotosho, M. Africa and South–South cooperation: Opportunities and challenges. In *The Rise of the Global South: Philosophical, Geopolitical and Economic Trends of the 21st Century*; World Scientific Publishing Co. Pte. Ltd.: Singapore, 2013; pp. 95–123.
85. SERVIR Amazonia. SERVIR Amazonia: Connecting Space to Village. 2023. Available online: <https://servir.ciat.cgiar.org/> (accessed on 30 August 2025).
86. Asante, K.O.; Macuácuá, E.; Mutanga, O. Capacity building in Earth Observation: Experiences from GMES & Africa. *Remote Sens. Appl. Soc. Environ.* **2021**, *23*, 100546. [\[CrossRef\]](#)
87. Landauer, J.; Klassen, S.; Wijker, A.P.; van der Kroon, J.; Jaszowski, A.; der Vaart, W.B.V.-V. Beyond the Greater Angkor Region: Automatic Large-Scale Mapping of Angkorian-Period Reservoirs in Satellite Imagery Using Deep Learning. *PLoS ONE* **2025**, *20*, e0320452. [\[CrossRef\]](#) [\[PubMed\]](#)
88. Sobotkova, A.; Kristensen-McLachlan, R.D.; Mallon, O.; Ross, S.A. Validating Predictions of Burial Mounds with Field Data: The Promise and Reality of Machine Learning. *J. Doc.* **2024**, *80*, 1167–1189. [\[CrossRef\]](#)
89. Fylaktos, P.; Petropoulos, G.P.; Lemesios, I. A Deep Learning Method for the Automated Mapping of Archaeological Structures from Geospatial Data: A Case Study of Delos Island. *ISPRS Int. J. Geo-Inform.* **2025**, *14*, 220. [\[CrossRef\]](#)
90. Chen, F.; Lasaponara, R.; Masini, N. An overview of satellite synthetic aperture radar remote sensing in archaeology: From site detection to monitoring. *J. Cult. Heritage* **2017**, *23*, 5–11. [\[CrossRef\]](#)
91. Bewley, R. Endangered archaeology in the Middle East and North Africa (EAMENA): Approach and possible solutions. In *Post-Conflict Archaeology and Cultural Heritage*; Routledge: New York, NY, USA, 2017; pp. 123–137.
92. Cuca, B. The contribution of earth observation technologies to monitoring strategies of cultural landscapes and sites. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 135–140. [\[CrossRef\]](#)
93. Xiao, W.; Mills, J.; Guidi, G.; Rodríguez-Gonzálvez, P.; Barsanti, S.G.; González-Aguilera, D. Geoinformatics for the conservation and promotion of cultural heritage in support of the UN Sustainable Development Goals. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 389–406. [\[CrossRef\]](#)
94. Anderson, K.; Ryan, B.; Sonntag, W.; Kavvada, A.; Friedl, L. Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-Spat. Inf. Sci.* **2017**, *20*, 77–96. [\[CrossRef\]](#)
95. Chen, F.; Zhou, W.; Xu, H.; Parcharidis, I.N.; Lin, H.; Fang, C. Space technology facilitates the preventive monitoring and preservation of the Great Wall of the Ming dynasty: A comparative study of the Qingtongxia and Zhangjiakou sections in China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5719–5729. [\[CrossRef\]](#)
96. Moise, C.; Lazar, A.-M.; Mihalache, C.E.; Dedulescu, L.A.; Negula, I.F.D.; Badea, A.; Poenaru, V.D.; Moise, R.; Ortan, A.R. Geomatics Technologies in the Framework of Multidisciplinary Project for Integrated Management of Cultural Heritage Sites. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *46*, 1477–1484. [\[CrossRef\]](#)
97. Cannelli, S.; de Veld, F.; Gehlot, M.; Nugent, T.K.; Thangavel, K. Earth’s orbits as a UNESCO World Heritage Site. In Proceedings of the 74th International Astronautical Congress, Baku, Azerbaijan, 2–6 October 2023.
98. Agapiou, A.; Lysandrou, V.; Cuca, B. Copernicus Earth Observation for Cultural Heritage. In Proceedings of the Joint International Event 9th ARQUEOLÓGICA 2.0 & 3rd GEORES, Valencia, Spain, 26–28 April 2021; pp. 259–266.
99. Zhong, Q.; Luo, J.; Fang, J.; Chen, Y. Two-stage cascaded vision transformer with spatial attention for dense settlement detection in remote sensing imagery. *npj Heritage Sci.* **2025**, *13*, 211. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Review

AI Chatbots and Remote Sensing Archaeology: Current Landscape, Technical Barriers, and Future Directions

Nicolas Melillos * and Athos Agapiou 

Department of Civil Engineering and Geomatics, Faculty of Engineering and Technology,
Cyprus University of Technology, Saripolou 2-8, 3036 Limassol, Cyprus; athos.agapiou@cut.ac.cy

* Correspondence: ng.melillos@edu.cut.ac.cy

Abstract

Chatbots have emerged as a promising interface for facilitating access to complex datasets, allowing users to pose questions in natural language rather than relying on specialized technical workflows. At the same time, remote sensing has transformed archaeological practice by producing vast amounts of imagery from LiDAR, drones, and satellites. While these advances have created unprecedented opportunities for discovery, they also pose significant challenges due to the scale, heterogeneity, and interpretative demands of the data. In related scientific domains, multimodal conversational systems capable of integrating natural language interaction with image-based analysis have advanced rapidly, supported by a growing body of survey and review literature documenting their architectures, datasets, and applications across multiple fields. By contrast, archaeological applications of chatbots remain limited to text-based prototypes, primarily focused on education, cultural heritage mediation or archival search. This review synthesizes the historical development of chatbots, examines their current use in remote sensing, and evaluates the barriers to adapting such systems for archaeology. Four major challenges are identified: data scale and heterogeneity, scarcity of training datasets, computational costs, and uncertainties around usability and adoption. By comparing experiences across domains, this review highlights both the opportunities and the limitations of integrating conversational AI into archaeological workflows. The central conclusion is that domain-specific adaptation is essential if multimodal chatbots are to become effective analytical partners in archaeology.

Keywords: multimodal chatbots; conversational AI; remote sensing; drone photogrammetry; digital archaeology; human–AI collaboration; archaeological data analysis



Academic Editor: Fulong Chen

Received: 19 November 2025

Revised: 24 December 2025

Accepted: 2 January 2026

Published: 16 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Archaeology has consistently advanced in parallel with technological innovation, with each new tool reshaping the questions that could be asked and the scales at which they could be explored [1]. The adoption of computers in the 1960s introduced early computational methods into archaeological practice [2], while the widespread use of Geographic Information Systems (GIS) in the 1990s enabled large-scale spatial analyses and new forms of data integration [3,4]. These developments have been described as successive “digital revolutions” that fundamentally transformed the discipline [2,5]. More recently, the exponential increase in digital datasets has created what Grosman terms a “point of no return,” where archaeology can no longer function without computational infrastructures [2]. Within this broader digital turn, remote sensing has emerged as a decisive milestone, providing archaeologists with unprecedented access to high-resolution satellite imagery,

airborne LiDAR and drone-based photogrammetry [2,6]. These techniques have revealed hidden landscapes, identified previously undetectable sites, and enabled investigations at regional scales. At the same time, they have generated vast and heterogeneous datasets, the interpretation of which requires significant technical expertise and computational resources [2,4].

In parallel with these technological transformations in archaeology, conversational systems—commonly referred to as chatbots—have undergone a profound evolution over the past six decades. The earliest chatbot, ELIZA, was developed by Joseph Weizenbaum at MIT in 1966 and simulated by a Rogerian psychotherapist by applying pattern-matching and substitution rules to user inputs [7,8]. Despite its simplicity, ELIZA demonstrated that limited linguistic cues could create the illusion of understanding, giving rise to what later became known as the “ELIZA effect.” The canonical CACM publication provides a detailed description of ELIZA’s pattern-matching logic and early human–computer dialogue behavior [8]. A few years later, PARRY (1972), created by Kenneth Colby, introduced rudimentary modelling of beliefs and emotions, representing one of the first attempts to simulate human reasoning [9]. Colby’s later work expanded this approach by introducing affective modelling and belief systems in conversation design [9]. These pioneering systems were followed by increasingly complex rule-based and retrieval-based models through the 1980s and 1990s, including Jabberwacky, ALICE, and Mitsuku, which used pattern databases and hand-crafted scripts to sustain conversation [10].

The emergence of statistical natural language processing in the early 2000s marked a decisive turning point. Data-driven models enabled chatbots to move beyond fixed templates and to learn from corpora of real dialogues [11,12]. With the introduction of deep learning and transformer architectures, modern conversational systems such as OpenAI’s GPT series [13], Google’s LaMDA [14], and Microsoft’s XiaoIce [15] now integrate context, memory, and multimodal capabilities that combine text, image, and speech [12,16]. This rapid progress has positioned chatbots as central tools across multiple sectors, including healthcare, education, environmental monitoring, and customer support [17].

However, despite this acceleration, applications in archaeology remain limited. Most existing archaeological chatbots such as the CHATBot developed by Casillo et al. [18] are text-based, designed for visitor engagement or archival retrieval rather than analytical interpretation of imagery. By contrast, in domains such as remote sensing, geospatial analysis, and environmental monitoring, multimodal conversational agents have already begun to assist in data analysis and interpretation. This discrepancy underscores a key gap: while the technology to integrate natural language interaction with image-based analysis exists, its adaptation to archaeological remote sensing has yet to be achieved.

The intersection of these two technological trajectories, remote sensing and conversational artificial intelligence, represents a new frontier for digital archaeology. Remote sensing has transformed archaeological discovery by revealing landscapes at unprecedented scales [2,6], while chatbots have reshaped human–computer interaction through natural language communication [10,15]. However, no integrated framework currently exists that combines these capabilities to support archaeological analysis. Existing archaeological chatbots remain text-based, designed for education or heritage communication [19–21]. In contrast, multimodal systems capable of processing imagery have advanced rapidly in other fields such as agriculture, environmental monitoring, and geospatial analysis [22–24]. This disparity reveals a clear research gap, since archaeology has not yet benefited from multimodal conversational agents that can analyze remote sensing imagery through dialogue.

The purpose of this review study is to synthesize the state of the art in chatbots for remote sensing and archaeology, to evaluate the challenges that constrain their adaptation, and to highlight opportunities for developing domain sensitive multimodal conversational

systems. By bridging these disciplines, this paper aims to demonstrate how conversational artificial intelligence can lower technical barriers, democratize access to analytical tools and enhance archaeological interpretation within the broader digital transformation of the field.

2. History of Chatbots

The concept of human–machine conversation has developed over more than half a century and reflects the broader evolution of artificial intelligence research. The earliest system regarded as a chatbot was ELIZA, created by Joseph Weizenbaum at the Massachusetts Institute of Technology in 1966 [8]. ELIZA simulated a Rogerian psychotherapist through simple pattern matching and text substitution, producing short replies that mirrored the user’s own statements. Although the system lacked any real semantic understanding, it demonstrated that minimal linguistic cues were sufficient to create the illusion of dialogue, a phenomenon later termed the ELIZA effect. The interaction style of the original ELIZA system is illustrated in Figure 1.

```

Welcome to
                EEEEE LL      IIII  ZZZZZZ  AAAAA
                EE      LL      II     ZZ    AA   AA
                EEEEE LL      II     ZZ    AAAAAA
                EE      LL      II     ZZ    AA   AA
                EEEEE LLLLLL  IIII  ZZZZZZ  AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:

```

Figure 1. Example dialogue from Joseph Weizenbaum’s ELIZA program (1966), demonstrating the simulated psychotherapy interaction that gave rise to the “ELIZA effect.” [25].

A few years later, Kenneth Colby designed PARRY (1972), which attempted to model human reasoning by representing beliefs, goals and emotions [26]. In controlled tests, psychiatrists were sometimes unable to distinguish between transcripts of PARRY and those of real patients. These early systems established the foundation for rule-based conversation engines that dominated chatbot design for several decades. This is illustrated in Figure 2, which shows a sample interaction between a psychiatrist and PARRY.

During the 1980s and 1990s, improvements in computing power and data storage encouraged the creation of more sophisticated pattern matching chatbots such as Jabberwacky, ALICE, and Mitsuku, which employed extensive databases of predefined responses [10,27]. Despite their limitations, these systems popularized conversational interfaces and inspired the first public chatbot competitions, including the Loebner Prize.

Psychiatrist: Good afternoon. My name is Doctor Jones.
 PARRY: Pleased to meet you.
 Psychiatrist: How are you?
 PARRY: All right I guess.
 Psychiatrist: Could you tell me your name?
 PARRY: Pat Smith

Figure 2. Interface example and simulated dialogue from PARRY (1972), an early chatbot developed by Kenneth Colby to emulate a patient with paranoid schizophrenia. The image illustrates the system's belief–emotion reasoning approach to conversation [9].

The emergence of statistical natural language processing in the early 2000s marked a turning point. Machine learning techniques enabled chatbots to move beyond static scripts and to learn from corpora of real dialogues [15]. The introduction of deep learning and transformer architectures further advanced the field, leading to models that can generate coherent, context aware dialogue. Contemporary systems such as OpenAI's GPT series, Google's LaMDA, and Microsoft's XiaoIce integrate context, memory and multimodal capacities that combine text, image, and speech [12,16]. This rapid evolution has expanded the role of chatbots far beyond information retrieval, enabling their application in healthcare, education, environmental monitoring, and customer support [17]. This technological progression from early rule-based systems to contemporary multimodal conversational models is summarized in Figure 3.

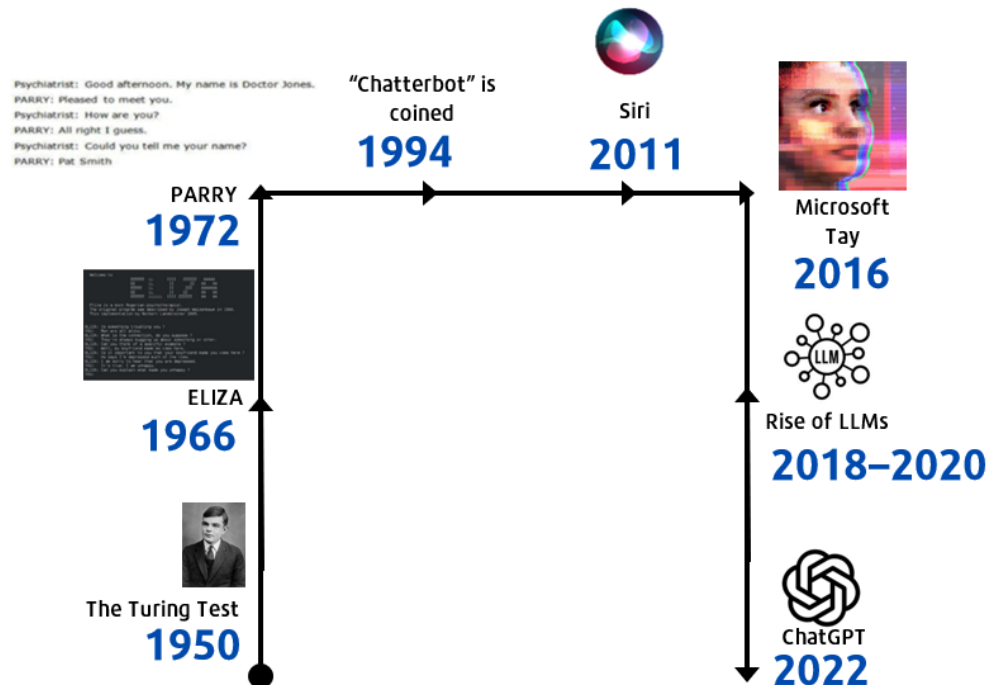


Figure 3. Evolution of chatbots and conversational artificial intelligence, from early rule-based systems such as ELIZA to contemporary domain-specific and multimodal conversational models, providing conceptual context for their potential application in archaeology and remote sensing.

2.1. Foundations and Architectures of Chatbots

Chatbots can be broadly divided into two principal categories: rulebased systems and datadriven or machinelearning systems.

Rulebased chatbots rely on predefined scripts, patternmatching templates and decision trees to produce answers that follow specific conversation paths [10]. Rulebased chatbots operate through explicitly defined linguistic patterns, where every possible user input is matched to a predefined template stored in the system's knowledge base. These early systems often relied on keyword detection and simple decision trees to trigger scripted responses using frameworks such as AIML (Artificial Intelligence Markup Language) [11]. Because the dialogue logic was entirely deterministic, these chatbots could simulate coherent exchanges within narrow domains but failed when queries deviated from expected inputs. Despite their rigidity, rule-based models laid the foundation for human-machine dialogue by formalizing the structure of conversational turns and illustrating the feasibility of computer mediated communication [8,26,28]. Although predictable and easy to implement, such systems are predictable and easy to implement but limited in their ability to manage unstructured queries. Early examples include ELIZA and PARRY, which applied deterministic rules to generate responses without real semantic processing [8,26]. Although they demonstrated that humanlike conversation was technically achievable, these chatbots were unable to interpret context or adapt to new topics.

Data driven approaches emerged with advances in natural language processing and statistical learning. These methods represent a fundamental shift from handcrafted linguistic rules to systems that learn conversational patterns directly from data. Instead of relying on predefined templates, these models use large dialogue corpora to infer statistical regularities in language. Early frameworks employed probabilistic methods such as Hidden Markov Models (HMMs) and ngram models to estimate the likelihood of word sequences, forming the basis for speech recognition and intent detection. Subsequent developments in machine learning introduced sequence to sequence architectures capable of mapping user inputs to responses through supervised training. This paradigm enabled chatbots to generalize beyond fixed scripts and generate more contextually relevant answers, marking the transition from symbolic to statistical natural language understanding [12,15]. In the 2000s, probabilistic models such as Hidden Markov Models and sequence to sequence networks allowed systems to learn linguistic structures directly from dialogue corpora [15]. These techniques were further refined through deep learning, which enabled models to represent language as high dimensional embeddings that capture syntactic and semantic relations [29]. Transformer-based architectures revolutionized natural language processing by introducing a self attention mechanism that enables models to weigh the importance of each word in relation to all others in a sentence [30]. Unlike recurrent or convolutional models, transformers process entire sequences in parallel, allowing them to capture long range dependencies and contextual nuances more efficiently [12]. This design underpins large language models (LLMs) such as GPT3, GPT4, LaMDA and XiaoIce, which are trained on massive text corpora containing billions of tokens the basic computational units of language that represent words, sub words or punctuation. During training, each token is converted into a vector in a high dimensional embedding space, allowing the model to learn grammatical relationships, semantic meaning and contextual patterns between them. The greater the number of tokens a model is exposed to, the richer and more generalizable its linguistic representation becomes, though this also increases computational cost and memory requirements. Consequently, transformer-based chatbots achieve remarkable fluency and contextual understanding but face ongoing challenges of bias, interpretability and efficiency [16,17,31].

While the transition from rule-based to transformer-based architectures represents a major technological leap, it also redefines the notion of a chatbot itself. Modern conversational agents are no longer fixed question-answer tools but adaptive systems capable of reasoning, content generation and multimodal interaction. This transformation forms the

technical basis upon which new domain specific applications, including those for remote sensing and archaeology, are now being developed.

Table 1 summarizes the main technological stages in chatbot evolution, from rule-based pattern matching to transformer-based multimodal systems. The comparison outlines how each generation introduced new capabilities—shifting from scripted dialogue to contextual reasoning and adaptive language generation—while also revealing persistent limitations such as restricted contextual understanding and increasing computational demands. Overall, the table highlights the steady transition from deterministic to data driven and pretrained models that define modern conversational AI.

Table 1. Evolution and Characteristics of Chatbot Architectures.

Generation	Period	Core Technology	Presentative Systems	Main Capabilities	Key Limitations	References
Rule-based	1960s–1990s	Pattern matching, decision trees	ELIZA [8], PARRY [26], ALICE [10]	Predefined dialogue, basic text substitution	No contextual understanding, static rules	[8,10,26]
Statistical/NLP-based	2000s	Probabilistic models, ngram, HMM	Early FAQ bots, retrieval models	Learned responses from corpora, improved relevance	Limited coherence, domain specific tuning required	[15]
Deep learning/Seq2Seq	2010s	Recurrent neural networks, attention	Early neural dialogue models	Contextual learning, adaptive phrasing	Poor long term memory, training data dependency	[12,15]
Transformer-based LLMs	2018–present	Self attention, largescale pretraining	GPT3/4 [13], LaMDA [13], XiaoIce [12]	Contextaware, multiturn dialogue, multimodal input	High computational cost, potential bias	[16,17]

2.2. Conversational Artificial Intelligence for Remote Sensing Analysis

Remote sensing generates vast and complex datasets that are often difficult for nonspecialists to interpret [32,33]. The increasing availability of high resolution satellite imagery, LiDAR surveys and drone photogrammetry has transformed how environmental and archaeological landscapes are analyzed, yet the diversity and scale of the data create substantial barriers for researchers who lack advanced computational expertise [23]. Recent advances in conversational artificial intelligence and large language models (LLMs) have introduced new possibilities for bridging this gap. Chatbots equipped with natural language interfaces can translate user queries into data processing tasks, enabling interaction with imagery and geospatial information through simple text input [34].

Early experiments such as LLMFind treated the language model as a decision making core rather than a text generator [34]. The system was able to retrieve geospatial data, generate code snippets for image processing and visualize outputs in real time through a QGIS plugin. A similar autonomous GIS agent [35] demonstrated end to end automation within QGIS, integrating natural language queries and workflow execution [35]. This approach demonstrated that conversational agents could operate as workflow managers, automating the technical steps between a user’s question and the resulting dataset.

