Validating predictions of burial mounds with field data: the promise and reality of machine learning

Adela Sobotkova
Department of History and Classical Studies, Aarhus Universitet, Aarhus, Denmark
Ross Deans Kristensen-McLachlan and Orla Mallon
Center for Humanities Computing, Aarhus Universitet, Aarhus, Denmark, and
Shawn Adrian Ross
Department of History and Archaeology, Macquarie University, Sydney, Australia

Received 10 August 2022
Revised 1 November 2023
Accepted 3 February 2024

Abstract

Purpose – This paper provides practical advice for archaeologists and heritage specialists wishing to use ML approaches to identify archaeological features in high-resolution satellite imagery (or other remotely sensed data sources). We seek to balance the disproportionately optimistic literature related to the application of ML to archaeological prospection through a discussion of limitations, challenges and other difficulties. We further seek to raise awareness among researchers of the time, effort, expertise and resources necessary to implement ML successfully, so that they can make an informed choice between ML and manual inspection approaches.

Design/methodology/approach – Automated object detection has been the holy grail of archaeological remote sensing for the last two decades. Machine learning (ML) models have proven able to detect uniform features across a consistent background, but more variegated imagery remains a challenge. We set out to detect burial mounds in satellite imagery from a diverse landscape in Central Bulgaria using a pre-trained Convolutional Neural Network (CNN) plus additional but low-touch training to improve performance. Training was accomplished using MOUND/NOT MOUND cutouts, and the model assessed arbitrary tiles of the same size from the image. Results were assessed using field data.

Findings – Validation of results against field data showed that self-reported success rates were misleadingly high, and that the model was misidentifying most features. Setting an identification threshold at 60%

This research was funded by the Aarhus University Digital Literacy Initiative. We thank Cormac Purcell for initial elaboration of manually trained CNN, which led us to our current approach. Data were acquired by the participants of the Tundzha Regional Archaeological Project (2009–2011) with the support of the Australian Research Council Linkage Projects Funding scheme LP0989901, University of Michigan International Grant, GeoEye Foundation, Kazanlak Historical Museum, Sofia University of St. Kliment Ohridski, University of New South Wales; Macquarie University; American Research Center, Sofia; and the Institute for the Study of Aegean Prehistory. All of the machine-learning computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

Code: Data processing and analysis was performed using R and Python and the scripts are available in public repositories: (1) Training data preparation and CNN prediction validation can be found in https://github.com/adivea/cnn-testing (2) 2021 CNN classifier training and mound prediction is implemented in https://github.com/centre-for-humanities-computing/burial-mounds (3) 2022 CNN classifier training and mound prediction is implemented in https://github.com/centre-for-humanities-computing/MoundDetection

Journal of Documentation
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-05-2022-0096

probability, and noting that we used an approach where the CNN assessed tiles of a fixed size, tile-based false negative rates were 95–96%, false positive rates were 87–95% of tagged tiles, while true positives were only 5–13%. Counterintuitively, the model provided with training data selected for highly visible mounds (rather than all mounds) performed worse. Development of the model, meanwhile, required approximately 135 person-hours of work.

Research limitations/implications – Our attempt to deploy a pre-trained CNN demonstrates the limitations of this approach when it is used to detect varied features of different sizes within a heterogeneous landscape that contains confounding natural and modern features, such as roads, forests and field boundaries. The model has detected incidental features rather than the mounds themselves, making external validation with field data an essential part of CNN workflows. Correcting the model would require refining the training data as well as adopting different approaches to model choice and execution, raising the computational requirements beyond the level of most cultural heritage practitioners.

Practical implications – Improving the pre-trained model's performance would require considerable time and resources, on top of the time already invested. The degree of manual intervention required – particularly around the subsetting and annotation of training data – is so significant that it raises the question of whether it would be more efficient to identify all of the mounds manually, either through brute-force inspection by experts or by crowdsourcing the analysis to trained – or even untrained – volunteers. Researchers and heritage specialists seeking efficient methods for extracting features from remotely sensed data should weigh the costs and benefits of ML versus manual approaches carefully.

Social implications – Our literature review indicates that use of artificial intelligence (AI) and ML approaches to archaeological prospection have grown exponentially in the past decade, approaching adoption levels associated with "crossing the chasm" from innovators and early adopters to the majority of researchers. The literature itself, however, is overwhelmingly positive, reflecting some combination of publication bias and a rhetoric of unconditional success. This paper presents the failure of a good-faith attempt to utilise these approaches as a counterbalance and cautionary tale to potential adopters of the technology. Early-majority adopters may find ML difficult to implement effectively in real-life scenarios.

Originality/value – Unlike many high-profile reports from well-funded projects, our paper represents a serious but modestly resourced attempt to apply an ML approach to archaeological remote sensing, using techniques like transfer learning that are promoted as solutions to time and cost problems associated with, e.g. annotating and manipulating training data. While the majority of articles uncritically promote ML, or only discuss how challenges were overcome, our paper investigates how – despite reasonable self-reported scores – the model failed to locate the target features when compared to field data. We also present time, expertise and resourcing requirements, a rarity in ML-for-archaeology publications.

Keywords Machine learning, Remote sensing, CNN, Satellite imagery, Archaeology

Paper type Article

Introduction

Machine learning (ML) has taken the world by a storm, and archaeologists too have tried to harness the power of Convolutional Neural Networks (CNN) and other "deep learning" approaches to extract archaeological features from remotely sensed data. A series of recent articles have promoted the success of CNNs in cultural heritage applications (Can *et al.*, 2021; Caspari, 2020; Caspari and Crespo, 2019; Ekim *et al.*, 2021; Ma *et al.*, 2021). High success rates reported for detecting and monitoring burial mounds in Siberia have highlighted CNNs as an effective approach for large-scale prospection (Caspari and Crespo, 2019). Enthusiasm arising from this study, and similar outcomes from Egypt (Woolf, 2018) must, however, be tempered by the fact that the authors targeted uniform features situated in environments with little variation in terrain or vegetation – indeed, with relatively little vegetation or other confounding factors at all. Fewer studies explore the challenges presented by more difficult environments where cultural heritage lies in diverse or thick vegetation, surrounded by obtrusive natural and artificial features (Doyle *et al.*, 2023; Fuentes-Carbajal *et al.*, 2023; Verschoof-van der Vaart *et al.*, 2020).

Although few publications report the time, expertise, or costs associated with applying ML to archaeological prospection, examples from projects trying to extract symbols and text from historical maps indicate that it can be labour-intensive (Can *et al.*, 2021; Ekim *et al.*, 2021; Ma *et al.*, 2021). Correct classification of linear road features with a pre-trained model required

1,250 h to digitise and annotate training datasets (Can *et al.*, 2021, p. 62,847). Without such investment, one can expect to detect only the most consistent map features (Groom *et al.*, 2021), despite the fact that maps offer a simplified version of reality and use standardised symbols, which are far less complex than the earth's surface as reflected in satellite imagery.

Projects applying ML approaches to satellite imagery or LiDAR data likewise report a lack of high-quality, well-annotated training data, often due to cost and time constraints (Albrecht *et al.*, 2019; Argyrou and Agapiou, 2022; Canedo *et al.*, 2023; Casini *et al.*, 2021; Gallwey *et al.*, 2019; Kadhim and Abed, 2023; Karamitrou *et al.*, 2022, 2023; Lambers *et al.*, 2019; Ofli *et al.*, 2016; Sech *et al.*, 2023; Verschoof-van der Vaart and Landauer, 2021). Transfer learning based on pre-trained models is sometimes proposed as solution to the problem of limited training data, as well as related problems like small dataset size (Casini *et al.*, 2021, 2022; Character *et al.*, 2021; Gallwey *et al.*, 2019; Sech *et al.*, 2023; Xiong *et al.*, 2020). Use of a pre-trained CNN potentially obviates the need to have large, high-quality, and representative datasets for ML training, and promises to bring CNN approaches within the reach of smaller-scale projects, or projects that hope to detect or monitor varied archaeological phenomena in heterogeneous landscapes. Relatively few projects applying pre-trained models, however, offer a sustained discussion of failures, limitations, or shortcomings of the approach (for exceptions see Casini *et al.*, 2023; Sech *et al.*, 2023).

This paper offers a cautionary tale about the challenges, limitations, and demands of ML applied to archaeological prospection. We set out to detect burial mounds in the Kazanlak Valley, Bulgaria, using IKONOS high-resolution satellite imagery. We developed a pre-trained CNN that was further trained using two datasets: (1) image cutouts of areas where mounds were discovered during pedestrian fieldwork, regardless of how visible the mounds were in the satellite imagery, and (2) image cutouts where a burial mound is clearly visible in the imagery to a human observer. Each of these CNNs was then used to predict the locations of burial mounds in the imagery from the study area. We compare the predictions from the two models and their reported success rates measured against ground-truthed data. The results show that even a sophisticated, pre-trained model that is subjected to additional training struggles when confronted by inconsistent and sometimes indistinct features in a varied landscape. Indeed, both models failed to identify burial mounds in our study area. Expert researchers, or even novice volunteers, would have been more reliable (Sobotkova *et al.*, 2023). As such, we offer corrective insights regarding the difficulties we faced deploying a pre-trained CNN, as well as suggestions about how archaeologists might improve ML approaches they choose to adopt.

Burial mounds as heritage under threat

Burial mounds are a ubiquitous feature of the Bulgarian landscape (Oltean, 2013; Škorpil, 1925). Thousands of such mounds exist in the country; estimates range between 8,000 – 19,000 surviving today, of perhaps 50,000 originally constructed (Kitov, 1993, pp. 41–43; Shkorpil and Shkorpil, 1989, p. 20). Such mounds were built from the Early Bronze Age through the Middle Ages in the western extensions of the Asian steppes and surrounding areas. Over time, burial mound agglomerations formed entire mounded landscapes. Often they dot elevated ridges to enhance their prominence. Where terrain is flat, mounds align in linear patterns to highlight subtle morphology of the surrounding landscapes and signal their presence to the passer-by Sobotkova and Weissova (2019).

These rounded, conical piles of earth and stones vary in diameter from 10 m to 100 m and <1 m to >20 m in height (see Plates 1 and 2). Their contents vary from nothing (cenotaphs), to simple burials with or without an enclosure, to elaborate stone or brick tombs with notable architectural and artistic refinement and intrinsically valuable grave goods (Dimitrova, 2006; Tsetskhladze, 1998; Vasileva, 2005).