The introduction of multimodal systems expanded these capabilities further. Visual ChatGPT, developed by Osco et al. [23], combined text and image processing in a single interactive framework, allowing users to upload imagery and request operations such as

edge detection, segmentation, or classification. Similarly, ChangeChat by Deng et al. [23] was trained on more than 87,000 instruction–response pairs to describe and quantify changes between two remote sensing images. ChangeChat constitutes the first bitemporal vision–language model designed specifically for interactive change analysis through multimodal instruction tuning [36]. An independent evaluation by Osco et al. (2023) [23] in Remote Sensing confirmed its ability to perform edge detection and segmentation on domain specific datasets. Complementary work by Pierdicca et al. demonstrated how large language models can function as natural language interfaces for translating user queries into GIS-based spatial analysis workflows, illustrating the potential of domain-specific task-oriented designs [37]. The model outperformed general purpose LLMs in binary classification and change detection accuracy, illustrating the potential of domain specific finetuning. Representative qualitative results from geospatially grounded conversational models are illustrated in Figure 4.

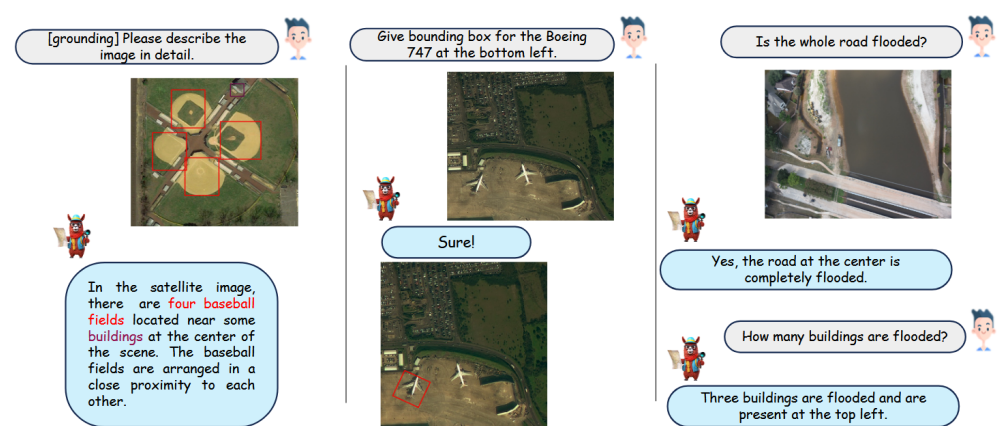


Figure 4. Qualitative results of GeoChat [22]. Results are shown on grounding, referring object detection, and disaster/damage detection. The user can provide task-specific tokens (e.g., [grounding]) to shape model responses according to the desired behavior. Red bounding boxes indicate regions identified by the model during grounding and referring object detection tasks.

Recent developments have also focused on geospatially grounded models such as GeoChat, which integrates large vision–language models for remote sensing question answering and image captioning [38]. GeoChat, presented recently at CVPR 2024, enables region specific dialogue and spatial reasoning on high resolution imagery, integrating grounded visual tokens for location aware responses [38]. These systems demonstrate that conversational AI can interpret complex visual data, summarize spatial patterns and even generate analytical outputs in accessible language. Complementary research by Pierdicca et al. [37] explored the use of large language models as natural language interfaces for translating plain language queries into GIS-based spatial analysis workflows. In parallel, Labba et al. [39] introduced IArch, an AI-based system designed to support interactive exploration and structured analysis of archaeological datasets through intelligent data access mechanisms. In a related study, Rane et al. [40] showed that ChatGPT could be employed to interpret remote sensing and GIS datasets on demand, summarizing temporal and spatial changes directly from imagery. Task oriented chatbots embedded in geoportals offer an HCl-centric approach that exposes persistent metadata and interoperability constraints [41].

Collectively, these examples reveal a clear trajectory in the field of remote sensing. Conversational agents have evolved from simple information brokers to multimodal assistants capable of performing spatial analyses and interpreting visual data. This evolution provides the conceptual and technical foundation for adapting similar approaches to archaeology, where comparable challenges of data scale, heterogeneity and accessibility

persist. Beyond individual system implementations, recent surveys and reviews demonstrate that these efforts form part of a broader and rapidly evolving research landscape in multimodal conversational AI and vision–language models, supported by diverse datasets, architectures, and evaluation frameworks across domains such as remote sensing and geospatial analysis.

While the systems discussed above primarily target analytical interaction with remote sensing and geospatial data, conversational agents have also been explored in educational and outreach-oriented contexts within archaeology. In these cases, chatbots are designed less as analytical tools and more as interfaces for communicating archaeological knowledge to non-specialist audiences, including students, visitors, and the general public.

2.3. Artificial Intelligence Applications in Archaeology

Artificial intelligence has been integrated into archaeological research primarily through machine learning, computer vision, and predictive modelling approaches rather than through conversational systems. Over the past two decades, AI-driven geospatial analysis has become a central component of archaeological prospection, landscape archaeology, and heritage management, particularly in the context of remote sensing and GIS-based workflows [2,4,42]. Neural networks, ensemble models, and probabilistic predictive frameworks have been widely applied to site detection, artefact classification, landscape pattern analysis, and modelling of human–environment interactions, often producing results that directly inform field strategies, excavation planning, and regional research agendas [42,43].

These approaches have demonstrated substantial impact across diverse archaeological contexts, enabling the analysis of large-scale spatial datasets that would be impractical to interpret manually. Machine learning techniques have been used to identify subtle anthropogenic features in satellite imagery and LiDAR data, to detect microtopographic anomalies, and to model settlement patterns across different environmental settings [23,42,44]. In many cases, these methods have shifted archaeological practice from opportunistic discovery toward hypothesis-driven and probabilistic exploration of landscapes. Compared to conversational systems, such AI methods have achieved broader disciplinary uptake and are now widely recognized as foundational tools within computational and geospatial archaeology [2,4].

By contrast, conversational systems in archaeology have so far seen limited uptake and remain largely experimental. Existing chatbot-based applications have largely focused on educational, outreach, or information-retrieval purposes, such as supporting visitor engagement at heritage sites or facilitating access to textual archives [19–21]. While these systems demonstrate the feasibility of natural language interaction with archaeological knowledge, they are best understood as experimental or proof-of-concept implementations rather than as general-purpose analytical tools. Their development has typically been project-specific, and they have not been widely embedded within routine research workflows or employed for systematic interpretation of remote sensing imagery [36,45].

The limited impact of chatbots in archaeology does not reflect a lack of technical potential, but rather a misalignment between conversational interfaces and the established analytical practices of the discipline. Archaeological reasoning commonly relies on the integration of spatial imagery, environmental data, and contextual textual evidence, a process that has so far been more effectively supported by machine learning pipelines and GIS-based analytical environments than by dialogue-based systems [2,42]. As a result, conversational agents currently occupy a peripheral position within archaeological AI ecosystems, functioning more as access points to information than as active components of analytical reasoning.

Nevertheless, recent advances in multimodal conversational systems in the broader field of remote sensing point toward a possible future trajectory for archaeology. Rather than competing with established AI methods, chatbots may act as mediating interfaces that link users to existing analytical infrastructures, translating archaeological questions into computational operations and lowering technical barriers to complex workflows [34,38,40]. From this perspective, the relevance of chatbots lies less in their current disciplinary impact and more in their potential to complement mature AI-driven geospatial methods by supporting more accessible, human-centred engagement with archaeological data and interpretative processes.

While these examples illustrate early and exploratory applications of conversational systems in archaeological contexts, they should not be interpreted as representative of widespread or mature disciplinary practice [19–21]. Rather than continuing with a case-by-case discussion, the following section deliberately shifts to a broader analytical perspective, situating chatbots within the wider landscape of artificial intelligence methods that have shaped archaeological research more substantially over the past two decades [2,4].

To synthesize the diverse ways in which conversational systems have been explored within archaeological contexts, Table 2 provides a conceptual overview of key chatbot categories, functionalities, and design characteristics. Rather than enumerating individual project-specific implementations, the table abstracts recurring patterns and roles of conversational agents across educational, outreach, information-retrieval, and research-oriented applications. This schematic perspective supports a comparative understanding of how chatbots have been positioned within archaeological practice while avoiding overemphasis on isolated or experimental case studies.

Table 2. Artificial intelligence approaches in archaeology and their analytical roles.

AI Approach Category	Primary Data Types	Typical Archaeological Applications	Level of Disciplinary Adoption	Analytical Role
Machine learning and predictive modelling	Spatial, environmental, remote sensing data	Site prediction, landscape modelling, settlement pattern analysis	High	Core analytical methodology
Computer vision and deep learning	Satellite imagery, LiDAR, photogrammetry	Feature detection, classification, automated mapping	Increasing	Automated interpretation
Rule-based and retrieval-oriented chatbots	Textual archives, metadata	Education, heritage communication, information access	Low	Interface and access support
Multimodal conversational systems	Text and imagery	Experimental analytical mediation, exploratory interaction	Emerging	Potential future interface

Summarizes the main categories of AI approaches in archaeology and their analytical roles.

3. Chatbots in Remote Sensing

The integration of conversational artificial intelligence into remote sensing workflows represents a growing area of research. This section examines how remote sensing technologies operate across different scientific domains and explores how these methods have been adopted in archaeological contexts. Section 3.1 provides an overview of remote sensing fields and analytical methods, while Section 3.2 focuses specifically on their application to archaeology and the emerging role of conversational systems in supporting image-based interpretation.

3.1. Remote Sensing Fields

Remote sensing has become one of the most powerful sources of spatial and environmental information, providing continuous coverage of the Earth's surface through multispectral, hyperspectral, radar and LiDAR sensors [4,6,46]. Recent syntheses emphasize that Earth observation has entered a bigdata era increasingly mediated by AI assisted interpretation [47,48]. The integration of these technologies enables the detection of physical and chemical characteristics of materials, the monitoring of land use changes and the analysis of vegetation indices that support ecological and agricultural research [24,40]. Over the past two decades, the combination of satellite, airborne and drone-based platforms has made it possible to collect high resolution imagery across large temporal and spatial scales, transforming how landscapes are analyzed [2].

Traditional remote sensing workflows involve a sequence of steps such as image preprocessing, segmentation, feature extraction and classification [49]. These procedures require both computational resources and specialized expertise, particularly when dealing with multisource data of varying resolutions [2,47,50]. The rapid increase in the availability of open access imagery from missions such as Landsat, Sentinel and World View has further intensified the need for automated processing pipelines [6,51]. Consequently, artificial intelligence has become a cornerstone of modern remote sensing analysis. Deep learning models such as convolutional neural networks (CNNs) and transformer architectures have been applied for classification, object detection and change detection tasks with notable success [23,24,29]. Complementary studies explore ChatGPT-3.5 driven interpretation pipelines that automate classification and reporting within RS and GIS workflows [40,52].

Recent research has also introduced conversational and multimodal frameworks that integrate natural language interaction with image-based processing. Examples include the LLMFind system [34], which allows for geospatial data retrieval via natural language queries, and GeoChat [38], a large vision-language model capable of interpreting satellite images and answering domain specific questions. Similarly, Visual ChatGPT [24,53] and ChangeChat [23] combine image recognition with dialogue-based reasoning, enabling users to perform segmentation and change analysis without manual coding. These developments demonstrate a growing convergence between remote sensing, machine learning, and conversational artificial intelligence [39,40].

As a result, remote sensing is no longer limited to data collection and image analysis but is evolving towards interactive systems that can understand context, summarise results and communicate them through natural language [17]. This progression not only enhances accessibility for nonspecialists but also provides a conceptual foundation for extending such approaches into other data intensive domains, including archaeology. The following subsection explores how remote sensing has been adopted within archaeological research and how its integration with conversational systems can further transform data interpretation.

3.2. Remote Sensing in Archaeology

Remote sensing has transformed archaeological practice by expanding the spatial and temporal scales at which past human activities can be observed and analyzed. The integration of satellite imagery, airborne LiDAR, multispectral data and drone-based photogrammetry allows archaeologists to detect features that are invisible at ground level and to reconstruct past landscapes with unprecedented precision [2,6]. Applications include site detection, mapping of settlement systems, monitoring of environmental change and the preservation of cultural heritage [42,46]. These technologies enable the identification of subtle anthropogenic signatures such as cropmarks, soil discoloration and microtopographic variations, which often indicate buried structures or ancient land use patterns [19].

The introduction of open access remote sensing datasets and cloud-based analytical tools has democratized archaeological investigation, allowing a wider community of researchers to access and process geospatial information [4,24]. However, the heterogeneity of data formats and the technical complexity of image processing continue to restrict broader adoption. Empirical evaluation of Visual ChatGPT in RS indicates that similar domain gaps could affect archaeological imagery interpretation [24]. Many archaeological projects rely on collaborations with remote sensing specialists, and the analytical results are often mediated through specialized software environments such as QGIS, ArcGIS or Python-based workflows [2,41], which do not explicitly report software version numbers. These requirements can create barriers for smaller institutions or projects that lack computational resources.

Recent research efforts have sought to overcome these limitations through automation and artificial intelligence. Machine learning and deep learning models have been applied to classify archaeological features, detect structures in LiDAR point clouds and identify patterns in satellite imagery [24,42]. The human–AI collaboration model proposed by Casini et al. [42] illustrates how interactive systems can combine human interpretative expertise with computational efficiency to improve site detection accuracy [43]. Despite these advances, most archaeological analyses still depend on expert driven interpretation rather than conversational or interactive frameworks.

The emergence of conversational artificial intelligence introduces a new opportunity to close this gap. Chatbots and multimodal systems capable of processing imagery and text could enable archaeologists to query datasets using natural language, simplifying interaction with complex digital infrastructures [39,40]. For instance, an archaeologist could ask a chatbot to identify possible enclosures in a LiDAR hill shade or to summarize landscape change across multiple temporal scenes. Such interaction would lower the technical threshold for data analysis and enhance accessibility without compromising scientific rigor. Grounded VLMs such as GeoChat demonstrate that image-based question answering can be applied to cultural landscape scenes [22].

In summary, remote sensing has become an indispensable component of modern archaeology, enabling largescale landscape reconstruction and long-term environmental analysis. The challenge that remains is to integrate these powerful techniques with conversational systems that can translate archaeological questions into analytical operations [34]. This convergence represents the conceptual foundation for the next stage of development, where multimodal chatbots could act as intelligent mediators between remote sensing data and archaeological interpretation.

4. Current Challenges and Research Gaps

Despite rapid advances in conversational artificial intelligence and remote sensing technologies, the integration of multimodal chatbots into archaeological workflows remains limited [2,4,24]. This section examines the key technical, methodological, and ethical challenges that constrain their adoption and identifies critical research gaps that must be addressed to enable meaningful and responsible use within archaeology [54].

4.1. Challenges

The integration of conversational artificial intelligence into archaeological remote sensing faces several structural and methodological challenges. The first concerns the scale and heterogeneity of remote sensing data per se. Archaeology increasingly depends on multisource imagery acquired from satellites, drones and LiDAR systems, which differ in spatial resolution, spectral range and data format [2,4,6]. These variations complicate data harmonization and require advanced preprocessing techniques, such as radiometric

and atmospheric corrections, which extend beyond typical interpretative workflows and are primarily relevant when imagery is integrated into automated or AI-driven analytical pipelines. The limited standardization of metadata and the partial interoperability of existing archives further restrict the ability of conversational systems to retrieve and interpret information reliably at scale [39].

A second challenge involves the limited availability of curated training datasets suitable for archaeological applications [55]. Unlike domains such as agriculture or environmental monitoring, archaeology has limited availability of standardized, annotated image corpora specifically designed to support the training of multimodal vision–language models [24,42]. Existing datasets are often small, site-specific, and created through manual labelling for particular research objectives, which constrains their scalability and generalization for machine learning applications. Without sufficient examples of archaeological features, conversational agents struggle to learn the subtle differences between natural and anthropogenic patterns. It is important to note that this limitation does not reflect a general lack of accessible remote sensing data in archaeology, but rather the additional requirements involved in transforming such data into curated training sets suitable for AI-driven analysis.

Another important limitation concerns computational costs and infrastructure requirements. High resolution imagery and LiDAR point clouds are computationally demanding, often requiring powerful hardware and specialized software environments [24]. Running large vision–language models such as Visual ChatGPT or ChangeChat in real time involves high memory usage and processing power, which may exceed the computational resources available in many archaeological research settings [23,40]. Cloud-based platforms can mitigate these issues, yet they raise concerns about sustainability, cost and long-term accessibility.

Finally, there are challenges related to usability, transparency and adoption. Archaeological workflows are shaped by disciplinary traditions and interpretative practices that may not easily adapt to automated systems [1]. Users must understand the limitations of conversational models and retain interpretative control over analytical outcomes through transparent workflows. The success of AI-assisted archaeology therefore depends not only on technological maturity, but also on the alignment of conversational systems with established archaeological workflows through intuitive interfaces, transparent algorithms, and interdisciplinary collaboration [50,56].

4.2. Gaps and Research Opportunities

This review of the existing literature reveals a clear asymmetry between advances in remote sensing and their practical adoption within archaeology. While multimodal chatbots have achieved remarkable progress in other scientific domains, archaeological applications remain primarily text oriented [44,57]. This gap results from both technical and disciplinary constraints. The absence of large, well annotated datasets limits the ability of models to recognize the subtle morphological and spectral characteristics of archaeological features [2,42]. Collaborative initiatives that promote open data sharing and standardized annotation could help overcome this limitation and provide the foundation for training domain sensitive conversational agents [4,6]. A recent technical note on archaeological object detection using airborne LiDAR identifies priorities for developing scalable and transferable workflows for automated feature recognition and site mapping [23,44]. These approaches emphasize the need for reproducible pipelines that can adapt to different landscapes and data resolutions, facilitating broader application across archaeological case studies.

Another major research opportunity lies in the integration of image-based and textual data. Archaeological reasoning typically relies on the combination of spatial imagery, excavation reports and historical records, yet most existing systems treat these sources separately [21]. Future research should aim to develop hybrid frameworks that allow conversational agents to interpret visual evidence while simultaneously linking it to contextual information. Such integration would enable archaeologists to query both images and texts through natural language, resulting in richer and more interpretatively coherent analyses [24,39].

There is also a need for lightweight and resource efficient prototypes that can operate under the computational constraints typical of archaeological institutions. Models such as Visual ChatGPT, which integrates ChatGPT with task-specific vision models without specifying a fixed language model version [24], and ChangeChat, which is built on the Vicuna-v1.5 large language model backbone [23], demonstrate the technical feasibility of multimodal analysis, although their high computational costs limit scalability. Optimizing these systems for smaller datasets and local execution environments would increase accessibility and sustainability, especially for projects in regions with limited infrastructure [4].

Finally, consideration must be given to usability and disciplinary adoption. Archaeologists should remain active participants in the design and evaluation of conversational systems to ensure that AI tools complement, rather than replace, expert interpretation [19,20]. Participatory design, transparency and explainability should therefore guide future development. Evaluations should not be confined to technical accuracy but should also assess whether the system genuinely supports archaeological reasoning and decision making.

Overall, the synthesis of these research gaps points to a clear direction for future work: the creation of domain specific, multimodal conversational agents capable of analyzing imagery, integrating textual context and operating within the interpretative frameworks of archaeology. Such systems have the potential to transform how archaeological data are explored, interpreted and shared, fostering a new generation of digital collaboration between humans and intelligent machines.

4.3. Ethical Considerations and Data Governance

The increasing integration of artificial intelligence into archaeological research raises ethical concerns that extend beyond technical feasibility and methodological integration. At a general level, debates on AI ethics emphasize issues of data provenance, intellectual property, transparency, and accountability, particularly in relation to systems trained on large and heterogeneous corpora of visual and textual data [58,59]. Multimodal conversational systems intensify these concerns [60] by enabling natural language access to analytical outputs, potentially obscuring the origins, assumptions, and biases embedded within training datasets and algorithmic decision-making processes [61].

Within archaeology, these ethical challenges acquire heightened significance due to the sensitive nature of archaeological and spatial information [62]. Machine learning and vision-language models capable of detecting, classifying, or mapping archaeological features from remote sensing imagery rely on training datasets that may contain precise locational data. While such information is essential for model performance, its integration into conversational interfaces risks lowering barriers to access in ways that may inadvertently facilitate site looting, illicit excavation, or unsustainable tourism [63,64]. The ability of users to query AI systems for site identification or spatial patterning from imagery underscores the need for careful control over data exposure and output resolution.

These risks are particularly acute in relation to Indigenous heritage and culturally sensitive landscapes. Not all archaeological sites are intended for public disclosure, and unrestricted dissemination of spatial knowledge may conflict with Indigenous data sovereignty,

community values, and ethical research practices, particularly in relation to sacred places and culturally restricted landscapes [65–67]. From this perspective, the deployment of conversational AI systems raises fundamental questions regarding authority over archaeological knowledge, responsibilities toward descendant communities, and the long-term implications of automated discovery and disclosure.

Addressing these concerns requires an ethics-by-design approach in which considerations of data sensitivity, cultural authority, and potential misuse are embedded into system architectures and governance frameworks from the outset. Practical measures may include spatial obfuscation of sensitive locations, tiered access models based on user roles, and the co-design of AI systems in collaboration with local and Indigenous communities [68,69]. Without such safeguards, the same AI technologies that promise analytical innovation risk undermining established principles of archaeological stewardship and responsible management of cultural heritage.

In addition to these concerns, the phenomenon of so-called “hallucinations” in generative and multimodal AI systems constitutes a critical epistemic risk for archaeological interpretation. Large language and vision–language models may generate plausible but incorrect descriptions, associations, or spatial inferences that are not grounded in the underlying data, potentially leading to misinterpretation of archaeological evidence [69,70]. In archaeology, where interpretation is inherently probabilistic and context-dependent, such errors risk reinforcing unfounded hypotheses or obscuring uncertainty rather than clarifying it. This reinforces the necessity of maintaining human oversight and expert validation in any AI-assisted analytical workflow, ensuring that conversational systems remain interpretative aids rather than authoritative sources [58]. Addressing hallucinations through transparency, uncertainty communication, and human-in-the-loop design is therefore essential for the responsible adoption of generative AI in archaeological research.