Plate 1.
Medium-sized burial mound in the Kaznlak Valley (TRAP inventory #2082)

Source(s): Shawn A Ross (CC-BY)

Plate 2.
Submeter mound in the Kazanlak Valley (TRAP inventory #3341)

Source(s): Shawn A Ross (CC-BY)

Despite the large number of burial mounds, they are endangered. Development in Bulgaria destroys dozens of mounds annually (Loulanski and Loulanski, 2017). Most regulated destructions result from formal rescue excavation in anticipation of housing or infrastructure construction, establishment of quarries, or similar development. In 2008, the last year for which data is available, burial mounds comprised nearly a quarter (57 of 257) of all excavations in Bulgaria (Cholakov and Chukalev, 2008, p. 91, Figure 2).

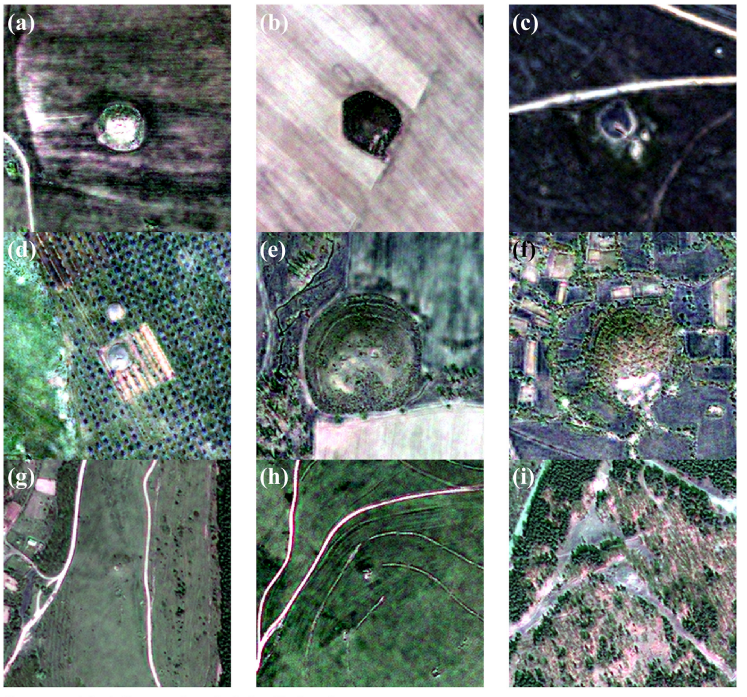
Despite administrative and legal measures to combat illicit artefact trafficking, such as a 2013 Memorandum of Understanding with the United States, looting probably damages more mounds than development (Eftimoski *et al.*, 2017; Lazar, 2015, pp. 107–124; Mateva, 2011). The Kazanlak Valley has been especially targeted by looters as it features an unusual concentration of conspicuous burial mounds (50–100 m in diameter) that include decorated ashlar tombs containing artwork, precious metals, high-quality ceramics, and other valuable objects (Dimitrova, 2006; Kitov, 1999). Although the majority of mounds in the valley are small in size and have modest contents, the potential payoff for looters, and failures of law and order, have fuelled looting since 1989 (Bailey, 1998).

Mounded landscapes also suffer slow and continuous wear from agricultural activities (Eftimoski *et al.*, 2017; Sobotkova and Weissova, 2020). Farmers plough and harrow arable fields annually, affecting thousands of mounds across Bulgaria. Unlike looting and (to a lesser extent) development, which capture public attention, damage from agriculture

generally goes unremarked. In the course of nearly 20 years of intermittent fieldwork in Bulgaria, the authors have seen few examples of mounds that had not been damaged either by development, looting, or agriculture, despite having inventoried over 2,000 of them across two Bulgarian provinces (Ross *et al.*, 2010, 2018; Sobotkova and Weissova, 2020).

Detecting archaeological features in satellite imagery

While burial mounds are readily identifiable on the ground due to their distinctive appearance, their visibility in satellite imagery depends on their size, surrounding terrain, and local land cover (see Figure 1). Large mounds in flat landscapes where a mound's surface contrasts with surrounding land cover (Figure 1a, b, e) are more visible than small mounds in hilly landscapes where their surfaces blend into the surroundings (Figure 1g-i). Human observers often look for crop, shadow, or soil marks to detect the less conspicuous mounds. Furthermore, mounds in forests can be obscured by the trees, while ploughed-over mounds blend into the surrounding fields. Visibility in satellite imagery is often marginal for small or even medium-sized mounds (Sobotkova and Ross, 2010). Conversely, image features such as lozenge-shaped clusters of trees in ploughed fields can be mistaken for mounds. When participants on the Tundzha Regional Archaeology Project (TRAP) previously attempted to detect mounds via manual inspection of satellite imagery, we encountered many false negatives, classification errors, and false positives. Failures, however, tended to be predictable, and some problems with manual inspection could be overcome by training and practice. Nevertheless, a field visit is usually required to confirm identification of a burial mound. After using satellite imagery for manual mound prospection since 2008, we were



Source(s): Figure by Adela Sobotkova

Figure 1.
150 by 150 m cutouts from the satellite imagery with mounds in the center used for ML training; (a–c) show medium-sized well-visible mounds, (d–f) large mounds covered in scrub and surrounded by varied vegetation, (g–i) show small mounds, poorly visible in satellite imagery

curious whether ML had, by 2022, reached the point where it could detect features as accurately and efficiently as a person.

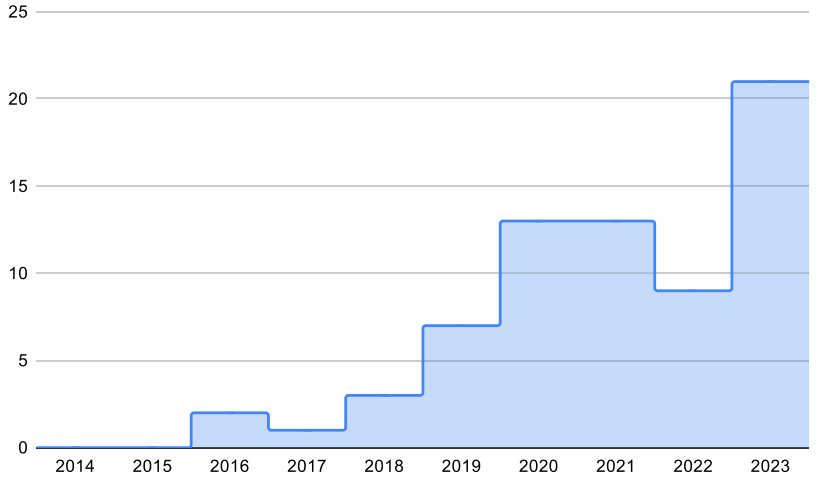
Automated approaches to remotely sensed data

Mapping cultural heritage at the scale of a country is a costly task whether it involves walking the landscape, digitising features from legacy sources like historical maps, or manually inspecting aerial or satellite images (Brophy and Cowley, 2005; Contreras, 2010; Crawford, 1929; De Laet *et al.*, 2007; De Laet and Lambers, 2009). Manual interpretation of remotely sensed imagery is not only tedious and time-consuming, but has limitations imposed by human bias and error (Bowen *et al.*, 2017). Today, however, more and more remotely sensed datasets are becoming available, requiring scalable and reproducible approaches to their analysis. Manual inspection is not without merit, and expert supervision is essential to – or may even outperform – automation (Bennett *et al.*, 2014; Casini *et al.*, 2023; Quintus *et al.*, 2017). Human observers, however, cannot manage the growing volume of remotely sensed data, or harness all the data contained in modern, global, high-resolution, multi- or hyperspectral imagery. Expertise in Geographic Information Systems (GIS), computational skills, and data management, as well as domain expertise, are all needed for manual digitisation, a shortage of which hampers crowdsourcing and citizen science projects (Duckers, 2013; Jessop, 2007). Projects that have applied manual inspection at scale ("GlobalXplorer", 2016; Lin *et al.*, 2014), struggled with: (1) the recruitment and retention of reliable participants (Wald *et al.*, 2016; West and Pateman, 2016), (2) the biases introduced by observers, and (3) the variability of naked-eye observations – problems that have only increased with the scale of the research (Cowley, 2012; Quintus *et al.*, 2017; Sadr, 2016).

In the last decade, a number of semi- and fully automated classification approaches have emerged to complement manual inspection of imagery. They range from template-matching (Kvamme, 2013; Trier *et al.*, 2009, 2015) to knowledge-based (D'Orazio *et al.*, 2012, 2015; Riley, 2009) and geographic object-based image analysis (Cerrillo-Cuenca, 2017; Davis *et al.*, 2019), to ML (Caspari *et al.*, 2014; Guyot *et al.*, 2018; Menze and Ur, 2012) and deep learning, a branch of ML which involves artificial neural networks such as CNNs (Berganzo-Besga *et al.*, 2023; Trier *et al.*, 2019; Verschoof-van der Vaart and Lambers, 2019). Ten years after Parcak's claim that "Tools cannot match the human eye" (2009, p. 110), artificial intelligence (AI) may be catching up. With a relatively small amount of high-quality, annotated data (circa 250 features), it is argued that archaeologists assisted by technologists may now extend their vision with digital eyes (Casana, 2020).

As a result, ML and other artificial intelligence approaches to remote sensing in archaeology are becoming ever more popular. Figure 2 displays the annual publication count in each of the past 10 years ($n = 70$; research articles and conference papers plus one preprint) in Web of Science for "topic" including "archaeology", and "remote sensing or satellite image", and "machine learning or artificial intelligence" (and variations thereof), manually reviewed to eliminate any publications that did not in fact discuss archaeology or ML/AI. This search reveals that the annual count of relevant publications has increased from zero in 2014 and 2015 to 21 in 2023. These 21 publications represent about 17% of the 2023 total ($n = 125$) for archaeological remote sensing (generated using the same query minus "machine learning or artificial intelligence"). Publication counts from Scopus using the same query are similar (cf. Argyrou and Agapiou, 2022, who analyse publications related to AI in archaeology more broadly and find a steadier rise in adoption beginning earlier).