5. Discussion

This synthesis of current research indicates that the convergence of conversational artificial intelligence and remote sensing has the potential to reshape the analytical landscape of archaeology. Remote sensing has already become an essential source of archaeological evidence, allowing researchers to reconstruct past environments, detect buried features and monitor cultural landscapes at multiple scales [2,4,6]. An editorial overview in *Remote Sensing* (2024) frames this convergence as a defining trend for AI-mediated Earth observation [59]. The recent introduction of large vision–language models and multimodal chatbots represents a logical extension of this digital transformation, offering new forms of interaction with complex geospatial data [23,24]. However, the practical implementation of such systems within archaeology remains in its infancy.

Several factors contribute to this situation. Technical complexity and the lack of standardized infrastructures hinder the smooth integration of AI systems into archaeological workflows [39]. Data related challenges, such as the limited availability of annotated training sets and the heterogeneity of remote sensing imagery, restrict the performance of multimodal chatbots [19,42,60]. Equally significant are cultural and methodological issues. Archaeological interpretation depends on human expertise, contextual reasoning and theoretical frameworks that are difficult to replicate algorithmically [19]. Therefore, the introduction of conversational systems should be understood as a form of augmentation rather than automation.

Future research should focus on the codesign of tools that align computational efficiency with disciplinary practice. Lightweight models capable of local deployment could address the resource limitations of smaller institutions, while hybrid frameworks linking imagery and textual archives would better reflect the nature of archaeological reason-

ing [40]. The ultimate challenge is to create systems that are not only technically proficient but also epistemologically compatible with archaeology's interpretative character. Achieving this balance will define the success of conversational AI as an assistive partner in the evolving ecosystem of digital archaeology. The maturity of multimodal conversational systems in other scientific domains [55–59] provides a comparative horizon for archaeology. Rather than framing adoption solely as a technological challenge, this comparison invites reflection on what new forms of archaeological questioning such systems might enable. In particular, AI-mediated interaction could support exploratory engagement with uncertainty, facilitate the dynamic integration of visual and textual evidence, and help articulate alternative hypotheses about past human activity, extending archaeological inquiry beyond efficiency-driven applications.

6. Conclusions

This review examined the intersection of conversational artificial intelligence and remote sensing in archaeology, highlighting the opportunities and challenges that arise from their integration. The analysis showed that while multimodal chatbots have demonstrated strong potential in fields such as environmental monitoring and geospatial analysis [23,24], their adoption in archaeology remains limited to text-based applications [19,20,40]. The primary barriers include data heterogeneity, the absence of curated training datasets, high computational requirements and uncertainties surrounding usability and disciplinary acceptance [2,4,39,42]. Addressing these challenges will require collaborative initiatives for dataset creation, the development of lightweight multimodal frameworks and participatory design processes involving archaeologists. The combination of conversational interfaces with remote sensing workflows has the potential to democratize access to complex imagery, enhance interpretative capacity and foster new forms of human–AI collaboration. By aligning technological innovation with archaeological practice, conversational AI can evolve from a communication tool into a genuine analytical partner that supports the future of digital archaeology.

Future research should extend beyond proof-of-concept studies to the development of operational multimodal chatbots capable of analyzing both spatial and textual archaeological datasets in real time [53,56]. Integrating conversational AI with image recognition models, GIS workflows and cloud-based repositories could enable the automatic retrieval and interpretation of excavation records, satellite imagery and sensor data. A further priority is the creation of domain specific datasets and evaluation benchmarks that reflect the diversity of archaeological contexts. Collaborative platforms involving archaeologists, computer scientists and heritage institutions are essential to ensure transparency, interpretability and ethical compliance. Ultimately, advancing these directions will transform conversational systems from assistive interfaces into autonomous analytical partners that enhance archaeological reasoning and support evidence-based cultural heritage management.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, N.M. and A.A.; supervision, A.A.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the CONNECTING project. The CONNECTING project has received funding from the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation of Cyprus under contract SMALL SCALE INFRASTRUCTURES/1222/0062.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: This research forms part of the Ph.D. thesis of N.M., conducted at the EOcult: Earth Observation Cultural Heritage Lab (<https://web.cut.ac.cy/eocult/>, accessed on 23 July 2025), Department of Civil Engineering and Geomatics, Cyprus University of Technology. The author acknowledges the use of AI-based assistants (ChatGPT (version GPT5) by OpenAI) to improve the readability and grammar of the manuscript. The authors have thoroughly reviewed and revised all AI-generated text and take full responsibility for the content and interpretations presented in this publication.

Conflicts of Interest: The author declares no conflicts of interest. The views and opinions expressed are, however, those of the authors only.

References

- Djindjian, F. Archaeology and Computers: A Long Story in the Making of Modern Archaeology. *Archeol. Calc.* **2019**, *30*, 13–20.
- Grosman, L. Reaching the Point of No Return: The Computational Revolution in Archaeology. *Annu. Rev. Anthropol.* **2016**, *45*, 129–145. [\[CrossRef\]](#)
- Hetrick, V.; Leventhal, R.M.; Read, D.M. A Design to Use Computational Science for Archaeological Interpretation. *Comput. Graph.* **1992**, *26*, 179–181. [\[CrossRef\]](#)
- McManamon, F.P.; Kintigh, K.W. Transforming Archaeological Data into Knowledge. *SAA Archaeol. Rec.* **2010**, *10*, 37–40.
- Cobb, P.J. Large Language Models and Generative AI, Oh My! Archaeology in the Time of ChatGPT, Midjourney, and Beyond. *Adv. Archaeol. Pract.* **2023**, *11*, 363–369. [\[CrossRef\]](#)
- Agapiou, A.; Lysandrou, V. Interacting with the Artificial Intelligence (AI) Language Model ChatGPT: A Synopsis of Earth Observation and Remote Sensing in Archaeology. *Heritage* **2023**, *6*, 4072–4085. [\[CrossRef\]](#)
- Abu Shawar, B.; Atwell, E. Chatbots: Are They Really Useful? *J. Lang. Technol. Comput. Linguist.* **2007**, *22*, 29–49. [\[CrossRef\]](#)
- Weizenbaum, J. ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45. [\[CrossRef\]](#)
- Colby, K. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*; Pergamon Press: Oxford, UK, 1975.
- Shum, H.Y.; He, X.; Li, D. From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 10–26. [\[CrossRef\]](#)
- Vidya, G.; Vernekar, N.K.; Kelkar, G. Comparative Study on Chatbot Frameworks. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* **2023**, *11*, 3602–3610. [\[CrossRef\]](#)
- Adamopoulou, E.; Moussiades, L. An Overview of Chatbot Technology. In *Artificial Intelligence Applications and Innovations (AIAI 2020)*; Maglogiannis, I., Iliadis, L., Pimenidis, E., Eds.; Springer: Cham, Switzerland, 2020; pp. 373–383. [\[CrossRef\]](#)
- Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:2201.08239. Available online: <https://arxiv.org/abs/2201.08239> (accessed on 3 August 2025).
- Casillo, M.; Clarizia, F.; D’Aniello, G.; De Santo, M.; Lombardi, M.; Santaniello, D. CHATBot: A Cultural Heritage Aware TellerBot for Supporting Touristic Experiences. *Pattern Recognit. Lett.* **2020**, *131*, 234–243. [\[CrossRef\]](#)
- AlAmin, M.; Ali, M.S.; Salam, A.; Khan, A.; Ali, A.; Ullah, A.; Alam, M.N.; Chowdhury, S.K. History of Generative Artificial Intelligence (AI) Chatbots: Past, Present, and Future Development. *arXiv* **2024**, arXiv:2402.05122. [\[CrossRef\]](#)
- Chakraborty, C.; Pal, S.; Bhattacharya, M.; Dash, S.; Lee, S.S. Overview of Chatbots with Special Emphasis on Artificial Intelligence-Enabled ChatGPT in Medical Science. *Front. Artif. Intell.* **2023**, *6*, 1237704. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shrager, J. ELIZA Reinterpreted: The World’s First Chatbot Was Not Intended as a Chatbot at All. *arXiv* **2024**, arXiv:2406.17650.
- Lombardi, M.; Pascale, F.; Santaniello, D. An Application for Cultural Heritage Using a Chatbot. In Proceedings of the 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 1–3 May 2019.
- Godewithana, N.; Kirindage, G. Intelligent Hybrid Chatbot Solution for Archaeological Sites Tracking. In Proceedings of the Sixth International Conference on Inventive Computation Technologies (ICICT 2021), Coimbatore, India, 20–22 January 2021; pp. 517–522. [\[CrossRef\]](#)
- Pajmans, H.; Brandsen, A. Searching in Archaeological Texts: Problems and Solutions Using an Artificial Intelligence Approach. *PalArch’s J. Archaeol. Egypt/Egyptol.* **2010**, *7*, 1–6.
- Kuska, M.T.; Wahabzada, M.; Paulus, S. AIChatbots for Agriculture, Where Can Large Language Models Provide Substantial Value? *Comput. Electron. Agric.* **2024**, *221*, 108924. [\[CrossRef\]](#)
- Deng, P.; Zhou, W.; Wu, H. ChangeChat: An Interactive Model for Remote Sensing Change Analysis via Multimodal Instruction Tuning. In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025.

23. Osco, L.P.; de Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.M.; Marcato Junior, J. The Potential of Visual ChatGPT for Remote Sensing. *Remote Sens.* **2023**, *15*, 3232. [CrossRef]
24. Wikipedia Contributors. ELIZA. In Wikipedia, The Free Encyclopedia; Wikimedia Foundation, Last Edited 31 October 2025. Available online: <https://en.wikipedia.org/wiki/ELIZA> (accessed on 15 June 2025).
25. Adamopoulou, E.; Moussiades, L. Chatbots: History, Technology, and Applications. *Mach. Learn. Appl.* **2020**, *2*, 100006. [CrossRef]
26. Rudolph, J.; Tan, S.; Tan, S. War of the Chatbots: Bard, Bing Chat, ChatGPT, Ernie and Beyond. The New AI Gold Rush and Its Impact on Higher Education. *J. Appl. Learn. Teach.* **2023**, *6*, 364–389. [CrossRef]
27. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
28. Dritsas, E.; Trigka, M. Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey. *Remote Sens.* **2025**, *17*, 550. [CrossRef]
29. Fiorucci, M.; Verschoofvan der Vaart, W.B.; Soleni, P.; Le Saux, B.; Traviglia, A. Deep Learning for Archaeological Object Detection on LiDAR: New Evaluation Measures and Insights. *Remote Sens.* **2022**, *14*, 1694. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
31. Ning, H.; Li, Z.; Akinboyewa, T.; Lessani, M.N. An Autonomous GIS Agent Framework for Geospatial Data Retrieval. *Int. J. Digit. Earth* **2025**, *18*, 2458688. [CrossRef]
32. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
33. Van der Meer, F.D.; van der Werff, H.M.A.; van Ruitenbeek, F.J.A.; Hecker, C.A.; Bakker, W.H.; Noomen, M.F.; van der Meijde, M.; Carranza, E.J.M.; de Smeth, J.B.; Woldai, T. Multi- and Hyperspectral Geologic Remote Sensing: A Review. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *14*, 112–128. [CrossRef]
34. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. GeoChat: Grounded Large Vision–Language Model for Remote Sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 27831–27840.
35. Vahidnia, M.H. Empowering Geoportals HCI with TaskOriented Chatbots Through NLP and Deep Transfer Learning. *Big Earth Data* **2024**, *8*, 608–648. [CrossRef]
36. Rane, J.; Kaya, Ö.; Mallick, S.K.; Rane, N.L. Artificial IntelligencePowered Spatial Analysis and ChatGPTDriven Interpretation of Remote Sensing and GIS Data. In *Generative Artificial Intelligence in Agriculture, Education, and Business*; Deep Science Publishing: San Francisco, CA, USA, 2024; pp. 162–217.
37. Pierdicca, R.; Muralikrishna, N.; Tonetto, F.; Ghianda, A. On the Use of Large Language Models for GIS-Based Spatial Analysis. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 401. [CrossRef]
38. Ciccone, G. ChatGPT as a Digital Assistant for Archaeology: Insights from the Smart Anomaly Detection Assistant Development. *Heritage* **2024**, *7*, 5428–5445. [CrossRef]
39. Labba, C.; Alcouffe, A.; Crubézy, E.; Boyer, A. IArch: An AI Tool for Digging Deeper into Archaeological Data. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA, 6–8 November 2023; 2023; pp. 22–29.
40. Bellat, M.; Orellana Figueroa, J.D.; Reeves, J.S.; TaghizadehMehrjardi, R.; Tennie, C.; Scholten, T. Machine Learning Applications in Archaeological Practices: A Review. *arXiv* **2025**, arXiv:2501.03840. [CrossRef]
41. Ali, G.; Mijwil, M.M.; Adamopoulos, I.; Ayad, J. Leveraging the Internet of Things, Remote Sensing and Artificial Intelligence for Sustainable Forest Management. *Babylonian J. Internet Things* **2025**, *1*, 1–65. [CrossRef] [PubMed]
42. BroniBediako, C.; Xia, J.; Yokoya, N. Real-Time Semantic Segmentation: A brief survey and comparative study in remote sensing. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 94–124. [CrossRef]
43. Patel, N. Generative Artificial Intelligence and Remote Sensing: A Perspective on the Past and the Future. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 86–88. [CrossRef]
44. Lin, H.; Hong, D.; Ge, S.; Luo, C.; Jiang, K.; Jin, H.; Wen, C. RSMoE: A VisionLanguage Model with Mixture of Experts for Remote Sensing Image Captioning and Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5614918. [CrossRef]
45. Casini, L.; Orr, V.; Montanucci, A.; Marchetti, N.; Rocchetti, M. Archaeological Sites Detection with a HumanAI Collaboration Workflow. *arXiv* **2023**, arXiv:2302.05286. [CrossRef]
46. Agapiou, A. Unequal Horizons: Global North–South Disparities in Archaeological Earth Observation (2000–2025). *Remote Sens.* **2025**, *17*, 3371. [CrossRef]
47. Drpinghaus, J.; Weil, V.; Rockenfeller, R.; Mangroliya, M.P. A Novel Approach Towards the Robustness of Centrality Measures in Networks. *Soc. Sci. Humanit. Open* **2025**, *11*, 101183. [CrossRef]
48. Hazan, S. The Promise of Pastiche: How AI Systems Pervade the Cultural Heritage Sector. In *The Arts and Computational Culture: Real and Virtual Worlds*; Giannini, T., Bowen, J.P., Eds.; Springer: Cham, Switzerland, 2024; pp. 253–274. [CrossRef]

49. AI in QGIS: Integration of LargeLanguageModel Plugins (Promptly, QLLM, and Others). QGIS Repository Documentation 2025. Available online: <https://plugins.qgis.org> (accessed on 3 August 2025).
50. Gattiglia, G. Managing Artificial Intelligence in Archaeology: An Overview. *J. Cult. Herit.* **2025**, *71*, 225–233. [\[CrossRef\]](#)
51. Zeynali, R.; Mandanici, E.; Bitelli, G. A Technical Note on AI-Driven Archaeological Object Detection in Airborne LiDAR Derivative Data, with CNN as the Leading Technique. *Remote Sens.* **2025**, *17*, 2733. [\[CrossRef\]](#)
52. Casini, L.; Marchetti, N.; Montanucci, A.; Orr, V.; Rocchetti, M. A HumanAI Collaboration Workflow for Archaeological Sites Detection. *Sci. Rep.* **2023**, *13*, 8699. [\[CrossRef\]](#)
53. Klein, K.; Müller, A.; Wohde, A.; Gorelik, A.V.; Heyd, V.; Lämmel, R.; Diekmann, Y.; Bami, M. AutArch: An AI-Assisted Workflow for Object Detection and Data Collection from Archaeological Catalogues. *J. Archaeol. Sci.* **2025**, *179*, 106244. [\[CrossRef\]](#)
54. El Behairy, R.A.; El Arwash, H.M.; El Baroudy, A.A.; Ibrahim, M.M.; Mohamed, E.S.; Rebouh, N.Y.; Shokr, M.S. Artificial Intelligence Integrated GIS for Land Suitability Assessment of Wheat Crop Growth in Arid Zones to Sustain Food Security. *Agronomy* **2023**, *13*, 1281. [\[CrossRef\]](#)
55. Sundar, A.; Heck, L. Multimodal Conversational AI: A Survey of Datasets and Approaches. In Proceedings of the 4th Workshop on NLP for Conversational AI. 2022. Available online: <https://aclanthology.org/2022.nlp4convai-1.12/> (accessed on 1 August 2025).
56. Carolan, K.; Fennelly, L.F.; Smeaton, A.F. A review of multimodal large language and vision models. *arXiv* **2024**, arXiv:2404.01322. [\[CrossRef\]](#)
57. Singh, G. A Review of Multimodal Vision–Language Models: Foundations, Applications, and Future Directions. 2025; *preprint*. [\[CrossRef\]](#)
58. Tao, L.; Zhang, H.; Jing, H.; Liu, Y.; Yan, D.; Wei, G.; Xue, X. Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques. *Remote Sens.* **2025**, *17*, 162. [\[CrossRef\]](#)
59. Xu, W.; Yu, R.; Xue, M.; Wang, X.; Zhang, Y.; Wei, Z.; Zhang, Z.; Peng, M.; Wu, Y. A survey on Earth observation multimodal large language models: Framework, core technologies, and future perspectives. *J. Radars*, 2025; *in press*. [\[CrossRef\]](#)
60. Wang, J.; Jiang, H.; Liu, Y.; Ma, C.; Zhang, X.; Pan, Y.; Liu, M.; Gu, P.; Xia, S.; Li, W.; et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv* **2024**, arXiv:2408.01319. [\[CrossRef\]](#)
61. Abate, N.; Visone, F.; Sileo, M.; Danese, M.; Minervino Amodio, A.; Lasaponara, R.; Masini, N. Potential Impact of Using ChatGPT3.5 in the Theoretical and Practical MultiLevel Approach to OpenSource RemoteSensing Archaeology: Preliminary Considerations. *Heritage* **2023**, *6*, 7640–7659. [\[CrossRef\]](#)
62. Floridi, L.; Cows, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [\[CrossRef\]](#)
63. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [\[CrossRef\]](#)
64. Tapete, D.; Cigna, F. Detection of archaeological looting from space: Methods, achievements and challenges. *Remote Sens.* **2019**, *11*, 2389. [\[CrossRef\]](#)
65. Carroll, S.; Garba, I.; Figueroa-Rodríguez, O.; Holbrook, J.; Lovett, R.; Materechera, S.; Parsons, M.; Raseroka, K.; Rodriguez-Lonebear, D.; Rowe, R.; et al. The CARE Principles for Indigenous Data Governance. *Data Sci. J.* **2020**, *19*, 43. [\[CrossRef\]](#)
66. Nicholas, G.P. Indigenous archaeology. In *Oxford Bibliographies in Anthropology*; Oxford University Press: Oxford, UK, 2020. [\[CrossRef\]](#)
67. Rainie, S.C.; Kukutai, T.; Walter, M.; Figueroa-Rodríguez, O.L.; Walker, J.; Axelsson, P. Indigenous data sovereignty. In *State of Open Data: Histories and Horizons*; Davies, T., Walker, S., Rubinstein, M., Perini, F., Eds.; African Minds: Cape Town, South Africa, 2019; pp. 300–319. [\[CrossRef\]](#)
68. Cochran, P.A.L.; Marshall, C.A.; Garcia-Downing, C.; Kendall, E.; Cook, D.; McCubbin, L.; Gover, R.M.S. Indigenous ways of knowing: Implications for participatory research and community engagement. *Am. J. Public Health* **2008**, *98*, 22–27. [\[CrossRef\]](#)
69. OpenAI. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [\[CrossRef\]](#)
70. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. A survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Learning from the Rare: Overcoming Class Imbalance in Archaeological Object Detection with Boosting Methods

Argyro Argyrou ^{1,*} , Federico Fasson ² , Emeri Farinetti ² , Apostolos Papakonstantinou ¹ ,
Dimitrios D. Alexakis ³  and Athos Agapiou ¹ 

¹ Department of Civil Engineering and Geomatics, Faculty of Engineering and Technology, Cyprus University of Technology, Saripolou 2-8, 3036 Achilleos 1 Building, 2nd Floor, Lemesos P.O. Box 50329, Cyprus; a.papakonstantinou@cut.ac.cy (A.P.); athos.agapiou@cut.ac.cy (A.A.)

² Department of Humanities, Roma Tre University, Via Ostiense 239, 00146 Rome, Italy; federico.fasson@uniroma3.it (F.F.); emeri.farinetti@uniroma3.it (E.F.)

³ Laboratory of Geophysics—Satellite Remote Sensing & Archaeoenvironment (GeoSat ReSeArch Lab), Institute for Mediterranean Studies, Foundation for Research and Technology—Hellas (FORTH), Nikiforou Foka 130 & Melissinou, 74100 Rethymno, Crete, Greece; dalexakis@ims.forth.gr

* Correspondence: argyro.argyrou@cut.ac.cy; Tel.: +357-25002542

Abstract

Detecting surface potsherds using low-altitude remote sensing is challenging due to severe class imbalance and limited training data. This study develops and validates a semi-automatic detection methodology that adapts threshold-optimized boosting classifiers (AdaBoost, XGBoost) to maximize ceramic detection recall under extreme class imbalance in the Western Megaris archeological landscape, Greece. Models were trained on only 15% of the available data to simulate realistic field conditions. Evaluation emphasized recall-oriented metrics (precision, recall, F1-score, AUC) for the minority class, addressing the accuracy paradox where high overall accuracy masks poor rare-class performance. Threshold optimization enabled AdaBoost and XGBoost to achieve substantially improved recall compared to baseline methods, with detection-to-ground-truth ratios of 2.5 and 3.2, respectively, reflecting deliberate prioritization of recall over precision for exploratory survey purposes. The results demonstrate that this methodological framework provides archeologically interpretable screening tools for identifying high-probability ceramic locations, supporting more efficient field survey design and heritage documentation workflows in Mediterranean landscapes.