If publication counts are used a proxy for research, this 17% figure indicates that AI/ML is on the cusp of "crossing the chasm" separating "innovators" and "early adopters" (together 16% of the population) from the "early majority", according to Rogers' diffusion of innovations paradigm as modified by Moore (Moore, 1991; Rogers, 2003). Indeed, a recent



Source(s): Figure by Shawn A Ross

Figure 2.
Annual publication count for AI or ML applications related to archaeological remote sensing

review article has argued that "[i]t is now outdated to focus on whether archeology should embrace automated detection . . . the aim should be to answer how to embrace AI tools to overcome the challenges of data overload and rapid destruction of the archaeological site landscapes" (Argyrou and Agapiou, 2022, p. 18). The appearance of ML code libraries, pre-trained models, and other commoditised ML tools, combined with the use of large-language models as research and programming assistants (e.g. ChatGPT or GitHub Copilot), makes it likely that the technology will continue to proliferate.

As these approaches spread to a broader cohort of researchers, potential adopters need to recognise their challenges and limitations. Critical assessment has, however, been somewhat neglected in the literature. Considering the 70 papers from the Web of Science mentioned above, 44 abstracts (63%) fail to mention any negative aspects of AI/ML approaches at all. Of the 26 papers (37%) with abstracts that mention some challenge or limitation, 11 state that they were overcome by the researchers, representing unqualified successes. Only 15 papers include specific or sustained critiques of ML approaches. Of those 15, seven (10% of the corpus) present qualified successes, while four (6%) discuss attempts to deploy ML that ended in partial or complete failures (Maxwell *et al.*, 2020; Roccetti *et al.*, 2020; Sech *et al.*, 2023; Verschoof-van der Vaart *et al.*, 2020). Two others (3%) present critical assessments of ML without the authors having undertaken it themselves (Casini *et al.*, 2021; Casana, 2020). Two (of four) review articles included critiques, one of which recognised specific challenges but generally reflected the positive tone of the literature (Argyrou and Agapiou, 2022), while the other offered a more sustained discussion of challenges and possible solutions (Kadhim and Abed, 2023). The overwhelmingly positive tone of these papers likely indicates a certain degree of "publication bias", where positive results are more likely to be published than negative (Brown *et al.*, 2017; Dickersin *et al.*, 1987; Harrison *et al.*, 2017; Ioannidis, 2005; Kühberger *et al.*, 2014; Møller and Jennions, 2001), or at the very least a reflection of the rhetorical shift in scientific research towards less qualified or uncertain presentation of outcomes (Vinkers *et al.*, 2015; Wheeler *et al.*, 2021; Yao *et al.*, 2023; Yuan and Yao, 2022). In this context, it is important to document unsuccessful attempts to apply ML techniques to archaeological remote sensing, or at least to highlight problems researchers are likely to face as they adopt the technology.

Data

Pedestrian survey

In this study we used a dataset of 773 mounds, collected by TRAP during 2009 – 2011 field survey in the Kazanlak Valley, Bulgaria (Sobotkova and Ross, 2018). This fieldwork covered some 85 sq km, inspected directly via pedestrian survey. TRAP survey identified many 0.5–20 m high (5–100 m diameter) conical features in the landscape as burial mounds (Figure 3). Each mound was recorded with a GPS point, height, diameter, and surface and surrounding land use, as well as preservation status using Likert scale and Wildesen (1982) classification (Sobotkova and Ross, 2018; Valchev and Sobotkova, 2019). While the identification of stony, submeter features (see Plate 2) as mounds was initially questioned, subsequent excavations demonstrated that these features indeed contained stone-lined graves with human remains, partly buried by colluvium or diminished by post-depositional processes (Nekhrizov *et al.*, 2013).

Satellite imagery

The satellite imagery used in this study consists of two IKONOS scenes covering 600 sq km delivered in geoTIFF format, which were acquired through a GeoEye Foundation grant in 2009. The scenes included a panchromatic band at 1 m resolution and a multispectral image (RGBNIR) at 4 m resolution. We fused the panchromatic and multispectral bands during 2009 fieldwork, to combine the high resolution of the panchromatic with the additional information of the multispectral bands in order to facilitate visual inspection of the images. Once fused, the two scenes were mosaiced and their edges cropped to form a rectangle. Although we had 600 sq km of imagery, all data for CNN model training and validation came from the 85 sq km TRAP study area within it.

Methods

Transfer learning

Rather than training our own model from scratch, we used a pre-trained CNN, a technique known as *transfer learning* (Casini *et al.*, 2022; Character *et al.*, 2021; Gallwey *et al.*, 2019; see also Kadhim and Abed, 2023; Pan and Yang, 2010; Sech *et al.*, 2023; for an overview see Weiss *et al.*, 2016; Xiong *et al.*, 2020). Transfer learning assumes that large, complex models can be pre-trained using data from one domain, then fine-tuned for a specific task in another domain. Transfer learning has become widely accepted as a standard approach to ML problems in areas like computer vision and natural language processing (Ruder *et al.*, 2019). In the case of transfer learning for image data, CNNs are often pre-trained on a subset of ImageNet (Deng *et al.*, 2009), learning combinations of weights and parameters which best extract information in order to make predictions between 1,000 different classes of image (Krizhevsky *et al.*, 2012). It is important to note that what is being transferred is not the ability to directly classify objects or make predictions about the contents of images. Instead, what pre-trained models appear to learn is the ability to perform a kind of *feature extraction*. The weights and parameters of the pre-trained CNN are used to extract dense, numerical vector representations of images – objects known as *image embeddings*. These image embeddings then act as the input to a classifier algorithm which learns to map between these abstract feature vectors and the provided labels (Akata *et al.*, 2015).

We adopted a transfer learning approach for this project for several reasons. First, pre-trained CNNs such as VGG16 (Simonyan and Zisserman, 2015), ResNet-50 (He *et al.*, 2016), and Inception-V3 (Szegedy *et al.*, 2016) have been shown to perform better (than a manually-trained model) on a wide range of downstream tasks such as image classification. Second, pre-trained CNNs have been trained on much larger and more diverse datasets than those

available to the average researcher, meaning that they should learn more sophisticated ways to generate image representations. Finally, the use of pre-trained models allows for faster prototyping and testing, an advantage when assessing its utility for particular field research.

After some preliminary experimentation with a range of different pre-trained models, we concluded that ResNet-50 seemed to perform best for our data. This model is one of the smaller pre-trained CNNs available, with only around 25.6m trainable parameters (for comparison, VGG16 has some 138.4m). The ResNet architecture has, however, been shown to learn high quality image embeddings more effectively due to the use of residual layers. The model was deployed using TensorFlow 2 in Python (see the GitHub repositories by Kristensen-McLachlan and Mallon, 2021, 2022).

During initial experiments, the model was found to overfit the training data, reporting close to 100% accuracy after only a few training epochs. Overfitting in such an extreme manner results in a model that does not generalise well to new data (Kadhim and Abed, 2023), meaning that such a classifier would not be useful. In order to counter this overfitting, we used proven data augmentation techniques to constrain model performance (Wang *et al.*, 2017). Data augmentation involves making random modifications and changes to the data during training, introducing noise and variability. The training data thus becomes more difficult to learn from, resulting in models that overfit less and producing a classifier that should generalise better to unseen data.

Additional CNN training

Training data preparation. Each ML training dataset requires two sets of images: positive data containing target features (MOUND) and negative data excluding the target features (NOT MOUND). Mound points taken during fieldwork were used as centroids for the generation of 150 × 150 m square polygons (150 × 150 pixels at 1 m resolution), which were clipped from the IKONOS imagery. This process yielded 773 MOUND cutouts, each centred on a mound (Figure 3). Cutouts were made from the four-band fused images. The NIR band was merged into the Red band for the classification. The mounds were always in the centre of the cutouts. No accommodation was made for the size of the mound; 100 m diameter mounds filled 34.9% of the cutout, while 10 m diameter mounds covered only 1.4%. Likewise, no accommodation was made for surrounding land cover, mound cover, or terrain. NOT MOUND data cutouts were generated in the same manner from areas excluding the 773 ground-truthed mound points, with no manual review. Both the satellite imagery and points were projected into the same EPSG: 32,635 coordinate reference system to be mutually compatible and to provide us with metric units of measurement. The ratio of positive to negative training data was approximately 1:2 (32%–68%).

In the 2021 run of the model, we used all 773 cutouts for training regardless of what was visible in the satellite image. In the 2022 run, we selected 249 cutouts where a mound was discernible with the naked eye. All of the training images were augmented using vertical and horizontal flip and random rotation. After processing, cutouts were divided into training, validation, and test sets following a 70:20:10 ratio for automated performance validation.

Training execution. We first guided the pre-trained CNN to learn to identify burial mounds from the positive (MOUND) and negative (NOT MOUND) training data. Once this additional training was completed (including validation and testing), the model was supplied with the entirety of the two IKONOS images (inclusive of training data), preprocessed into 150 × 150 pixel tiles. The model evaluated each tile and returned a MOUND probability prediction on a 0–1 scale, with 1 being 100%.

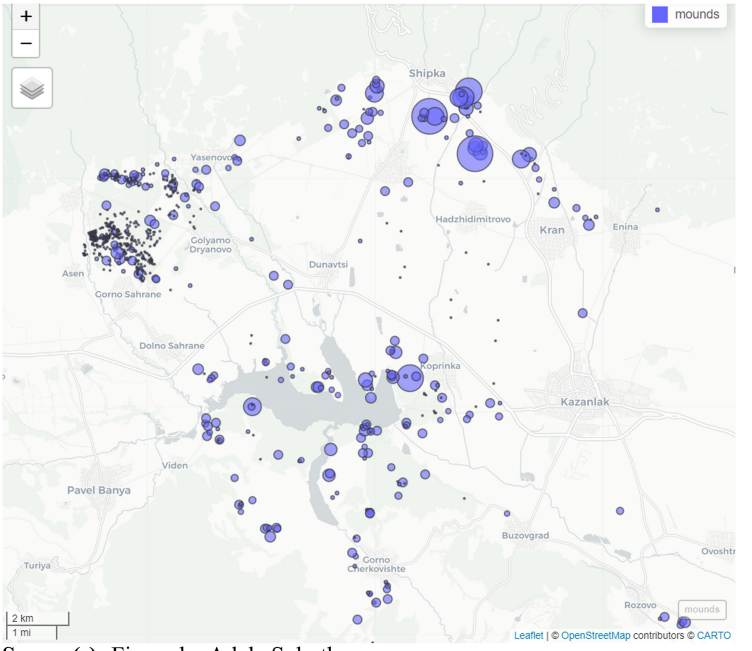


Figure 3.
Map of 773 mounds documented during pedestrian survey in the Kazanlak Valley; circle radius represents height from 0.5 to 20 m

Source(s): Figure by Adela Sobotkova

Assessment

Performance evaluation. The performance of the model was assessed using common evaluation criteria: We calculated F1 scores, precision, recall, and accuracy. These evaluation metrics are the model's own answers to the question "how well does the model predict MOUND/NOT MOUND in the test data, based on what it has learned from the training data?"