Keywords: archeology; remote sensing; accuracy paradox; machine learning; boosting algorithms; surface potsherds; Megara; drone



Academic Editors: Nicola Masini and Fulong Chen

Received: 18 January 2026

Revised: 22 February 2026

Accepted: 27 February 2026

Published: 2 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Remote Sensing in Archeological Research

Surface potsherds play a crucial role in archeological research, offering significant insights into past human activity, trade networks, and cultural practices. Their durability and widespread occurrence make them essential indicators for reconstructing settlement patterns and understanding the dynamics of human occupation [1]. These remains serve as material evidence of both historic and prehistoric activities [2]. As noted by Orengo and Garcia-Molsosa (2019) [3], analyzing the dispersion of surface artifacts can inform researchers about potential changes in land use and the destruction or transformation of archeological sites. Surface surveys are a foundational method in archeology, enabling the

identification of settlement patterns and behaviors across the landscape. Additionally, they offer a means to investigate the relationship between past human populations and their natural environments, supporting the discovery and protection of cultural heritage in the context of rapidly changing modern landscapes.

Despite their value, traditional ground-based surface surveys present several limitations: (a) they are time-consuming, (b) require extensive training, (c) rely on grid-based sampling methods, (d) detect only those artifacts exposed at the surface, (e) may not always meet survey objectives due to methodological constraints, and (f) face limitations related to environmental conditions such as terrain, vegetation, lighting, and weather [3].

In recent years, remote sensing technologies have become increasingly valuable in archeological investigations [4,5]. Technological advancements, particularly in the spatial and spectral resolution of satellite sensors, and the open access to datasets such as Landsat and Sentinel have facilitated their broader use in identifying archeological features across vast landscapes [6]. The integration of low-cost, low-altitude platforms, especially drones, has further enhanced the capabilities of archeologists by enabling efficient documentation and monitoring of archeological sites [7].

Unmanned Aerial Vehicles (UAVs) allow for the collection of high-resolution imagery and spatial data from previously inaccessible areas, offering new and comprehensive perspectives on archeological landscapes. Recent studies have highlighted the effectiveness of combining drone imagery with machine learning classifiers to detect artifact scatters, particularly surface potsherds, thereby improving both the precision and efficiency of archeological surveys [2].

Alongside advances in imaging platforms, archeology has also undergone a digital transformation in data analysis. Researchers have increasingly shifted from traditional desktop software to scalable, cloud-based platforms powered by advanced artificial intelligence (AI) algorithms [3]. A pioneering study by [3] demonstrated the use of Google Earth Engine and Random Forest classifiers to detect archeological potsherds from centimeter-resolution drone imagery.

Machine learning (ML) classifiers' predictive power enables archeologists to process large datasets, evaluate excavation strategies, and support preservation planning [8–32]. Studies by Agapiou et al. [9] and [3] illustrate how deep learning techniques applied to UAV imagery can dramatically improve outcomes in field research, especially in contexts with limited time or resources—by delivering faster results and higher analytical precision.

Despite the appeal of these technologies, traditional artifact surface surveys remain rooted in systematic, human-led observation. Fieldwalking relies on structured, direct visual inspection, guided by established protocols and expert experience. While digital tools support data management and visualization, initial identification and interpretation typically depend on trained surveyors. Programs like the Fieldwalker's Integrated Documentation system (FIDo) and long-standing projects featured in the IMSW workshop series (e.g., Fasti Online Survey) showcase how systematic field surveys continue to evolve while remaining foundational to archeological practice [33].

There is substantial potential for automated detection methods to complement—not replace—traditional survey techniques. Their effectiveness depends on being rigorously tested and trained using real-world survey data. This article seeks to contribute to that dialog by examining how technologies such as ML and multi-class classification can enhance pedestrian survey workflows, particularly by improving area selection, site potential evaluation, and artifact recognition.

1.2. Challenges and Limitations

Although there have been initial examples of machine learning (ML) applications in archeology, particularly in classification and detection tasks [2], the issue of class imbalance has not been adequately addressed [34]. As emphasized by Sun et al. [35], class imbalance arises when one class significantly outnumbers the other(s), resulting in skewed learning outcomes that fail to represent the minority class sufficiently. This issue is particularly relevant in archeology when detecting rare surface potsherds (minority class) within large, complex landscapes dominated by non-surface potsherds data (majority class).

Most standard classification algorithms are developed assuming a relatively balanced class distribution (Figure 1A). However, when applied to datasets with imbalanced distributions (Figure 1B), their performance often fades, resulting in biased predictions that favor the majority class [34]. In a classification dataset, the majority class is the one with the highest number of instances, while the minority class has significantly fewer instances. This imbalance causes many machine learning models to focus on the dominant class, leading to poor performance in detecting or predicting the minority class.

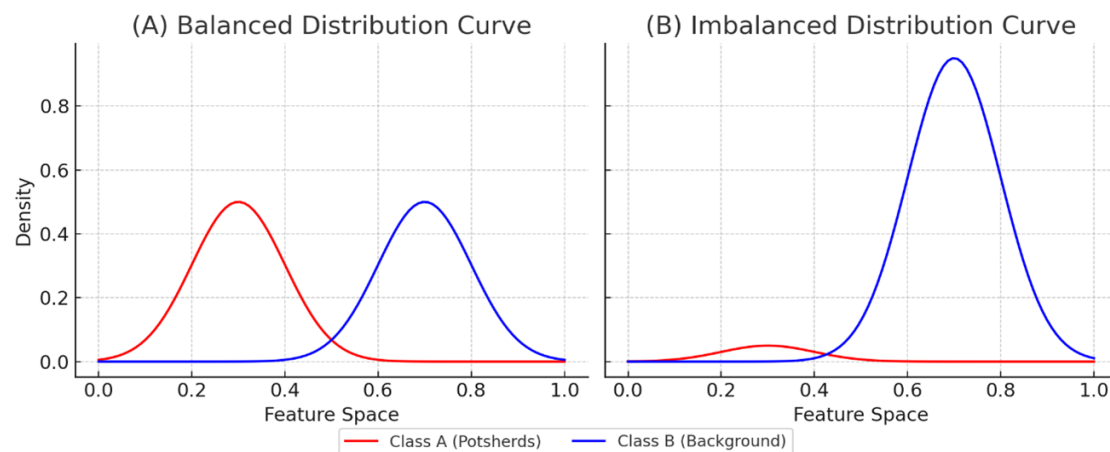


Figure 1. (A) shows balanced class distribution; (B) shows class imbalance, which can cause classifiers to favor the majority class and reduce accuracy for the minority class.

The class imbalance problem is a well-documented challenge in machine learning and data mining, with critical implications across domains such as medical diagnosis, satellite image analysis, and risk assessment.

Traditional classification methods—including decision trees, Support Vector Machines (SVMs), neural networks, Bayesian models, k-nearest neighbors, and associative classifiers—are known to struggle with class imbalance [35]. These models often optimize for overall accuracy, which can mask poor performance on the minority class, leading to misleading evaluation results and reduced model reliability. Due to its substantial impact on generalizability and fairness, class imbalance has become a central focus in recent research on robust AI and machine learning systems [35].

An imbalanced dataset is characterized by unequal class proportions, where the majority class dominates and the minority class is underrepresented (Figure 2). According to Sun et al. (2009) [35], the challenges of modeling minority classes—such as surface potsherds in this study—extend beyond skewed distributions. Additional factors such as small sample size, low class separability, and the presence of within-class sub-concepts further complicate the learning process. Effectively addressing these challenges is essential for achieving meaningful and interpretable results in ML-driven archeological research.

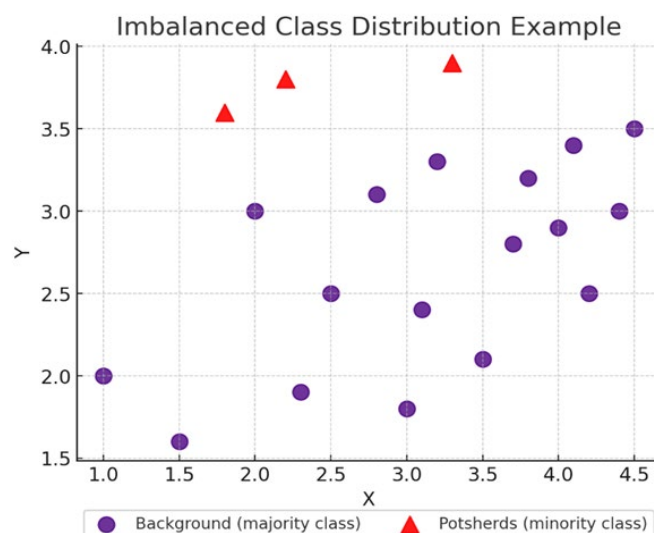


Figure 2. An imbalanced data set example with unequal class proportions.

1.3. Evaluating Metrics, the Accuracy Paradox, and the Challenge of Small Object Detection

Evaluation metrics play a dual role in surface potsherds detection: they inform model development and serve as indicators of classification performance. The accuracy paradox demonstrates that high overall accuracy does not necessarily reflect the true effectiveness of a classifier, particularly in imbalanced datasets [36,37]. While widely used metrics such as overall accuracy and the kappa coefficient [38–40] have been critiqued for their limitations in imbalanced settings [41], more informative measures, including precision, recall, F1-score, and the Matthews Correlation Coefficient (MCC), provide a more reliable assessment of classifier performance. When dealing with class imbalance—where one class is rare but of high importance—accuracy becomes misleading, as the majority class dominates the metric [42,43]. In our case study, surface potsherd detection is a binary classification problem with significant imbalance, where the minority class (surface potsherds) is critical despite its low representation in the data. As discussed by [35], post-classification results can be categorized into true positives, false positives, true negatives, and false negatives, summarized in the confusion matrix (see Table 1).

Table 1. Categories of samples.

	Predicted as Positive	Predicted as Negative
Actually Positive	True positives (TPs)	False negatives (FNs)
Actually Negative	False positive (FP)	True negatives (TNs)

A further challenge, largely overlooked in the archeological remote sensing literature, concerns the physical scale of the target objects themselves. Archeological ceramic fragments are inherently small relative to the spatial footprint of UAV imagery pixels, even at very high resolutions. At a GSD of 2.5 cm/px, a typical potsherd of 5–10 cm occupies only 4–16 pixels, placing it firmly within the “small object” category as defined in computer vision, where objects typically occupy fewer than 32×32 pixels. Small object detection in UAV imagery presents well-documented difficulties: sparse pixel representation, limited discriminative spectral and textural features, and high sensitivity to background clutter [3]. These challenges directly compound the class imbalance problem: small, underrepresented objects are simultaneously harder to detect and easier to misclassify as background. The practical consequence is that a classifier may achieve high overall accuracy simply by predicting background everywhere, completely missing the target class, which is precisely the accuracy paradox described above. Addressing both challenges of class imbalance

and small object scale is therefore essential for developing operationally viable ceramic detection systems from UAV imagery.

Several measures can be derived using the confusion matrix table (Table 2). Therefore, classification performance should be evaluated by different measures depending on the different classification objectives. Clearly, neither of these measures is adequate by itself, and several other evaluation metrics are devised.

Table 2. Confusion matrix table.

	Predicted Positive	Predicted Negative	
Actual Positive	TP (<i>true positive</i>)	FN (<i>False Negative</i>)	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP (<i>false positive</i>)	TN (<i>True Negative</i>)	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{(TP+TN)}{(TP+TN+FP+FN)}$

1.4. Objectives and Contributions of This Study

This study develops a semi-automatic methodological framework for detecting surface ceramic fragments in Mediterranean archeological landscapes, addressing the operational challenge of extreme class imbalance (<1% ceramic pixels) that characterizes UAV-based prospection data. Rather than introducing novel algorithms, we adapt existing boosting classifiers (AdaBoost, XGBoost) through application-specific threshold optimization to prioritize minority-class recall—a requirement fundamental to archeological survey where undetected artifacts represent irreversible information loss. This framework is validated using systematic field survey data from the Western Megaris landscape, demonstrating how computational methods can enhance field efficiency while maintaining archeological interpretability. Our research explores how machine learning (ML) algorithms can be used to classify and detect archeological surface potsherds using low-altitude sensor cameras. Additionally, it is examined how an interdisciplinary approach that combines remote sensing and artificial intelligence can tackle the “accuracy paradox” often encountered when dealing with imbalanced remote sensing data.

Boosting algorithms are particularly suited to archeological ceramic detection because they iteratively reweight misclassified minority instances, progressively focusing computational effort on difficult-to-detect rare objects [35]. While solutions for binary class imbalance exist in computer science literature, their translation into archeologically meaningful workflows remains limited. This is achieved by assigning greater weights to important surface potsherds samples. While research has investigated solutions for binary class imbalance, these methods have not yet been applied in the field of archeology. This paper aims to implement boosting techniques to develop a methodology for the semi-automatic detection of surface potsherds using low-altitude, low-cost sensors. The study addresses the challenges of detecting minority classes in archeological remote sensing by focusing on classifying surface potsherds within highly imbalanced datasets. Specifically, the objectives are to:

- Evaluate the effectiveness of boosting-based classifiers—such as AdaBoost and XGBoost—in detecting surface potsherds, which represent the minority class, and compare their performance against commonly used classifiers.
- Assess the performance of these boosting algorithms under limited training data conditions and investigate whether they can significantly enhance the detection of minority class instances.

- (c) Emphasize the importance of using evaluation metrics beyond overall accuracy—such as precision, recall, F1-score, and Area Under the Curve (AUC) to provide a more reliable assessment of classifier performance in imbalanced classification settings.
- (d) Develop an operational methodological framework that integrates threshold-optimized boosting classifiers with UAV-based multispectral data and systematic ground-truth validation, providing archeologists with practical screening tools for identifying high-probability ceramic locations and optimizing field survey resource allocation in Mediterranean landscapes.

This research contributes a methodologically transparent framework for translating class-imbalanced ML classifiers into archeological prospection workflows. By explicitly addressing the trade-offs between detection recall and false positive rates, and by validating algorithmic outputs against systematic pedestrian survey data, this research demonstrates how computational approaches can augment—rather than replace—expert archeological interpretation in heritage landscape documentation.

2. Study Area: The Megara Landscape

2.1. Geographical and Historical Context of Megara

Megara is a historic city in Greece, located in the western part of Attica towards the Isthmus of Corinth. It is approximately 34 km west of Athens and 37 km east of Corinth. Its strategic position—bridging central Greece and the Peloponnese and providing access to both the Saronic Gulf (via its port, Nisaea) and the Corinthian Gulf (via Pagae)—played a crucial role in its historical development and significance. The Megaris region holds topographic and geopolitical importance, lying along vital routes of maritime and overland trade.

The WeMALP project [33] seeks to reconstruct the long-term landscape transformation of Western Megaris, a region of considerable geographical and historical interest within the wider Megaris. Bounded to the north by the Geraneia mountains and stretching toward the Saronic Gulf, this area functioned as a key corridor for commerce, communication, and cultural exchange. The project's primary objectives include investigating long-term rural land use, exploring the interactions between city and countryside, identifying contemporary settlements from the ancient phases of Megara, and highlighting the continued importance of this natural bridge connecting Attica and Corinthia.

Since 2019, WeMALP has conducted intensive surface archeological surveys in key areas to answer historical questions and delineate settlement pattern trajectories, such as Vrestiza, Chani Derveni, Marmara, and Dami Mouchlizi. These surveys provide valuable insights into ancient settlement patterns, agricultural practices, and responses to environmental changes, significantly enriching our understanding of Greek historical landscapes and regional dynamics [44].

In particular, systematic surface research on the area of Dami Mouchlizi, extending to the immediate West of the walled city of ancient Megara, has offered a window into the socio-economic strategies of rural settlements along key routes between the city of Megara and the Geraneia ridge towards Corinth. In particular, the fortified farm site at Dami Mouchlizi [45] has illustrated the multifaceted roles of elite estates in agricultural production and regional governance (Figures 3 and 4). The site known as Dami tou Mouchlizi (DM1) features structures built on a terrace and a circular tower, situated on a low hill with excellent visibility over the gulf. The architectural remains visible on the surface were documented through plans and elevations during previous field campaigns, and most likely represent an example of a so-called fortified farmhouse dating to the Classical–Hellenistic period—a type well attested, especially in nearby Attica.

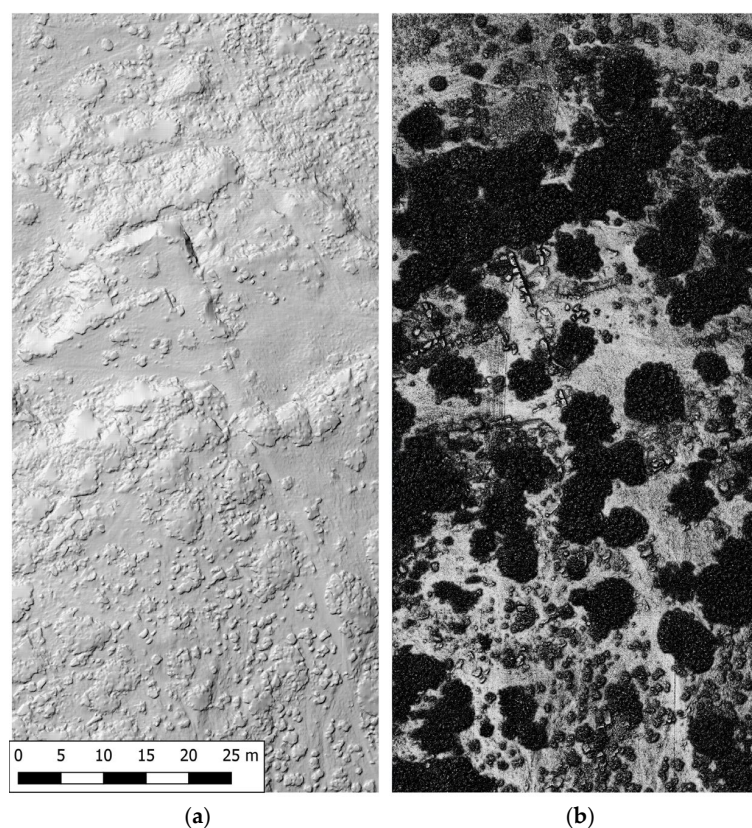


Figure 3. LiDAR mapping of the site of *Dami tou Mouchlizi* (DM1) after processing: (a) DEM (0.05 m/px) generated using DJI Terra software (v4.4.6); (b) VAT (Visualization for Archeological Topography) based on the DSM (0.05 m/px) derived from the unfiltered point cloud (elab. F. Fasson; © SAIA Archives, V/230). Note: LiDAR data were acquired within the framework of the WeMALP project and are presented here solely to provide archeological context for the study area. These data are independent of the UAV-RGB dataset collected for the present study.

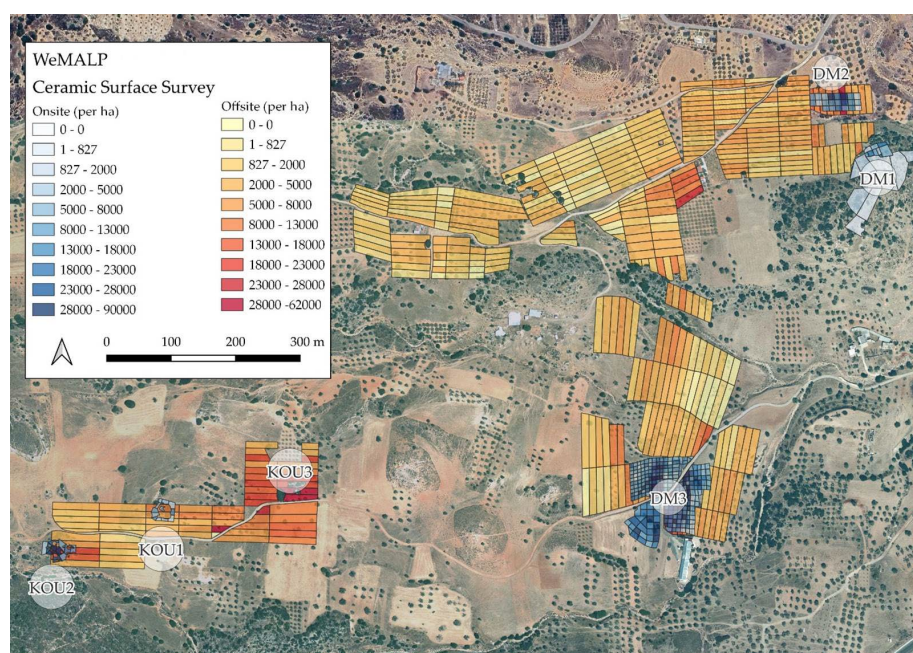


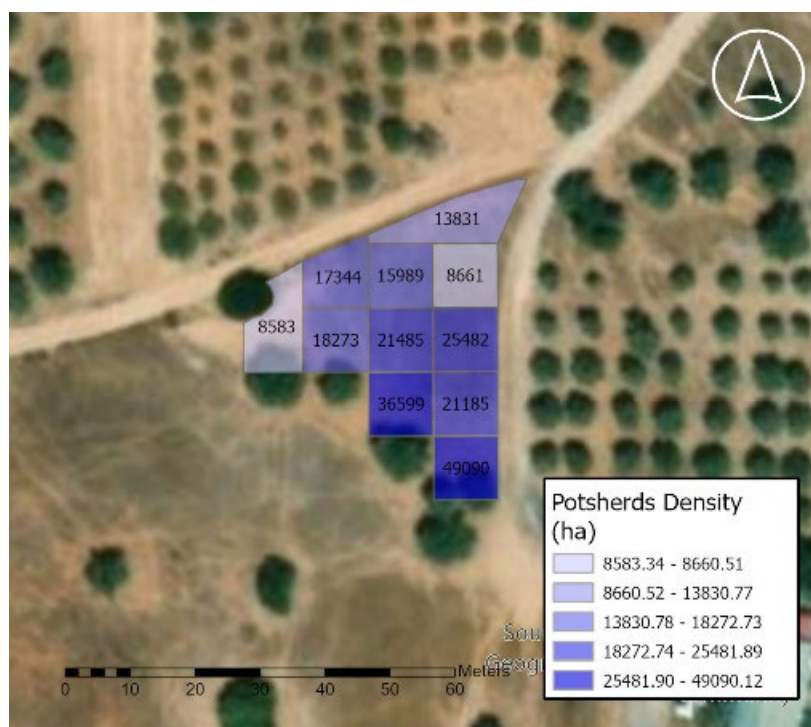
Figure 4. Area of Dami Mouchlizi DM1 and DM3 (graphic design by M. Bottoni, E. Farinetti and M. Rossi; © SAIA Archives, V/228[M1])el. grafica M. Bottoni, E. Farinetti e M. Rossi; ©Archivi SAIA, V/228).