Model validation against field data. After the initial performance evaluation, we manually validated model performance. To do so, we verified mound predictions using points marking ground-truthed mounds in the 85 sq km section of the satellite image that had been systematically surveyed during TRAP fieldwalking.

The CNN's predictions were first associated with the tile name and a pair of origin coordinates marking the bottom left corner of the tile. We selected prediction records whose mound-probability exceeded 0.599 and used the coordinates to generate square polygons of 150 m side. We intersected these polygons with the points marking ground-truthed mounds. Mounds that fell inside the tiles were considered successfully detected (true positives). Mounds outside the flagged tiles were missed by the model (false negatives). Tiles flagged by the model that did not contain mounds constituted false positives.

Results

First run (2021): full training dataset

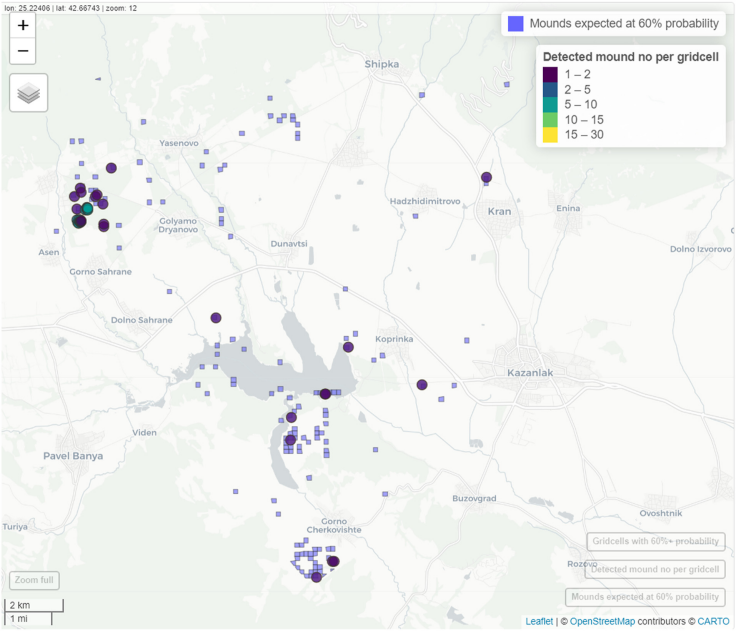
In the first run of the model, it was trained using all 773 known mounds, whether or not a mound was visible in the satellite image. After image augmentation, the model reported good learning and model fit (F1 = 0.87). This F1 score indicated that the use of a pre-trained model improved performance by 0.05 compared to a previous, manually trained model (pers.comm Cormac Purcell; Kristensen-McLachlan and Mallon, 2021).

Nevertheless, only 19 out of 148 tiles (12.8%) tagged by the model with at least a 60% chance of having a mound actually contained one (see Figure 4). Some 129 of the tagged tiles (87.1%) were false positives. The 19 true-positive tiles contained 38 mounds (1–9 mounds per tile), out of 773 in the study area (4.9%), while the remaining 735 mounds went undetected. Undetected mounds were located in 381 tiles (1–20 mounds per tile) out of 400 tiles that actually contained mounds, a false negative rate of 95.3% (Sobotkova, 2022).

During a visual examination of the model predictions, we saw that the model avoided the Koprinka reservoir in the middle of the valley. It correctly detected some mounds around the reservoir as well as a few in the northeastern necropolis in the Valley. In the north and northwest, however, it missed many, including both small mounds and large royal mounds that are clearly visible during manual inspection (see Figure 1e, f). It is also puzzling that in the densely mounded landscape of the northeastern necropolis, the algorithm managed to select an area devoid of mounds. To the south of the reservoir, the algorithm incorrectly selected a lot of forest, beaches, and roads, all of which increased the false positive rate. Overall, the model seemed to select bright lines and edges (forest, roads), rather than round shapes more likely to represent mounds. The greatest surprise was that the model failed to detect the largest mounds in the valley. These round, symmetrical features stand out against surrounding agricultural fields, and are crystal clear to any human viewer (Figure 5).

Second run (2022): training data filtered for visible mounds only

The second run was trained on 249 mounds that were clearly visible during manual inspection of the satellite imagery. The rest of the method remained the same.



Source(s): Figure by Adela Sobotkova

Figure 4.
Results of 2021 CNN run, blue squares represent tiles with 60%+ probability of containing mounds, circles mark true positives (actual mounds detected)