In the surrounding area, the DM3 site emerged as one of the most significant Late Roman rural settlements in the region, with evidence also indicating occupation during the Classical–Hellenistic period. It appears to have succeeded DM1 in the agricultural exploitation and management of the territory.

The site (Figure 5a,b), whose surface potsherd material was collected (Figure 6) during an intensive and systematic survey, is still under study, suggesting the presence of an extensive production center belonging to an elite landowner.



(a)



(b)

Figure 5. (a) Panoramic photo of the DM3 site, which occupies the olive groves and pistachio orchard in the foreground (photo by F. Fasson; ©SAIA Archives, U/13911); (b) density map of surface potsherds material in the selected area (elab. by M. Bottoni, M. Rossi; ©SAIA Archives, V/229).



Figure 6. Example of a potsherd assemblage from a survey unit at the DM3 site (photo by S. Forgione; ©SAIA Archives, U/13912).

2.2. Environmental and Imaging Conditions Relevant to Drone-Based Detection

Between September and October 2024, fieldwork conducted with UAVs focused on high-precision mapping to support civil engineering and geomatics research for cultural heritage conservation. The survey was conducted in a cultivated field planted with olive and pistachio trees, characterized by regularly tilled and well-maintained terrain. The surface was scattered with lithic material of a reddish hue, often indistinguishable in color from the surrounding soil. This chromatic homogeneity posed a significant challenge for identifying surface potsherds fragments—especially those partially embedded or coated in soil—as they did not naturally stand out in the RGB spectrum.

To address this, flight planning prioritized maximizing chromatic and textural contrast through careful timing and altitude selection. Data acquisition took place under stable weather conditions during morning hours to minimize shadows and avoid strong solar reflections common around midday. Shadows cast by trees or uneven terrain can obscure surface details and hinder feature detection; likewise, overexposed images may cause the loss of subtle textural cues critical for distinguishing surface potsherds from stone.

To achieve sufficiently high spatial resolution for detecting small surface potsherds, multiple photogrammetric datasets were acquired at different altitudes to evaluate the optimal balance between altitude, area coverage, and dataset size. Additionally, flight planning considered soil moisture levels: flights were avoided immediately after rainfall or irrigation to prevent reflectance changes and to reduce the risk of surface potsherds fragments being obscured by mud or debris.

The data collected during this campaign are utilized in this study, which aims to develop an automated methodology for detecting surface archeological potsherds while using the fewest possible samples for classification. The results will be compared with findings from the University of Rome team to facilitate international peer review and cross-validation.

3. Materials and Methods

This research integrates several recent technological advancements that have transformed archeological workflows. The increasing accessibility and performance of low-cost, low-altitude UAVs have enabled more autonomous and efficient aerial surveys. At the same time, digital photogrammetry has become more user-friendly through the development of semi-automated workflows, which are now widely adopted in archeological documentation. A further breakthrough lies in the application of machine learning (ML), a subset of artificial intelligence, which enables software to make accurate predictions and perform object detection without explicit programming. ML has emerged as a powerful tool for data mining and pattern recognition in archeological contexts. Finally, the use of multiple evaluation metrics—such as precision, recall, F1-score, and AUC—beyond simple overall accuracy allows for more reliable assessment and modeling of classification algorithms, particularly when addressing imbalanced datasets.

The proposed approach has been tested in simulated environments in Alampira village and at an archeological site in Kophinou village, both located in Cyprus (Argyrou et al., 2023 [46]; Argyrou et al., 2025 [47]). The research aims to tackle the challenges posed by the time-consuming nature of data processing and the accuracy issues in detecting surface potsherds, especially concerning the minority class of these artifacts and the phenomenon known as the “accuracy paradox.”

This study integrated drone-based remote sensing, machine learning (ML) classification, and GIS spatial analysis to identify surface potsherds artifacts in an archeological context. The methodology is a semi-automated framework of image classification (Figure 7) using a low-altitude and low-cost drone, enabling precise documentation of the surveyed areas. High-resolution drone imagery was processed into a georeferenced orthomosaic using Terra software (v4.4.6), from which spectral bands (RGB values) and spatial coordinates (x, y) were extracted. In ArcGIS Pro 3.5.0, labeled training samples were created for surface potsherds (Class 1) and non-surface potsherds (Class 0) regions, with each sample including mean RGB values and geographic locations. Additionally, rigorous model training, including hyperparameter tuning with machine learning and boosting algorithms, was conducted to detect and classify surface potsherds fragments through photomosaic analysis. Five supervised classifiers (Random Forest (RF), Support Vector Machine (SVM), and Maximum Likelihood algorithm (MLC), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost)) were evaluated using precision, recall, and F1-scores after splitting the dataset into training and testing subsets.

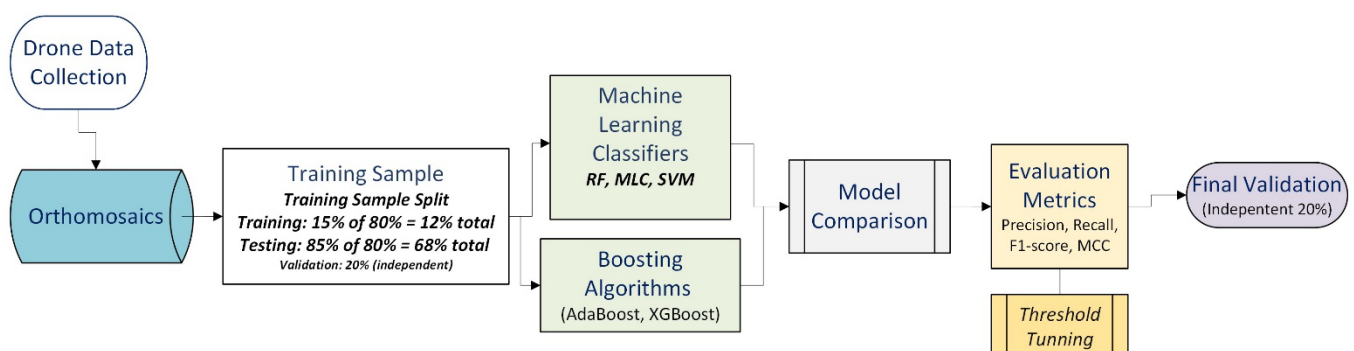


Figure 7. Workflow implemented in this study.

Post-processing was carried out in ArcGIS Pro 3.5.0 using a reproducible workflow. The target class (potsherds) was extracted with the Reclassify tool, followed by the application of the Majority Filter to reduce classification noise. The resulting raster was converted

into polygons representing contiguous ceramic areas, and subsequently into point features using polygon centroids.

3.1. UAV Image Acquisition

During the survey, a low-altitude drone acquired a high-resolution photographic dataset for terrain analysis using a DJI Matrice 350 RTK (SZ DJI Technology Co., Ltd., Shenzhen, China) equipped with a Zenmuse L2 sensor (SZ DJI Technology Co., Ltd., Shenzhen, China) (20 MP, 4/3-inch CMOS, max resolution 5280×3956 pixels). Images were captured at various altitudes: 12.5, 15, and 20 m, to assess altitude effects on ground resolution and image quality. Flights occurred between 10:00 a.m. and 11:00 a.m. under favorable lighting, minimizing shadows and reflections to reduce post-processing and maintain chromatic consistency and contrast. Ground resolution varied with altitude but generally remained within a few centimeters per pixel (2.5 cm/px–3cm/px), suitable for detailed analysis. Among the three acquisition altitudes, the 12.5 m dataset (GSD = 2.5 cm/px) was selected for model training and evaluation, as it provided the highest spatial resolution for ceramic fragment detection.

3.2. Photogrammetric Processing and Computational Processing

The next step included computational processing (ML techniques) to identify and isolate surface potsherds fragments using the orthomosaic of 293 captured images. The photogrammetric processing of the photos involved the orthorectification of all photographs and combining them into a single orthomosaic using Terra software Version Number: 4.4.6. Orthorectifying the image involves ensuring that the images are geometrically accurate and corrected from lens distortion, camera tilt, perspective, and topographic relief. Therefore, the images were orthorectified and merged into an orthomosaic map using the photos' metadata, which contained information like drone model, camera sensor, lens types, and GNSS coordinates. After the mosaic was created (Figure 8), image processing techniques were used to identify surface potsherds in the area of interest, which is a grid of 4×100 sq. m where the field survey was conducted.

The full dataset consisted of 250 labeled samples. A 20% subset (50 samples) was reserved for final validation. The remaining 200 samples were used as follows:

- 15% (30 samples) for model training;
- 85% (170 samples) for testing.

All classifiers were trained on the same 30-sample training set and evaluated on the 170-sample test set.

All analyses were carried out using R version 4.4.0 (R Core Team, 2025) and ArcGIS Pro 3.5.0. Specifically, all machine learning classifiers (Random Forest, SVM, XGBoost, AdaBoost, and MLC) were implemented and trained in R software; ArcGIS Pro 3.5.0 was used exclusively for data preparation prior to training and for post-processing and visualization of classification results. Computational processing was performed on a laptop equipped with an 11th Generation Intel® Core™ i7-1165G7 processor operating at 2.80 GHz and 32.0 GB of RAM, running a 64-bit Windows operating system. ArcGIS Image Analyst tools of the ArcGIS Pro 3.5.0 software were used for computational processing. Within the ArcGIS Pro 3.5.0 environment, a training sample was given using the Training Samples Manager in the Classification Tools, by selecting areas of interest (points) for two classes: 'surface potsherds' (class 1) and 'No surface potsherds' (class 0). The training sample file included a class name indicating the name of the class category and a class value containing the integer value for each class category (class 1 = 1, class 0 = 0). The creation of the training data consisted of assigning to each class the values of the pixels delimited by the points in each composite band.

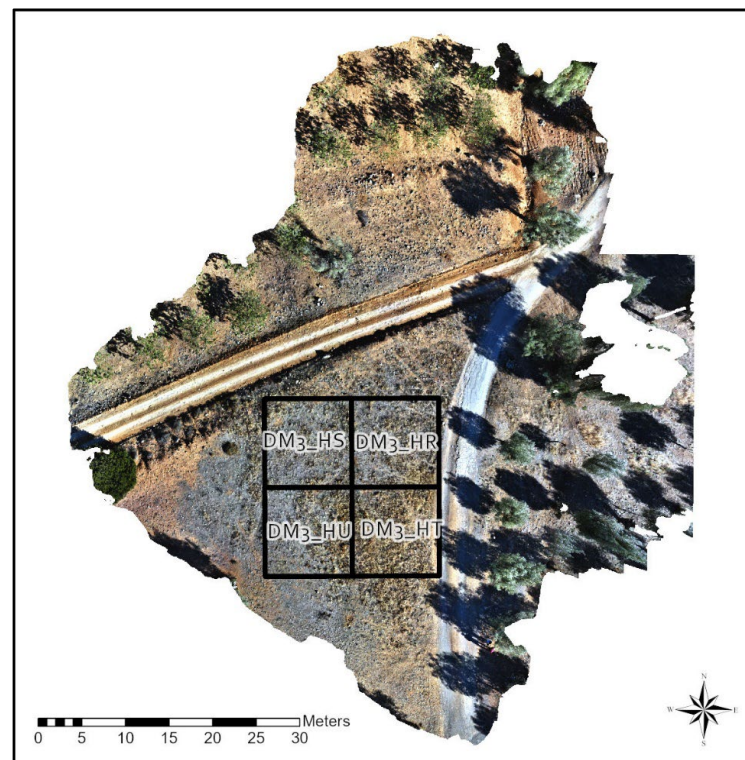


Figure 8. Orthomosaic of the Dami 3 area in Megara. The grid with a black outline indicates the area of interest used in this study.

To address the inherent class imbalance in the dataset, where surface potsherds remain represent the minority class, a two-stage stratified sampling approach was employed to preserve class distribution across all data subsets. In the first stage, 20% of the total dataset was randomly selected using stratified sampling and withheld as a hold-out validation set, entirely unseen during model training and hyperparameter tuning. This set is an independent benchmark for evaluating generalization performance after model development. In the second stage, the remaining 80% of the data was again stratified and split into a 15% training subset and an 85% internal test subset with a skewed distribution of the two classes (80% majority class and 20% minority class). This setup simulates real-world constraints faced in archeological surveys, where training data collection—such as verified surface potsherds fragments—is limited. This stratified approach ensures a proportional representation of both classes in all evaluation phases. Furthermore, enables a robust performance comparison across classifiers, particularly for detecting the minority surface potsherd class. The deliberately small training subset (15%, $n = 30$) was retained to evaluate classifier behavior under extreme class imbalance conditions, reflecting realistic archaeological survey constraints rather than optimized data availability.

3.3. Machine Learning Classifiers and Training Parameters

Five supervised machine learning classifiers were implemented to perform the binary classification using the three spectral input bands (red, green, blue) extracted from the orthophoto. RF, SVM, and MLC were applied initially, followed by two boosting algorithms, AdaBoost and XGBoost.

The Random Forest (RF) classifier (Breiman, 2001) [48] was trained using 500 decision trees ($n_{tree} = 500$). The number of features randomly selected at each split (m_{try}) was set to the default value equal to the square root of the total number of predictors ($\sqrt{3} \approx 1$). The Gini impurity index was used as the split criterion, and bootstrap sampling was applied for each tree. RF was implemented using this standard configuration to serve as a robust

bagging-based baseline. More advanced Random Forest variants (e.g., class-weighted or balanced RF) were not explored in order to avoid confounding tuning effects and to maintain a clear methodological contrast between bagging-based and boosting-based approaches under extreme class imbalance conditions. Maximum Likelihood Classification (MLC) was implemented via the Naïve Bayes algorithm (John & Langley, 1995) [49], which assumes conditional independence among predictors. MLC is included as a classical remote-sensing reference method, providing continuity with earlier archeological studies rather than competitive performance with modern machine learning classifiers. For each class, the probability distribution of the input features was modeled as Gaussian. No hyperparameter tuning was required, and all other parameters were kept at their default settings. SVM with a radial basis function (RBF) kernel (Cortes & Vapnik, 1995) [50] was used. Prior to training, the input features were standardized via centering and scaling. Hyperparameters (C and sigma) were tuned automatically using a 10-fold cross-validation procedure. Automatic hyperparameter optimization via cross-validation was used to obtain a representative and competitive SVM configuration. This k-fold cross-validation was applied exclusively during hyperparameter tuning; final model performance was evaluated using independent stratified training, test, and validation splits to assess generalization under extreme class imbalance and to avoid information leakage. Additional SVM variants or manually defined hyperparameter scenarios were not explored, as the aim was to contrast margin-based learning with bagging- and boosting-based approaches under extreme class imbalance, rather than to exhaustively tune individual classifiers.

Threshold optimization was applied specifically to boosting algorithms based on their theoretical foundation for addressing class imbalance. Unlike conventional classifiers that treat all instances equally, boosting methods are designed to iteratively reweight misclassified minority instances during training, progressively focusing computational effort on difficult-to-classify rare objects. This inherent design makes boosting algorithms particularly suitable for threshold-based optimization in archeological prospection, where maximizing minority-class detection (recall) is prioritized over overall accuracy. AdaBoost and XGBoost were therefore selected as the primary focus for threshold tuning to evaluate whether their theoretical advantages for imbalanced learning translate to practical archeological surface potsherd detection. RF, SVM, and MLC were retained at default thresholds to serve as baseline comparisons representing standard archeological remote sensing practice, enabling assessment of boosting algorithms' incremental value for minority-class detection in heritage applications.

In this study, boosting classifiers were utilized to enhance the detection of class 1 in an imbalanced dataset derived from UAV data. Key hyperparameters, such as the number of iterations, learning rate, and maximum depth, were optimized through grid search. Probability thresholds were systematically evaluated in increments of 0.01, ranging from 0.05 to 0.95, to maximize the F1-score for class 1 during cross-validation. Additionally, the threshold that maximized the Matthews Correlation Coefficient (MCC) was also recorded. The optimal threshold, based on the F1-score, was then applied during the validation and test stages to improve the detection performance of the minority class.

3.4. Overview of Boosting Classifiers: AdaBoost, XGBoost

As Sun et al. (2009) [35] referenced, several boosting algorithms have been reported as meta-techniques applicable to a wide range of classifier learning algorithms. These approaches typically assign higher misclassification costs to samples from rare classes and aim to minimize high-cost errors. Consequently, the principal concept involves utilizing a boosting strategy that emphasizes surface potsherds samples of higher identification importance. These algorithms are considered data-level solutions since they update the

data space by adjusting sample weights. While research exploring binary class imbalance solutions is well-documented in the literature, such techniques have yet to be applied within the archeological domain. This paper aims to implement these boosting methodologies, adapting them for semi-automatic detection of surface potsherds.

Boosting represents a robust ensemble machine learning technique integrating multiple weak learners—models that slightly outperform random guessing—into a single, highly accurate predictive model. In 1995, researchers Schapire [51] and Freund [52] introduced the AdaBoost algorithm to enhance the classification performance of existing learning algorithms. AdaBoost is recognized as the first Adaptive Boosting method, dynamically adjusting its parameters based on the data's performance, including iterative re-weighting of data and calculation of aggregation weights. The boosting process involves iteratively training a weak learner on “n” weighted iterations of the training dataset. In this way, previously misclassified instances are given increased importance [53]. The model concentrates more effectively on challenging examples through successive iterations, resulting in improved performance. Final predictions are typically achieved by aggregating the outputs of all weak learners via weighted majority voting. Boosting algorithms can employ a variety of base classifiers. Well-known examples include (AdaBoost), which is effective for binary problems and focuses on difficult-to-classify samples; (XGBoost), which enhances minority class impact and performs efficiently on large, imbalanced datasets; Gradient Boosting, which allows tuning to prioritize rare classes; Light Gradient-Boosting Machine (LightGBM), offering advanced capabilities for imbalanced data; and Categorical Boosting (CatBoost), which delivers strong results with minimal configuration even in the presence of class imbalance [51,52].

In this study, AdaBoost and XGBoost were used to detect the minority class 1. These boosting methods can improve the recall and F1-score for underrepresented classes by focusing on difficult cases. AdaBoost was selected for its simplicity and easy-to-understand results, while XGBoost was used for its better performance and efficiency with imbalanced data [54]. Both models were fine-tuned by adjusting their settings and optimizing thresholds to better detect class 1 surface potsherds, making them effective tools for reliable classification in heritage mapping. A summary of all classifiers and their key parameters is included in the table below (Table 3).

Table 3. Summary of classifier training parameters.

Classifier	Key Parameters	Defaults/Notes
Random Forest (RF)	N tree = 500	mtry = 1 ($\sqrt{3}$), bootstrap sampling, Gini split
Maximum Likelihood (Naïve Bayes proxy)	None set explicitly	Gaussian assumption for r, g, b
Support Vector Machine (SVM, Radial)	10-fold CV, metric = ROC, Pre-Proc = center & scale	Kernel parameters tuned automatically, RBF kernel
AdaBoost	N Iter = 100	Base = decision stump, exponential loss, probability threshold tuned
XGBoost	Objective = binary: logistic, eval metric = log loss, n rounds = 100	Learning rate = 0.3, max depth = 6, Gamma = 0, probability threshold tuned

3.5. Evaluation Metrics: Recall, Precision, F1-Score, Specificity, AUC, and Per-Class Confusion Matrices

As Prati et al. [55] noted, achieving an accuracy of 99.9% is straightforward in domains where the majority class comprises 99.9% of observations, simply by consistently predicting the majority class. This phenomenon, called the “accuracy paradox,” shows that these classifiers do not work well, especially when there is an imbalanced class distribution. In these situations, accurately identifying minority classes, like archeological surface potsherds, is the main goal.

To ensure robust model evaluation, each classifier’s performance was assessed using an extensive suite of metrics: accuracy, precision, recall (sensitivity), specificity, F1-score, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). Comparative Receiver Operating Characteristic (ROC) curves and precision–recall (PR) curves were generated and exported to Excel for further analysis.

Accuracy measures the proportion of correctly classified instances (true positives and negatives) out of the total predictions [41]. Although accuracy is simple and commonly used, it can be misleading in imbalanced datasets. For example, in a dataset with 95% negative instances, a model that predicts all instances as negative would achieve an accuracy of 95%, even though it completely fails to identify any positive instances. Thus, accuracy should be complemented with other metrics in classification tasks [56,57].

It is calculated as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Precision, recall (sensitivity), and specificity metrics evaluate different aspects of classifier performance.

Precision (Positive Predictive Value) measures the proportion of true positives among predicted positives. High precision indicates fewer false alarms.

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

Recall (sensitivity/true positive rate) measures the proportion of actual positives correctly identified. High recall indicates fewer missed positives.

$$Recall \text{ (Sensitivity)} = \frac{TP}{(TP + FN)} \quad (3)$$

Specificity (True Negative Rate) measures the percentage of actual negative cases that have been correctly identified.

$$Specificity = \frac{TN}{(TN + FP)} \quad (4)$$

F1-score computes the harmonic mean between precision and recall, balancing both metrics:

$$F1 \text{ Score} = \frac{2TP}{2TP + FP + FN} \quad (5)$$

Matthews Correlation Coefficient (MCC) quantifies classifier performance across all four confusion matrix categories (TP, TN, FP, FN) with a value between -1 (total disagreement) and $+1$ (perfect prediction). Unlike the F1-score, Matthew’s Correlation

Coefficient (MCC) takes into account the imbalance between classes and works well even when one class is much larger than the other.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

Area Under the ROC Curve (AUC) evaluates the classifier's ability to distinguish between classes across all possible thresholds. AUC is typically calculated from the ROC curve using numerical integration. The ROC curve plots the true positive rate (recall) against the false positive rate ($1 - \text{specificity}$), and AUC ranges from 0.5 (random guessing) to 1.0 (perfect discrimination). AUC is threshold-independent, making it ideal for comparing models (Hanley & McNeil, 1982) [58].

There is no closed-form formula for TP, TN, FP, or FN, but it represents the following:

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (7)$$

4. Results

4.1. Test Set Performance

The initial evaluation of the classifiers on the test dataset, using default parameters, reveals significant differences in performance among the models. The SVM achieved the highest overall accuracy of 0.8412 and an AUC of 0.8173, indicating strong general discrimination capability. Meanwhile, RF and XGBoost also performed reasonably well, with accuracies of 0.8235 and 0.8294, respectively.

However, when focusing on detecting the target class (class 1, which represents potsherds), we observe a trade-off between precision and recall. MLC exhibited the highest precision for class 1 at 0.7059, indicating fewer false positives; however, its recall was relatively low at 0.4068, suggesting that many class 1 instances were missed. In contrast, XGBoost demonstrated the highest recall for class 1 at 0.8571, successfully detecting most positive instances, but this came at the cost of very low precision (0.1765), indicating many false positives.

RF and SVM delivered a more balanced performance for class 1, with moderate precision (0.2941) and higher recall values (0.625 for RF and 0.7692 for SVM). AdaBoost showed relatively lower overall performance, with an accuracy of 0.7882 and modest metrics for class 1, suggesting that the default parameters may not effectively capture class 1 instances. A summary of the key performance metrics for each model on the test data is illustrated in Table 4 below.

Table 4. Classification metrics for each model on the test data. (Bold values indicate the best-performing model for each evaluation metric).

Model	Accuracy	Kappa	AUC	MCC	Precision Class 1	Recall Class 1	Specificity Class 1
RF	0.82	0.31	0.75	0.34	0.29	0.63	0.84
MLC	0.74	0.35	0.77	0.38	0.71	0.41	0.91
SVM	0.84	0.35	0.82	0.41	0.29	0.77	0.85
ADA	0.79	0.17	0.73	0.19	0.21	0.44	0.82
XGB	0.84	0.24	0.71	0.34	0.18	0.86	0.83

4.2. Validation Set Performance

Using default settings, the validation dataset results (Table 5) show some clear trends and a few performance drops for certain classifiers compared to the test set. The SVM achieved the highest accuracy of 0.84 and an AUC of 0.8275. It also had a strong recall of 0.75 and an average precision of 0.30. This means SVM is the most reliable model for

detecting class 1. The RF model had a moderate accuracy of 0.80 and a decent specificity of 0.8261. However, its precision for class 1 was low at 0.20, making it hard to label ceramics correctly. The MLC model focused on precision, scoring 0.60, but its recall was only 0.33. This resulted in fewer positive detections but also fewer false positives.

Table 5. Classification metrics for each model on the validation dataset. (Bold values indicate the best-performing model for each evaluation metric).

Model	Accuracy	Kappa	AUC	MCC	Precision Class 1	Recall Class 1	Specificity Class 1
RF	0.8	0.19	0.71	0.22	0.2	0.50	0.83
MLC	0.68	0.23	0.73	0.25	0.6	0.33	0.88
SVM	0.84	0.35	0.83	0.41	0.3	0.75	0.85
ADA	0.7	0.03	0.61	0.03	0.2	0.22	0.80
XGB	0.78	0.07	0.65	0.08	0.1	0.33	0.81

Neither boosting method performed well. AdaBoost had the lowest metrics overall, with only 0.70 accuracy, poor recall of 0.22, and almost negligible MCC and kappa values, showing it is not very robust. XGBoost also had trouble with a low precision of 0.10 and a modest recall of 0.33, showing it struggles to identify ceramics using the default settings.

4.3. Comparative Analysis: Test vs. Validation Performance

A comparison of the test (Table 4) and validation (Table 5) results (Figure 9) reveals critical differences in model generalization. Overall accuracy remained stable (SVM: 0.84 in both sets; RF: 0.82 to 0.80), but minority-class metrics showed substantial variation. Boosting algorithms experienced dramatic degradation: XGBoost's recall dropped from 0.86 to 0.33 (−62%), while AdaBoost declined from 0.44 to 0.22 (−50%). These drops reflect overfitting to training-site characteristics and recall to distribution shifts between archeological contexts.

In contrast, SVM demonstrated exceptional stability (recall: 0.77 to 0.75; MCC: 0.41 to 0.41), though at lower absolute detection rates. The MCC metric particularly illuminates generalization capacity: SVM's consistent performance indicates transferable predictive power, while boosting algorithms' MCC collapse (AdaBoost: 0.19 to 0.03; XGBoost: 0.34 to 0.08) reveals site-specific dependencies under default parameters.

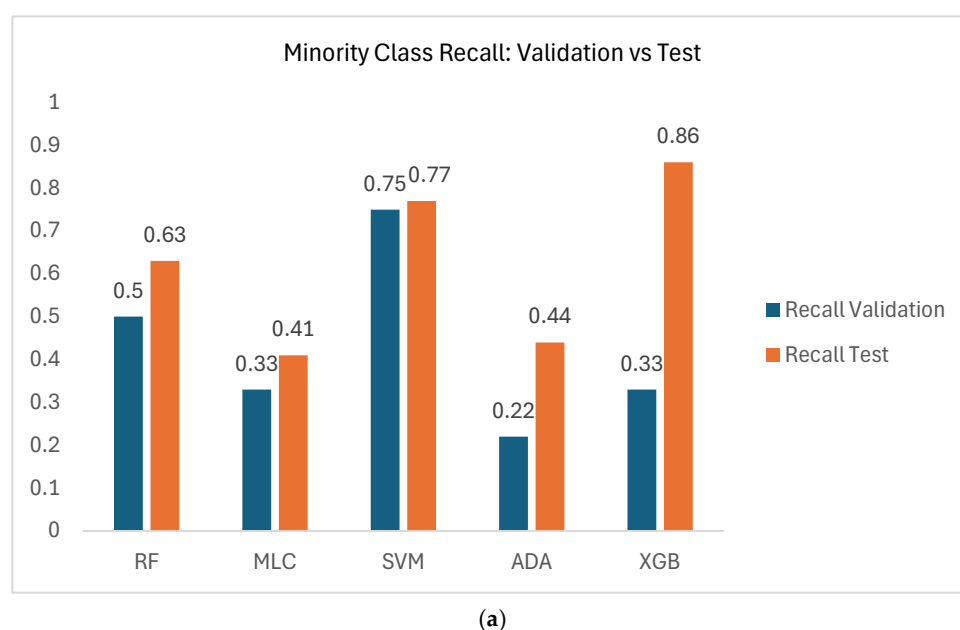


Figure 9. Cont.

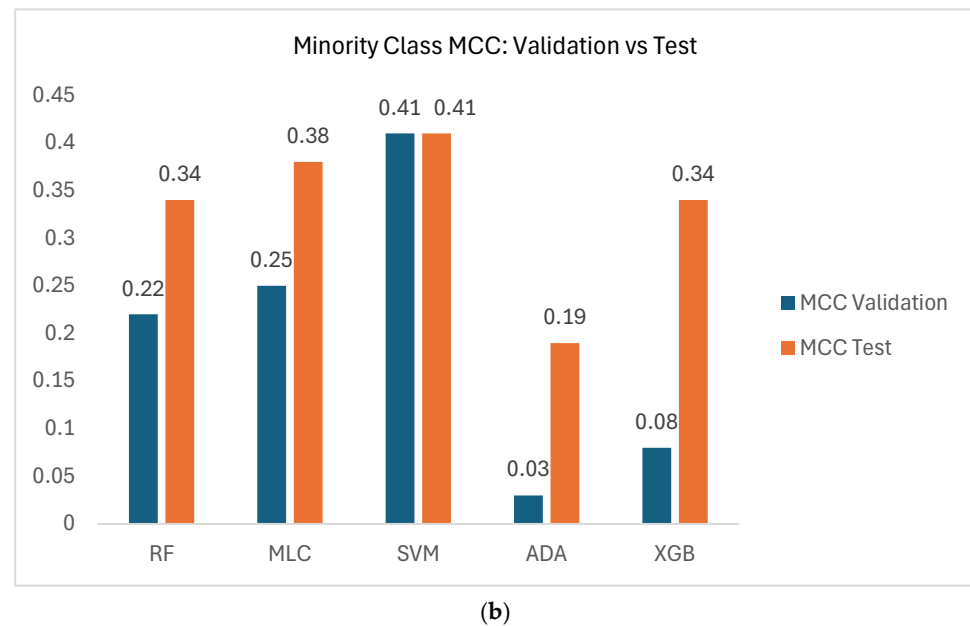


Figure 9. Classifier performance under default parameters across test and validation datasets: (a) recall for the minority class (ceramics), (b) Matthews Correlation Coefficient (MCC).

These patterns carry important practical implications. Validation results using default parameters underscore that boosting algorithms require site-specific threshold optimization (Section 4.3) to maintain operational effectiveness. When properly calibrated, these algorithms recover performance—AdaBoost achieves 0.88 recall after tuning—demonstrating that threshold selection is essential for translating models across varied archeological landscapes.

4.4. Threshold Tuning Analysis

Threshold tuning for boosting classifiers has a significant impact on the detection of minority classes. In the case of AdaBoost, adjusting the threshold to maximize the F1-score resulted in an optimal threshold of 0.34. This adjustment improved recall and balanced the detection of class 1 instances, although it also led to an increase in false positives. On the other hand, selecting the threshold to maximize the Matthews Correlation Coefficient (MCC) resulted in a higher threshold of 0.6 (Figure 10). This approach favored precision and specificity but drastically reduced recall, causing a substantial number of true positives to be missed. This scenario highlights the precision-recall trade-off that comes with threshold selection.

For XGBoost, both the best F1 and MCC thresholds were low (0.11), reflecting the need to lower the default threshold of 0.5 to adequately capture the minority class (Figure 11). At this threshold, XGBoost demonstrated a better balance between true positive detection and false positive control, improving the recall for minority classes without significantly increasing false positives.

AdaBoost, which initially struggled with both low recall and precision under default parameters, achieved a very high recall of 0.88 after tuning, correctly detecting nearly thirty out of thirty-four potsherds' instances in the test dataset. Although this improvement came at the cost of lower specificity due to more false positives, the model proved highly effective for maximizing class 1 detection, which is a key objective of this study. XGBoost also benefited from threshold adjustment, improving recall to 0.68 while maintaining a stronger balance with specificity (0.74) compared to AdaBoost.

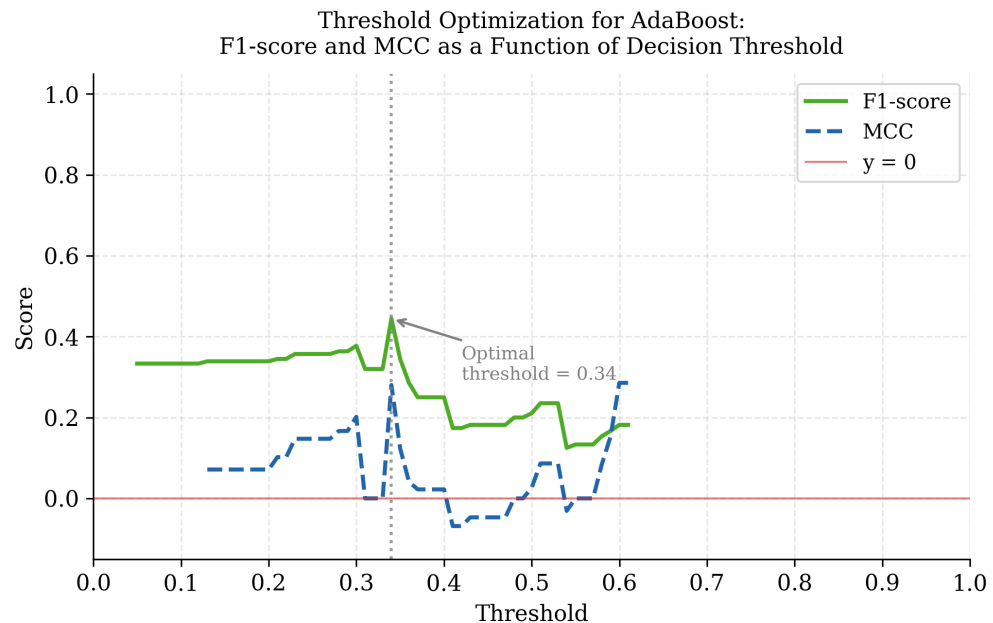


Figure 10. F1-score and MCC across decision thresholds for AdaBoost, illustrating the trade-off between classification performance metrics over the full threshold range [0, 1]. The vertical dashed line indicates the optimal threshold.

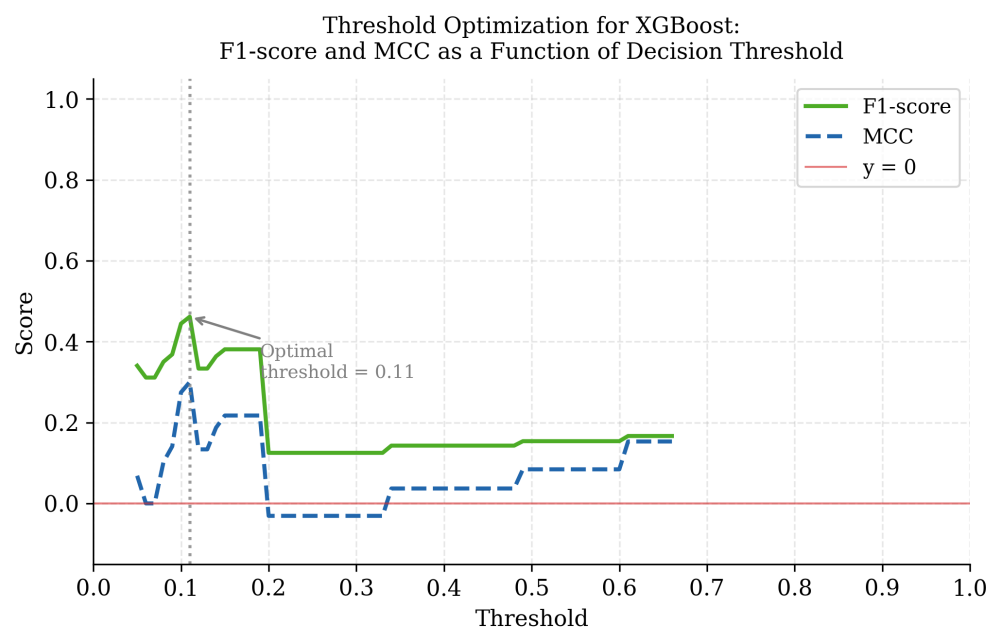


Figure 11. F1-score and MCC across decision thresholds for XGBoost, illustrating the trade-off between classification performance metrics over the full threshold range [0, 1]. The vertical dashed line indicates the optimal threshold.

Overall, the tuned boosting models (Table 6) demonstrated clear improvements over their default configurations and, importantly, enhanced class 1 detection relative to the baseline classifiers. AdaBoost proved particularly effective for prioritizing recall (capturing most ceramics but also producing a greater number of false positives), while XGBoost offered a more balanced trade-off between detecting ceramics and limiting false positives. These results underscore the value of threshold optimization in adapting boosting algorithms to imbalanced archeological remote sensing data, and they highlight boosting as a promising approach for ceramic detection when the primary goal is to minimize missed positives.

Table 6. Classification metrics with Adaboost and XGBoost threshold tuning.

Model	Accuracy	MCC	Precision Class 1	Recall Class 1	Specificity Class 1
RF	0.8	0.22	0.20	0.50	0.83
MLC	0.68	0.25	0.60	0.33	0.88
SVM	0.84	0.41	0.30	0.75	0.85
ADA	0.62	0.27	0.33	0.88	0.56
XGB	0.72	0.32	0.39	0.68	0.74

Overall, these results emphasize that tuning thresholds is essential for boosting classifiers, especially when the aim is to detect rare classes. This allows the model to be adjusted for either higher recall (F1 optimization) or higher precision and overall correlation (MCC optimization). The difference in optimal thresholds between AdaBoost (0.34) and XGBoost (0.11) highlights how these two models learn differently. XGBoost uses gradient-boosted trees, which tend to make more cautious predictions.

4.5. Detection of Potsherds in High-Resolution RGB Mosaic

At the final post-processing stage, a structured workflow was implemented in ArcGIS Pro 3.5.0 to refine the classification results and prepare them for subsequent spatial analysis. The target class (potsherds) was first extracted from the classified raster using the Reclassify tool. The Majority Filter was applied to minimize classification noise and enhance spatial coherence. The filtered raster was then converted into polygon features to delineate contiguous ceramic areas.

After the classified raster was converted to polygon features, an additional field named Circularity was computed within the polygon dataset to describe the geometric regularity of each fragment quantitatively. The circularity index was calculated using the standard formula:

$$Circularity = \frac{4\pi \text{Shape Area}}{\text{Perimeter}^2} \quad (8)$$

where values closer to 1 indicate shapes approaching perfect circles, while lower values correspond to more elongated or irregular fragments.

In ArcGIS Pro 3.5.0, the field was created using the Field Calculator with the following expression:

$$4\pi \text{ Shape Area} / \text{Shape length}^2 \quad (9)$$

Circularity was selected as the primary shape descriptor because it provides a robust and scale-independent measure of geometric regularity, directly linked to the morphological integrity of ceramic fragments.

A definition query was applied to the dataset to isolate a specific subset of potsherds for detailed analysis. This query established quantitative criteria to filter fragments based on their size and morphological characteristics. This criterion targeted sherds exhibiting a specific range of circularity ($Circularity \geq 0.6$ and $Circularity \leq 0.75$), thereby focusing the study on fragments that are more regularly shaped. The resulting polygons were subsequently transformed into point features by calculating polygon centroids. This reproducible workflow produces both vector polygons and points, thereby facilitating accurate mapping, quantitative analysis, and integration with other spatial datasets. Examples from the results are presented in Figure 12, where manually selected ground-truth samples are compared with the prediction output obtained after model training.

The outcomes of this analysis are shown in the Annex (Figures A1–A5, in Appendix A). Potsherds are shown as colored dots, whereas the different colors indicate the classifier used for the detection. The field survey recorded different numbers of potsherds across the sampled grids. In grid DM3_HR, 26 potsherds were documented, resulting in a VIS_count of 86.7. In grid DM3_HS, 80 potsherds were recorded with a VIS_count of

160. Grid DM3_HU contained 86 potsherds with a VIS_count of 215, while grid DM3_HT documented 51 potsherds with a VIS_count of 255.



(a)



(b)

Figure 12. *Cont.*



(c)



(d)

Figure 12. *Cont.*



(e)



(f)

Figure 12. Examples from two different samples of potsherd with a true–false detection, and true–true detection for each classifier, (a) for orthophoto, (b) for RF, (c) for MLC, (d) for SVM, (e) for Adaboost, and (f) for XGBoost. (Colored dots show ceramics correctly detected by each classifier.; examples are highlighted with red circles).

Across the entire area of interest (AOI), 243 potsherds were recorded during the field survey, with a cumulative VIS_count of 716.7. These counts derive from systematic pedestrian

surveys and are treated as operational ground truth, rather than as an exhaustive census of all ceramic fragments physically present in the area. In comparison, the classifiers produced predictions ranging from 210 to 1965 detections, depending on the algorithm and classification threshold applied (Table 7). UAV-based detections represent probabilistic candidate locations identified under a recall-oriented exploratory strategy and should not be interpreted as direct artifact counts. The detected points may include both ceramic fragments overlooked during pedestrian survey and false positives (e.g., stones or soil features), which require targeted field verification to resolve. This discrepancy between automated classifier outputs and ground-truth field observations is further examined in Section 5.

Table 7. Summarized results from the AOI foot survey and classifier predictions. (Note: classifier results correspond to the total detections of potsherds predicted across the AOI for each algorithm).

Survey Type		Foot Survey		Detection of Potsherds with Machine Learning Applied to UAV Image				
Grid	Potsherd Count	VIS Count	Density (m ²)	RF	MLC	SVM	AdaBoost	XGBoost
DM_HS	80	160	1.60	83	36	85	500	601
DM_HR	26	86.7	0.87	92	66	69	314	446
DM_HU	86	215	2.15	106	25	118	452	488
DM_HT	51	255	2.55	598	83	501	299	430
AOI Total	243	716.7	1.79	879	210	773	1565	1965

5. Discussion

The primary contribution of this research lies not in algorithmic innovation but in developing a methodologically transparent framework for adapting existing ML techniques to archeological operational requirements. Heritage landscape documentation demands detection strategies that prioritize comprehensive coverage over precision, accepting higher false positive rates as an acceptable cost for minimizing undetected artifacts. This study demonstrates how threshold optimization transforms general-purpose classifiers into archeologically interpretable prospection tools.

Detecting surface potsherds in remote sensing data is challenging due to class imbalance, limited annotated datasets, and spectral similarities between ceramics and surrounding soil [9]. Differences between UAV-based classifications and ground-truth counts (Table 7) reflect these limitations and the significant variations in the results in the different environments (i.e., grids). Higher spatial resolution, achievable by lowering flight altitude, could improve detection, though at the cost of coverage. Similarly, enhanced spectral imaging may help distinguish potsherds from soil, while field-walkers face analogous difficulties under vegetation, rocks, dust, or variable lighting. Although multispectral imagery can improve material separability, the principal limitation in potsherd detection lies in extreme class imbalance rather than sensor capability alone. The present study, therefore, emphasizes methodological optimization for minority-class detection, demonstrating that recall-oriented boosting strategies remain essential even under constrained spectral conditions.

This study implements a semi-automatic detection approach, extending and validating previous works applied and evaluated at the Alampra and Kophinou sites in Cyprus [46,47]. The results demonstrate that boosting algorithms, particularly AdaBoost and XGBoost, significantly improved minority-class detection despite working with a limited training sample. While these methods initially underperformed under default parameter settings, threshold optimization substantially enhanced their recall to archeological materials. Field surveys recorded 215 potsherds, whereas classifier predictions ranged from 215 to 2033 detections. This considerable variation reflects fundamental trade-offs inherent

in classification approaches. Boosting models tend to produce elevated detection at lower probability thresholds, primarily due to spectral confusion between ceramic materials and similar soil backgrounds. Conversely, more conservative classifiers such as Maximum Likelihood Classification (MLC) often fail to identify subtle or partially buried ceramic fragments, resulting in underestimating archeological presence.

Post-processing workflows in ArcGIS Pro 3.5.0—including spatial filtering, reclassification, and conversion of raster outputs to polygon and centroid geometries—proved essential for visual validation and systematic comparison with ground-truth field data. These procedures enabled more detailed interpretation of classifier outputs and facilitated the identification of spatially coherent detection clusters that correspond to genuine archeological features.

5.1. Sensitivity Archeological Detection (SAD)

Building on the boosting performance results, we further assessed the performance of each algorithm's detection capacity by developing the so-called "Sensitivity Archeological Detection (SAD)", which quantifies recall to potential ceramic sites while considering error and variability across survey areas. The goal was to prioritize high recall and coverage over strict agreement with visual counts. The SAD is defined as a linear combination of three normalized metrics: (1) the mean ratio of ML detections to VIS counts (MeanRatio)—a positive indicator where higher values reflect stronger exploratory detection, (2) the root mean squared error (RMSE) as a measure of absolute deviation from VIS counts—a negative indicator where lower values are better, and (3) the standard deviation across areas (StdDev)—a negative indicator reflecting consistency between the different areas (grids), with lower values being preferable.

Normalization was performed by dividing each positive metric by its maximum value

$$NormMeanRatio = MeanRatio / \max(MeanRatio) \quad (10)$$

and inverting the negative metrics

$$NormRMSE = 1 - RMSE / \max(RMSE) \quad (11)$$

$$NormStdDev = 1 - StdDev / \max(StdDev) \quad (12)$$

ensuring all normalized values ranged from 0 to 1 and pointed in the same direction (1 = best). The final SAD was then calculated as a weighted sum to favor exploratory recall as follows (Equation (13)):

$$SAD = 0.6NormMeanRatio + 0.3NormRMSE + 0.1NormStdDev \quad (13)$$

The weighting scheme (0.6, 0.3, 0.1) was established through expert judgment informed by archeological remote sensing priorities for detecting ceramic surfaces. The specific values reflect the relative importance assigned to each criterion: detection recall ($w_1 = 0.6$) is prioritized as approximately twice as critical as spatial precision ($w_2 = 0.3$), which in turn is three times more important than cross-plot consistency ($w_3 = 0.1$). This hierarchical structure reflects the operational requirement to identify candidate ceramic locations for subsequent field verification rather than achieve precise quantitative agreement with visual counts. These weights are context-dependent and may need to be adjusted for different survey environments or research objectives.

The resulting SAD values indicate that, under this evaluation framework, XGBoost ranks highest (SAD = 0.60), followed by AdaBoost (SAD $\approx$ 0.58), supporting XGBoost as

the most exploratory detector, while SVM and RF are more conservative, and MLC highly under-detects (Table 8).

Table 8. Normalized performance metrics and SAD for each ML algorithm against VIS counts.

Algorithm	Mean Ratio	StdDev	RMSE	Norm. Mean Ratio ¹	Norm. Std Dev ²	Norm. RMSE ³	SAD ⁴
SVM	0.96	0.68	137.72	0.30	0.56	0.58	0.41
MLC	0.36	0.28	142.73	0.11	0.82	0.56	0.32
RF	1.10	0.87	184.04	0.34	0.44	0.44	0.38
AdaBoost	2.51	1.09	237.37	0.78	0.30	0.28	0.58
XGBoost	3.21	1.55	327.39	1.00	0.00	0.00	0.60

¹ NormMeanRatio = MeanRatio/3.21 (maximum MeanRatio, XGBoost). ² NormStdDev = 1 – (StdDev/1.55) (maximum StdDev, XGBoost). ³ NormRMSE = 1 – (RMSE/327.39) (maximum RMSE, XGBoost). ⁴ SAD is the weighted sum as defined above. Interpretation/Practical Implications.

Based on the SAD, XGBoost achieved the highest score (0.60), followed by AdaBoost (0.58), indicating that—when exploratory detection of potential ceramic sites is prioritized—XGBoost provides the best balance between maximal coverage (recall) and acceptable error. In contrast, SVM and RF are more conservative, while MLC consistently shows low exploratory capability. Therefore, for studies aiming to maximize potential finds for subsequent field verification, XGBoost (or AdaBoost as an alternative) is recommended, with careful qualitative validation of areas showing high detection density.

5.2. Rationale for Pixel-Based Classification

The choice to employ pixel-based classification with boosting algorithms, rather than object-based image analysis (OBIA) or deep learning object detection frameworks, reflects fundamental methodological considerations specific to archeological ceramic detection from UAV imagery. At operational flight altitudes (10–15 m) with standard RGB sensors, surface ceramic sherds typically span 2–5 pixels. An additional factor supporting the pixel-based approach is spectral mixing. Due to the very small physical size of surface potsherds, individual pixels frequently capture a mixture of ceramic material and the surrounding soil matrix. This mixed spectral response reduces the effectiveness of object-based segmentation and bounding-box detection approaches and reinforces the suitability of pixel-level classification strategies that operate directly on subtle spectral variation. This sub-object scale presents inherent limitations for both OBIA and deep learning approaches. OBIA segmentation algorithms require meaningful object boundaries that do not exist at this resolution, while deep learning object detection frameworks (YOLO, Faster R-CNN) are optimized for objects spanning minimally 10–20 pixels and rely on bounding box representations that become geometrically unstable at pixel-scale dimensions. The pixel-based approach more appropriately matches the spatial scale at which ceramic artifacts appear in the imagery.

Additionally, deep learning approaches face practical constraints in archeological contexts. Thousands of annotated examples are typically required for robust deep learning model training, and transfer learning from standard computer vision datasets does not readily generalize to the distinctive spectral-textural characteristics of weathered ceramics. The post-processing protocols employed in this study, including spatial filtering and minimum mapping unit thresholds, effectively addressed salt-and-pepper noise and transformed pixel classifications into coherent archeological feature candidates.

While future research may benefit from evaluating hybrid approaches or emerging architectures designed for small object detection, the pixel-based boosting methodology employed here offers a practical and theoretically grounded approach given current data and technological constraints.

5.3. Robustness Analysis and Methodological Limitations

5.3.1. Sensitivity Analysis of SAD Metric Weights

The hierarchical weighting scheme (0.6, 0.3, 0.1) employed in the SAD metric reflects established archeological remote sensing principles: $w_1 = 0.6$ (Mean Ratio) prioritizes detection sensitivity as the primary objective in archeological prospection; $w_2 = 0.3$ (RMSE) acknowledges that spatial precision is important but secondary to detection capability; and $w_3 = 0.1$ (StdDev) recognizes that consistency across survey areas is desirable but should not override superior mean performance.

However, as the SAD metric (Equation (13)) is a linear composite of three normalized criteria, weight selection may introduce subjectivity. To assess the robustness of SAD rankings to weight uncertainty, we conducted a sensitivity analysis using established guidance for composite indicators and multi-criteria decision analysis (MCDA), which recommends Monte Carlo uncertainty propagation to quantify ranking stability and potential rank reversals.

We generated $N = 200,000$ admissible weight vectors $w = (w_{MR}, w_{RMSE}, w_{SD})$ constrained to $w_i \geq 0$ and $\sum w_i = 1$, using a Dirichlet distribution centered on the nominal weights (0.6, 0.3, 0.1). For each sampled weight vector, SAD scores were recomputed for all classifiers, and the resulting rankings were recorded.

Under these local perturbations, the Top-2 set (XGBoost, AdaBoost) remained unchanged in 98.32% of samples; XGBoost ranked first in 69.99% and AdaBoost in 29.94%, indicating that rank reversal is largely confined to swapping between these two algorithms. An analytic threshold analysis further shows that for $w_{MR} \geq 0.577$, XGBoost is guaranteed to remain top-ranked irrespective of the split between RMSE and StdDev weights. The nominal configuration ($w_{MR} = 0.6$) lies within this robust region (Figure 13).

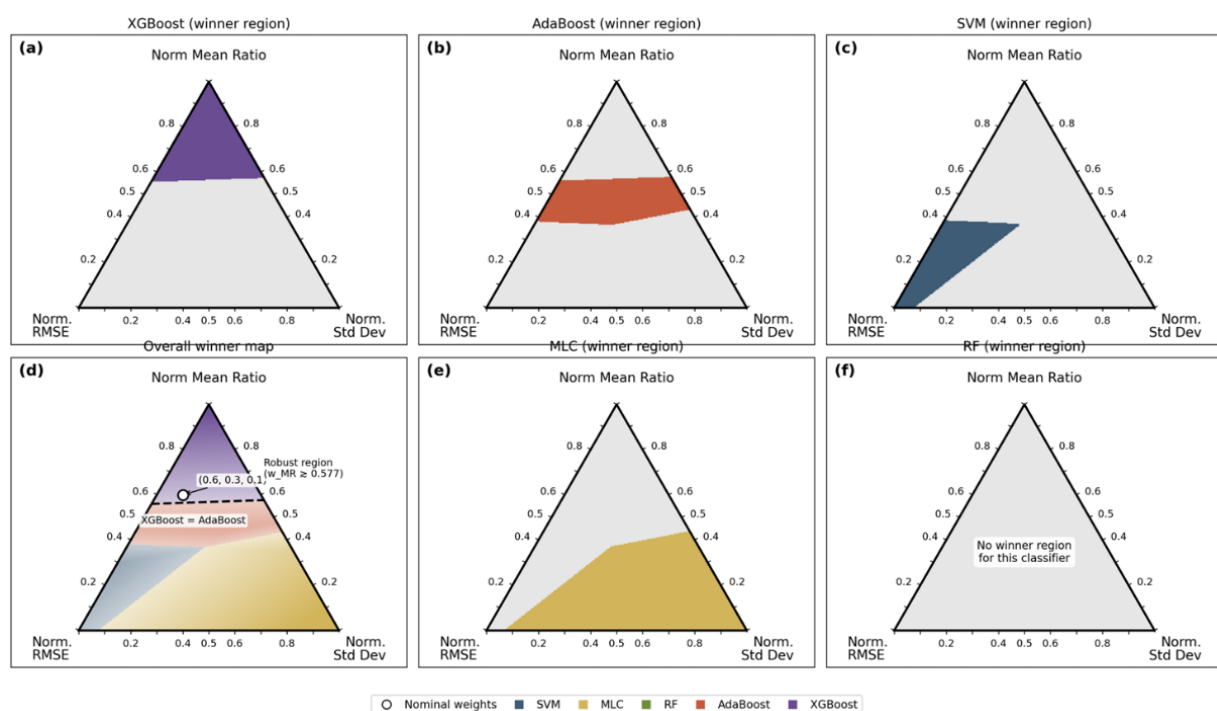


Figure 13. Weight-sensitivity maps of SAD rankings under varying weight configurations. Panels (a–c) show the regions of the weight simplex (w_{MR}, w_{RMSE}, w_{SD}) with $w_i \geq 0$ and $\sum w_i = 1$, where XGBoost, AdaBoost, and SVM, respectively, achieve the highest SAD score (winner regions). Panel (d) presents the overall winner map (The colored regions correspond to the winner model in each panel, as indicated in the legend) and marks the nominal weight configuration (0.6, 0.3, 0.1), together with the analytical boundary where XGBoost and AdaBoost swap ranks, and the robust region $w_{MR} \geq 0.577$ in which XGBoost remains top-ranked regardless of the split between the other two

weights. Panel (e) shows the winner region for MLC. Panel (f) indicates that RF does not attain rank #1 for any feasible weight combination given the normalized criteria in Table 8.

Overall, SAD conclusions are stable to realistic uncertainty around the proposed weight structure, while transparently indicating the conditions under which the top rank may switch between XGBoost and AdaBoost.

5.3.2. Spatial Generalizability Considerations

The threshold optimization is sensitive to specific data distributions, raising questions about cross-site generalizability. The proposed methodology was tested across four spatially distinct survey grids (DM_HS, DM_HR, DM_HU, DM_HT) within the Megara study area, which exhibit substantial variability in potsherd density (0.87–2.55 m²), total counts (26–86), and environmental conditions (Table 7). The standard deviation component in the SAD metric ($w_3 = 0.1$) explicitly quantifies inter-grid consistency, providing a measure of algorithm stability across these variable micro-environments.

Although these grids were not employed as independent train/test partitions for threshold determination, the consistent superior performance of boosting algorithms across grids with nearly threefold density variation suggests reasonable robustness to local site characteristics within a single landscape context.

However, it is acknowledged that full generalizability requires more extensive validation. Future work should include the following: (1) cross-site validation using geographically independent archeological areas (e.g., training on Cyprus sites and testing on Greek sites, or vice versa); (2) grid-based cross-validation where thresholds are optimized on subset grids and validated on spatially separated held-out grids within the same landscape; and (3) systematic testing across different soil types, vegetation cover densities, and seasonal illumination conditions.

The current results demonstrate proof-of-concept performance across variable micro-environments within a single archeological landscape, but we recognize that establishing broader applicability necessitates testing across environmentally and geographically distinct study areas with independent training and validation datasets.

5.3.3. Methodological Boundary Conditions and Applicable Scope

The effectiveness of the UAV-based detection method is shaped by environmental and operational factors that set its limits.

The method works best when vegetation is sparse and ceramic surfaces are directly visible, such as in plowed fields or bare soil. Dense vegetation or crop cover greatly reduces detection. Strong spectral contrast between ceramics and soil improves results. Midday data collection minimizes shadows, while morning or evening light increases shadow variability. Recently disturbed soils exposed more artifacts; older, compacted soils may hide them.

Detection is limited when canopy cover exceeds 30%, when ceramics and soil have similar colors, or when the terrain is uneven. Stones, debris, and variable soil moisture can cause false positives. Small ceramic fragments (under 2 cm²) are hard to detect at standard survey heights, nearing the limits of UAV sensors.

This method is an exploratory tool to identify high-potential areas for later ground checks, not for exact artifact counts. It favors high recall and accepts more false positives to maximize coverage. It is best used for rapid, large-area screening where follow-up fieldwork is possible and precise artifact counts are less important.

These findings apply mainly to Mediterranean landscapes similar to Megara—agricultural areas with soil disturbance, low-to-moderate vegetation, and ceramics that contrast with local soils. Densely vegetated or spectrally uniform areas would need different methods.

6. Conclusions

This study demonstrates that threshold-optimized boosting classifiers can be effectively adapted for archeological ceramic detection, providing an operational framework that maximizes recall for exploratory landscape survey. While this approach generates higher false positive rates than conventional classifiers, this trade-off aligns with archeological prospection priorities where comprehensive coverage outweighs precision. The integration of UAV-based detection with systematic ground-truth validation ensures archeologically interpretable outputs and guards against over-reliance on algorithmic predictions. This methodological framework acknowledges that remote sensing serves as a prospection tool rather than a definitive inventory method, requiring expert validation to transform statistical predictions into meaningful archeological knowledge. Future refinements should explore ensemble combinations, spatial-textural feature integration, multispectral data incorporation, and cross-landscape validation to enhance the framework's robustness and transferability across diverse Mediterranean archeological contexts.

The overall findings are aligned with previous works, such as those reported by Orengo and Garcia-Molsosa [3] and Agapiou et al. [9], indicating that low-altitude remote sensing sensors can play a significant role in archeological field surveys. These methods can potentially support future archeological fieldwork cost-effectively, particularly in situations with limited research timelines and a need to document rapidly disappearing archeological landscapes for many reasons.

This study validates a semi-automatic detection methodology that adapts existing boosting algorithms (AdaBoost, XGBoost) through threshold optimization for archeological ceramic prospection. Working with limited training data (15% subset), the framework demonstrates that recall-oriented optimization can enhance minority-class detection despite elevated false positive rates—a methodological trade-off appropriate for exploratory archeological survey where undetected artifacts represent greater information loss than controlled over-prediction. The results confirm that minority-class recall metrics provide more archeologically meaningful performance indicators than overall accuracy in prospection contexts, prioritizing comprehensive site coverage over precise artifact enumeration.

Optimizing the threshold proved crucial for improving classifier recall during prospection. The optimal thresholds were 0.34 for AdaBoost and 0.11 for XGBoost, resulting in detection-to-ground-truth ratios of 2.5 and 3.2, respectively. While these thresholds substantially increased detection rates, they also produced higher false positives—a deliberate trade-off that aligns with the exploratory objectives of archeological landscape survey. These findings highlight the need to move beyond default classification parameters and to adopt data-driven optimization strategies that explicitly balance recall and specificity according to survey objectives.

The study establishes a methodological framework for archeological remote sensing that acknowledges fundamental trade-offs between detection recall and precision. The Sensitivity Archeological Detection (SAD) introduced here provides a metrics-based approach for evaluating algorithms based on exploratory detection capacity rather than strict accuracy, making it applicable to rare-object prospection problems in archeological landscape investigations. However, integrating UAV-based multispectral detection with systematic ground-truth surveys remains essential to ensure archeologically realistic interpretation and guard against over-reliance on algorithmic outputs.

This research demonstrates that UAV-based boosting classifiers can effectively flag areas for targeted archeological field investigation, enabling archeologists to prioritize survey efforts more efficiently than traditional systematic coverage alone. The approach functions as a prospection tool requiring expert validation rather than a standalone detection system.

Future advancements should explore ensemble combinations of classifiers, the integration of spatial and textural features, enhanced multispectral data, and validation across diverse landscapes in order to improve the robustness and generalizability of ceramic prospection from remote sensing imagery. Recent advances in remote sensing have increasingly explored few-shot and meta-learning approaches to address scenarios with limited training data [59,60]. While these methods are promising for annotation-scarce settings, the present study focuses on a complementary challenge—extreme class imbalance—demonstrating that recall-oriented boosting and threshold optimization remain critical for effective minority-class detection even outside episodic learning frameworks.

This work supports the growing application of machine learning in archeology and represents a significant step toward developing practical, operational computational tools for archeological landscape prospection using low-altitude aerial imagery. Boosting classifiers, combined with threshold optimization and hybrid validation workflows, offer archeologists an accessible means to enhance field survey efficiency while preserving the essential role of expert interpretation in transforming statistical predictions into meaningful archeological knowledge.

Author Contributions: Conceptualization, A.A. (Argyro Argyrou); methodology, A.A. (Argyro Argyrou); software, A.A. (Argyro Argyrou), F.F. and A.P.; validation, E.F., A.P. and D.D.A.; formal analysis, A.A. (Argyro Argyrou); investigation, A.A. (Argyro Argyrou); resources, A.A. (Athos Agapiou), E.F. and F.F.; data curation, A.A. (Argyro Argyrou); writing—original draft preparation, A.A. (Argyro Argyrou); writing—review and editing, all authors.; visualization, A.A. (Argyro Argyrou); supervision, A.A. (Athos Agapiou); project administration, A.A. (Athos Agapiou); funding acquisition, A.A. (Athos Agapiou). All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Cyprus University of Technology internal funding, as well as the CONNECTING project. The CONNECTING project has received funding from the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation of Cyprus under contract SMALL SCALE INFRASTRUCTURES/1222/0062. Disclaimer: Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the UKRI. Neither the European Union nor the UKRI can be held responsible for them.

Data Availability Statement: All classifiers were implemented using R software, while post-processing was conducted within the ArcGIS Pro 3.5.0 environment. The complete R code for machine learning classification, threshold optimization, and analytical workflows, together with the associated results tables, is openly available in the Cyprus University of Technology institutional repository at: <https://doi.org/10.82463/6xww-1p15>. The RGB mosaic of the Dami area in Megara, Greece, is available upon request from E.F. and F.F. (this mosaic was created under the WeMALP project). The WeMALP project is a collaborative initiative that brings together researchers and institutions from Greece and abroad. It has been operating since 2019 under the scientific direction of Christina Merkouri and Panagiota Avgerinou from the Ephorate of Antiquities of West Attica, as well as Emeri Farinetti from the Department of Humanities at the University of Roma Tre, with the support of the Italian Archaeological School in Athens.

Acknowledgments: The authors acknowledge the ENSURE project (innovative survey techniques for detection of surface and sub-surface archeological remains) and CUT internal funding, as well as the Connecting project (SMALL SCALE INFRASTRUCTURES/1222/0062). This paper is part of the PhD dissertation of A. Argyrou, supervised by A. Agapiou, and co-supervised by A. Papakonstantinou and D. Alexakis. A. Argyrou would like to thank Emeri Farinetti for providing the data and the archeological context and Federico Fasson for the drone images. The authors acknowledge the use of language editing tools (Grammarly (2025)) and AI-based assistants (ChatGPT (version GPT-4o) by OpenAI, Claude (version Sonnet 4.5) by Anthropic) to improve the readability and grammar of the manuscript. The authors have reviewed and edited the output and take full responsibility for the

content of this publication. The authors would like to thank the three anonymous reviewers for their criticism and suggestions for improvements during the review process.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

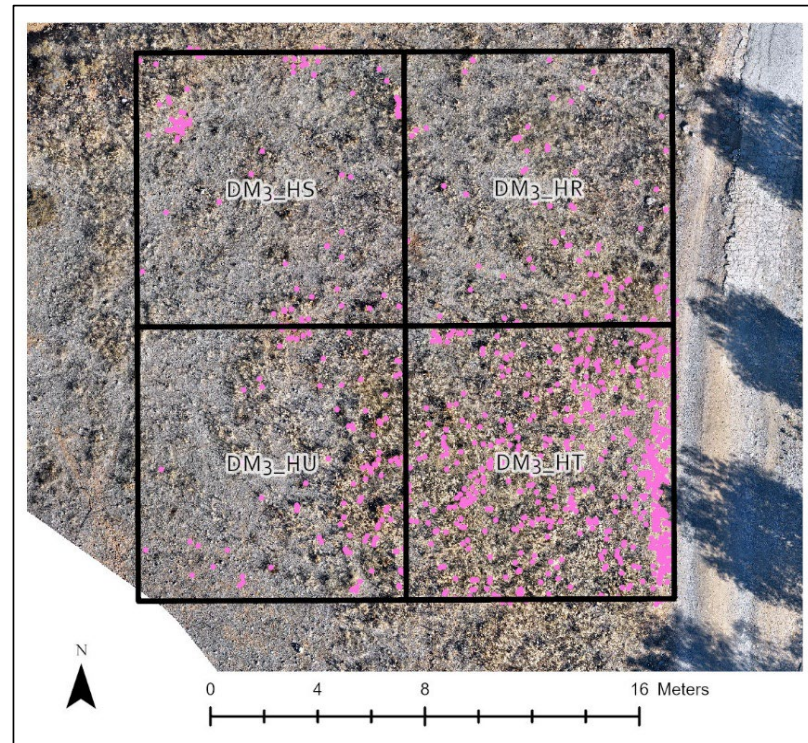


Figure A1. Potsherd detection (points—pink color) over the AOI from the Random Forest classification.

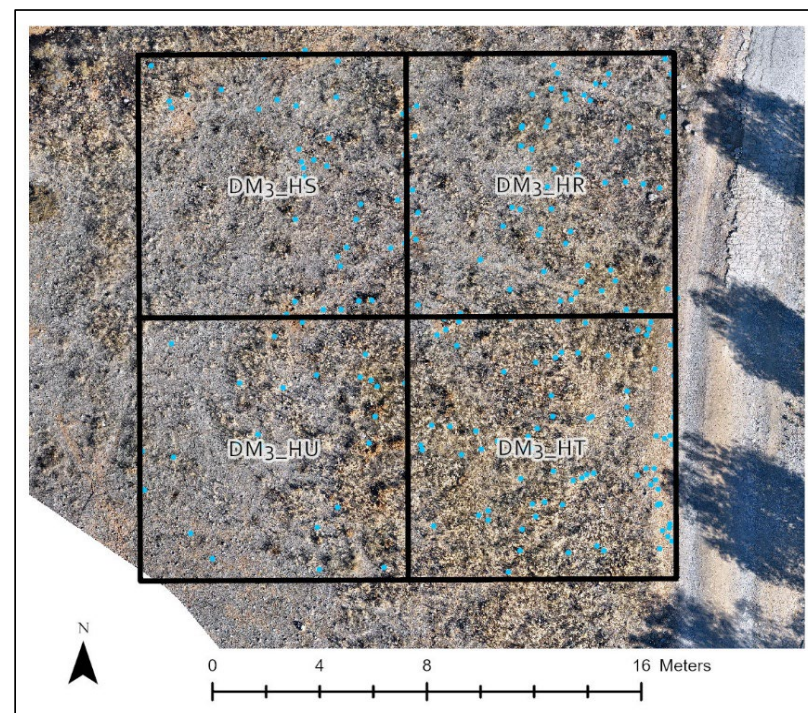


Figure A2. Potsherd detection (points—light cyan color) over the AOI from the Maximum Likelihood Classification.

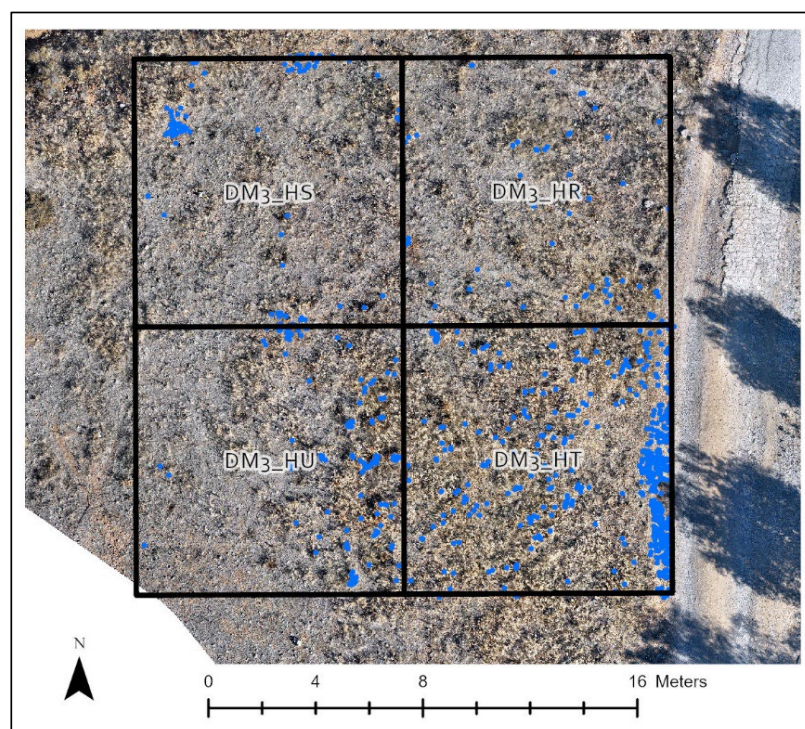


Figure A3. Potsherd detection (points—blue color) over the AOI from the Support Vector Machine classification.

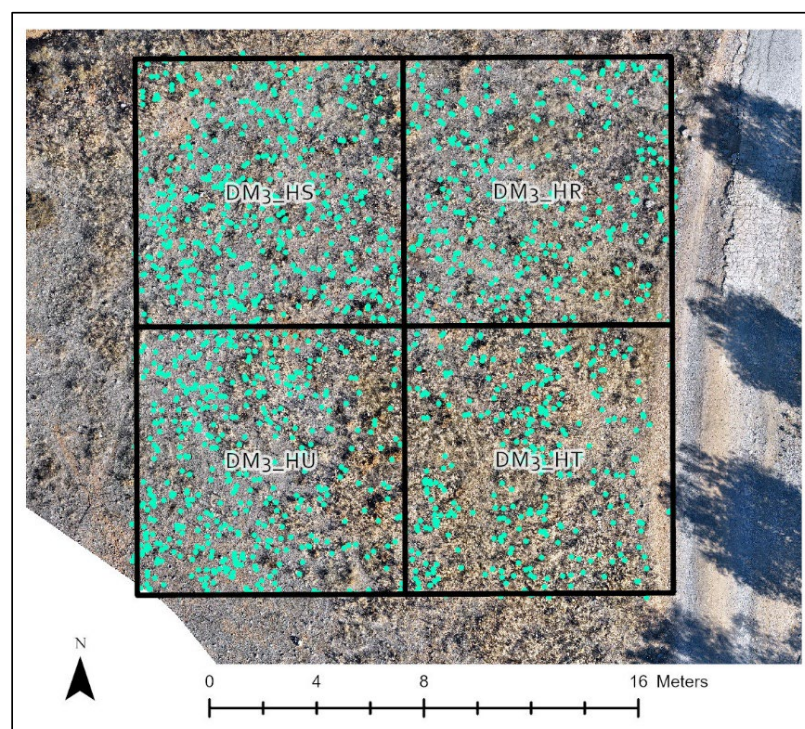


Figure A4. Potsherd detection (points—light green color) over the AOI from the Adaptive Boosting classification.

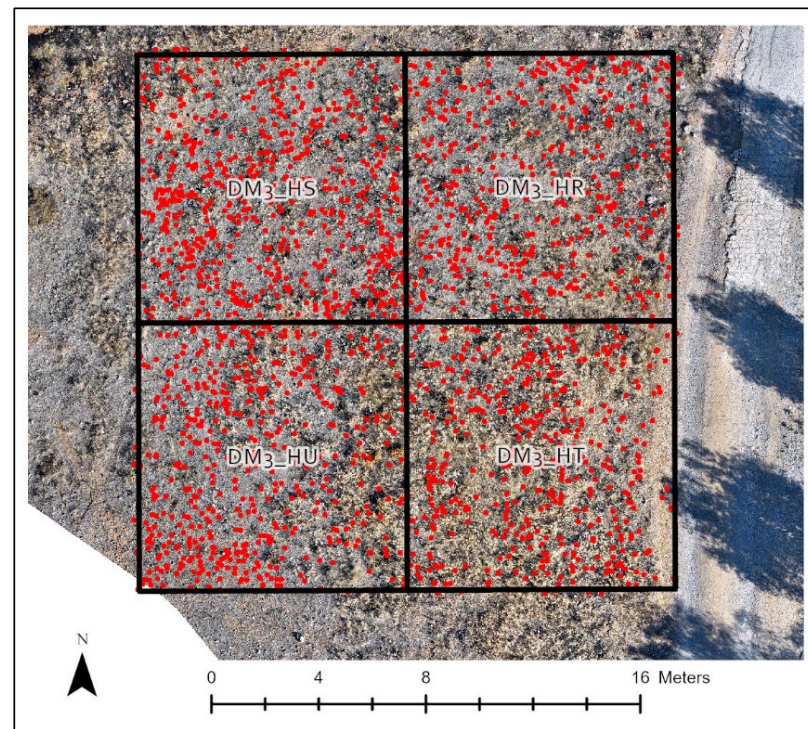


Figure A5. Potsherd detection (points—red color) over the AOI from the Extreme Gradient Boosting classification.

References

1. Attema, P.; Bintliff, J.; van Leusen, M.; Bes, P.; de Haas, T.; Donev, D.; Jongman, W.; Kaptijn, E.; Mayoral, V.; Menchelli, S.; et al. A guide to good practice in Mediterranean surface survey projects. *J. Greek Archaeol.* **2020**, *5*, 1–62. [\[CrossRef\]](#)
2. Argyrou, A.; Agapiou, A. A Review of Artificial Intelligence and Remote Sensing for Archaeological Research. *Remote Sens.* **2022**, *14*, 6000. [\[CrossRef\]](#)
3. Orengo, H.A.; Garcia-Molsosa, A. A brave new world for archaeological survey: Automated machine learning-based pot-sherd detection using high-resolution drone imagery. *J. Archaeol. Sci.* **2019**, *112*, 105013. [\[CrossRef\]](#)
4. Traviglia, A.; Torsello, A. Landscape pattern detection in archaeological remote sensing. *Geosciences* **2017**, *7*, 128. [\[CrossRef\]](#)
5. Tapete, D. Remote Sensing and Geosciences for Archaeology. *Geosciences* **2018**, *8*, 41. [\[CrossRef\]](#)
6. Lasaponara, R.; Masini, N. Satellite remote sensing in archaeology: Past, present and future perspectives. *J. Archaeol. Sci.* **2011**, *38*, 1995–2002. [\[CrossRef\]](#)
7. Materazzi, F.; Pacifici, M. Archaeological crop marks detection through drone multispectral remote sensing and vegetation indices: A new approach tested on the Italian pre-Roman city of Veii. *J. Archaeol. Sci. Rep.* **2022**, *41*, 103235. [\[CrossRef\]](#)
8. Trier, Ø.D.; Cowley, D.C.; Waldeland, A.U. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeol. Prospect.* **2019**, *26*, 165–175. [\[CrossRef\]](#)
9. Agapiou, A.; Vionis, A.; Papantoniou, G. Detection of Archaeological Surface Ceramics Using Deep Learning Image-Based Methods and Very High-Resolution UAV Imagery. *Land* **2021**, *10*, 1365. [\[CrossRef\]](#)
10. Altaweel, M.; Khelifi, A.; Li, Z.; Squitieri, A.; Basmaji, T.; Ghazal, M. Automated Archaeological Feature Detection Using Deep Learning on Optical UAV Imagery: Preliminary Results. *Remote Sens.* **2022**, *14*, 553. [\[CrossRef\]](#)
11. Bickler, S.H. Machine Learning Identification and Classification of Historic Ceramics. *Archaeol. N. Zeal.* **2018**, *61*, 20–32. Available online: <https://www.researchgate.net/publication/323302055> (accessed on 10 April 2025).
12. Bickler, S.H. Prospects for Machine Learning for Shell Midden Analysis. *Archaeol. N. Zeal.* **2018**, *61*, 48–58. Available online: <https://www.researchgate.net/publication/323468156> (accessed on 10 April 2025).
13. Verschoof-van der Vaart, W.B.; Lambers, K. Learning to Look at LiDAR: The Use of R-CNN in the Automated Detection of Archaeological Objects in LiDAR Data from the Netherlands. *J. Comput. Appl. Archaeol.* **2019**, *2*, 31–40. Available online: <https://www.researchgate.net/publication/331874666> (accessed on 10 April 2025). [\[CrossRef\]](#)
14. Reese, K.M. Deep learning artificial neural networks for non-destructive archaeological site dating. *J. Archaeol. Sci.* **2021**, *132*, 105413. [\[CrossRef\]](#)

15. Bonhage, A.; Eltaher, M.; Raab, T.; Breuß, M.; Raab, A.; Schneider, A. A modified Mask region-based convolutional neural network approach for the automated detection of archaeological sites on high-resolution light detection and ranging-derived digital elevation models in the North German Lowland. *Archaeol. Prospect.* **2021**, *28*, 177–186. [\[CrossRef\]](#)
16. Pawłowicz, L.M.; Downum, C.E. Applications of deep learning to decorated ceramic typology and classification: A case study using Tusayan White Ware from Northeast Arizona. *J. Archaeol. Sci.* **2021**, *130*, 105375. [\[CrossRef\]](#)
17. Davis, D.S. Defining what we study: The contribution of machine automation in archaeological research. *Digit. Appl. Archaeol. Cult. Herit.* **2020**, *18*, e00152. [\[CrossRef\]](#)
18. Olivier, M.; Verschoof-van der Vaart, W. Implementing State-of-the-Art Deep Learning Approaches for Archaeological Object Detection in Remotely-Sensed Data: The Results of Cross-Domain Collaboration. *J. Comput. Appl. Archaeol.* **2021**, *4*, 274–289. [\[CrossRef\]](#)
19. Richards-Rissetto, F.; Newton, D.; Al-Zadjali, A. A 3D point cloud Deep Learning approach using Lidar to identify ancient Maya archaeological sites. In Proceedings of the 28th CIPA Symposium “Great Learning & Digital Emotion”, Beijing, China, 28 August–1 September 2021.
20. Berganzo-Besga, I.; Orenge, H.A.; Lumbreras, F.; Carrero-Pazos, M.; Fonte, J.; Vilas-Estévez, B. Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel-2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia. *Remote Sens.* **2021**, *13*, 4181. [\[CrossRef\]](#)
21. Verschoof-van der Vaart, W.B.; Lambers, K.; Kowalczyk, W.; Bourgeois, Q.P. Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 293. [\[CrossRef\]](#)
22. Trier, V.D.; Reksten, J.H.; Løseth, K. Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *95*, 102241. [\[CrossRef\]](#)
23. Somrak, M.; Sašo Džeroski, S.; Kokalj, Z. Learning to Classify Structures in ALS-Derived Visualizations of Ancient Maya Settlements with CNN. *Remote Sens.* **2020**, *12*, 2215. [\[CrossRef\]](#)
24. Maxwell, A.E.; Pourmohammadi, P.; Poyner, J.D. Mapping the Topographic Features of Mining-Related Valley Fills Using Mask R-CNN Deep Learning and Digital Elevation Data. *Remote Sens.* **2020**, *12*, 547. [\[CrossRef\]](#)
25. Martin-Abadal, M.; Piñar-Molina, M.; Martorell-Torres, A.; Oliver-Codina, G.; Gonzalez-Cid, Y. Underwater Pipe and Valve 3D Recognition Using Deep Learning Segmentation. *J. Mar. Sci. Eng.* **2021**, *9*, 5. [\[CrossRef\]](#)
26. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *J. Appl. Remote Sens.* **2017**, *11*, 042609. [\[CrossRef\]](#)
27. Fu, T.; Ma, L.; Li, M.; Johnson, B.A. Using convolutional neural network to identify irregular segmentation objects from very high-resolution remote sensing imagery. *J. Appl. Remote Sens.* **2018**, *12*, 025010. [\[CrossRef\]](#)
28. Guyot, A.; Hubert-Moy, L.; Lorho, T. Detecting neolithic burial mounds from lidar derived elevation data using a multi-scale approach and machine learning techniques. *Remote Sens.* **2018**, *10*, 225. [\[CrossRef\]](#)
29. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#)
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [\[CrossRef\]](#)
31. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [\[CrossRef\]](#)
32. Domínguez Rodrigo, M.; Cifuentes Alcobendas, G.; Jiménez García, B.; Abellán, N.; Pizarro Monzo, M.; Organista, E.; Baquedano, E. Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications. *Sci. Rep.* **2020**, *10*, 18862. [\[CrossRef\]](#)
33. Rossi, M. 213. At the interface between GIS mobile applications and reflexive archaeology: Using QField app in archaeological surveys (CAA 2023). In *Book of Abstracts. CAA 2023: 50 Years of Synergy*; Brandsen, A., Hacigüzeller, P., van der Knaap, L., Taelman, D., Verhagen, P., Visser, R., Waagen, J., Eds.; Zenodo: Amsterdam, Switzerland, 2023; pp. 264–266. [\[CrossRef\]](#)
34. Argyrou, A.; Agapiou, A. Archaeological Surface Ceramics Detection Using Low Altitude Images Based on Weak Learners. In *Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Athens, Greece, 7–12 July 2024; IEEE: New York, NY, USA, 2024; pp. 3380–3383. [\[CrossRef\]](#)
35. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of Imbalanced Data: A Review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [\[CrossRef\]](#)
36. Valverde-Albacete, F.J.; Peláez-Moreno, C. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. *PLoS ONE* **2014**, *9*, e84217. [\[CrossRef\]](#)
37. Afonja, J. Accuracy Paradox. Towards Data Science. 10 December 2018. Available online: <https://medium.com/data-science/accuracy-paradox-897a69e2dd9b> (accessed on 17 February 2025).
38. Davis, K.; Maiden, R. The Importance of Understanding False Discoveries and the Accuracy Paradox When Evaluating Quantitative Studies. *Stud. Soc. Sci. Res.* **2021**, *2*, 1–8. [\[CrossRef\]](#)

39. Congalton, R.G.; Oderwald, R.G.; Mead, R.A. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogramm. Eng. Remote Sens.* **1983**, *49*, 1671–1678.
40. Fitzgerald, R.W.; Lees, B.G. Assessing the classification accuracy of multisource remote sensing data. *Remote Sens. Environ.* **1994**, *47*, 362–368. [\[CrossRef\]](#)
41. Foody, G.M. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sens. Environ.* **2020**, *239*, 111630. [\[CrossRef\]](#)
42. Joshi, M.V.; Kumar, V.; Agarwal, R.C. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. In Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01), San Jose, CA, USA, 29 November–2 December 2001.
43. Weiss, G. Mining with rarity: A unifying framework. *SIGKDD Explor. Newsl.* **2004**, *6*, 7–19. [\[CrossRef\]](#)
44. Cavalazzi, M.; Johnson, T.; Putzolu, C. *EAA 2024-Rome 28-31 August Surveying in Archaeology in the Digital Era: Rethinking Old Procedures and Imagining New Standards CALL FOR PAPERS*; Sapienza University: Rome, Italy, 2024.
45. Van Leusen, M.; Mayoral Herrera, V.; Farinetti, E.; de Haas, T.; Vandam, R. FIDO Guidelines for Fieldwalking Documentation 1.1. Available online: https://www.academia.edu/109603068/FIDO_Guidelines_for_Fieldwalking_Documentation_1_1 (accessed on 10 April 2025).
46. Argyrou, A.; Agapiou, A.; Papakonstantinou, A.; Alexakis, D.D. Comparison of Machine Learning Pixel-Based Classifiers for Detecting Archaeological Ceramics. *Drones* **2023**, *7*, 578. [\[CrossRef\]](#)
47. Argyrou, A.; Vionis, A.; Papakonstantinou, A.; Christoforou, M.; Alexakis, D.; Agapiou, A. Utilizing Low-Altitude Imagery and Weak Learner Algorithms for the Detection of Archaeological Surface Ceramics: The Kophinou case study. In Proceedings of the 52nd CAA International Conference, Athens, Greece, 5–9 May 2025.
48. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
49. John, G.H.; Langley, P. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–20 August 1995*; Morgan Kaufmann Publishers: San Francisco, CA, USA, 1995; pp. 338–345.
50. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
51. Schapire, R.E. The Strength of Weak Learnability. *Mach. Learn.* **1990**, *5*, 197–227. [\[CrossRef\]](#)
52. Freund, Y. Boosting a Weak Learning Algorithm by Majority. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT), University of Rochester, Rochester, NY, USA, 6–8 August 1990*; Fulk, M.A., Case, J., Eds.; Morgan Kaufmann Publishers Inc.: San Mateo, CA, USA, 1990; pp. 202–216.
53. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [\[CrossRef\]](#)
54. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [\[CrossRef\]](#)
55. Prati, R.C.; Batista, G.E.A.P.A. Class imbalances versus class overlapping: An analysis of a learning system behavior. In *Proceedings of the Mexican International Conference on Artificial Intelligence (MICAI), Mexico City, Mexico, 26–30 April 2004*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 312–321.
56. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010*; IEEE: New York, NY, USA, 2010.
57. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [\[CrossRef\]](#)
58. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [\[CrossRef\]](#)
59. Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research Progress on Few-Shot Learning for Remote Sensing Image Interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [\[CrossRef\]](#)
60. Zhou, Z.; Cui, Z.; Tang, K.; Tian, Y.; Pi, Y.; Cao, Z. ISPRS Journal of Photogrammetry and Remote Sensing Gaussian meta-feature balanced aggregation for few-shot synthetic aperture radar target detection. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 89–106. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.