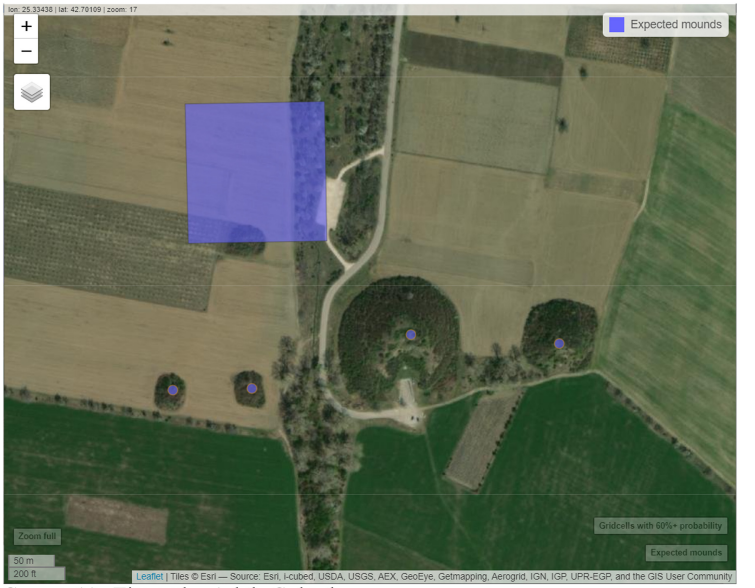


Figure 5.
Poor detection of large mounds at the royal necropolis of Shipka in the Kazanlak Valley; blue dots mark the actual mounds, blue square represents tile with 60%+ probability of containing a mound

Source(s): Figure by Adela Sobotkova

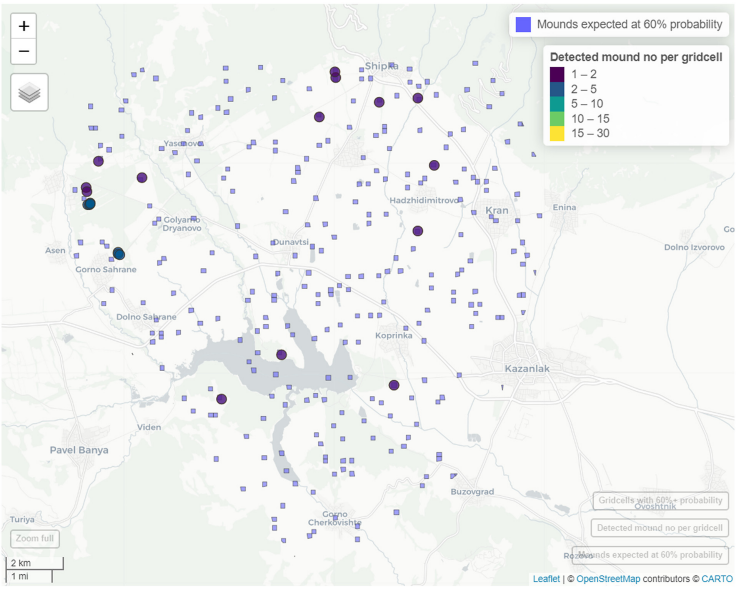
The second model's performance declined to an F1 score of 0.62 (Kristensen-McLachlan and Mallon, 2022). Validation revealed that only 21 of 773 mounds (2.7%) were detected, while 752 mounds (97.3%) remained undetected. The number of tiles within the TRAP study area flagged as containing a mound (at a >60% probability) increased from 148 in the first run to 288 here. Only 15 of these 288 tiles (5.2%), however, were true positives, containing the 21 detected mounds (1–4 mounds per tile; see Figure 6). The remaining 273 of 288 tiles were false positives (94.8%). The undetected 752 mounds lay in 384 tiles (1–28 mounds per tile) out of 399 tiles that actually contained mounds, a false negative rate of 96.2% (Sobotkova, 2022).

Small mounds were detected at a lower rate than in the 2021 model, demonstrated especially in the lack of predictions in the northwestern necropolis, which contained many mounds <1 m high. In addition, the large, archetypal mounds around the village of Shipka (north-centre of the study area, Figure 5) also remained undetected (as in the first run of the model). False positive tiles show that edges of forest, beach, and sections of roads are again classified as mounds. Furthermore, the model flagged parts of the reservoir as a mound with >60% probability, despite the homogeneous water surface, bringing the question "what is the CNN actually detecting?" to the fore. Overall, despite additional curation of the training data, the second run was even less successful than the first; the false positives and false negatives increased and fewer mounds were detected.

Discussion

Limitations and challenges of pre-trained CNNs

A number of factors contribute to the poor success rate of our CNN model when attempting to detect burial mounds in satellite imagery. The challenge centres around the variability of the appearance of the mounds themselves, combined with the heterogeneous landscape surrounding them and the noise of non-mound features.



Source(s): Figure by Adela Sobotkova

Figure 6.
Results of 2022 CNN run; blue squares show tiles with 60%+ probability of containing mounds, circles mark true positives (actual mounds detected)

In terms of CNN mechanics, when detecting mounds the model split the 600 sq km geoTIFF into non-overlapping 150 × 150 pixel tiles rather than analyse the mosaiced image via a moving window of variable size. As a result, mounds could sit on a tile boundary and elude detection. Many of the large undetected mounds, for example, fell on the edge of a tile. Second, an intersection of our CNN process and the nature of the training data produces low sensitivity to different mound sizes. The training data with the smallest mounds leaves a lot of non-target pixels in the tile (ca. 98%), while introducing other repeating and prominent features that confuse the classifier, such as the sharp edges of different land-use zones (e.g. stands of trees versus annual agriculture or pasture) and reflective linear features like roads and shorelines. Models have been observed to predict class membership based on incidental background features in the periphery of the tiles that happen to accompany the target phenomenon in the centre. Peter Haas pointed to this problem in "The Real Reason to be Afraid of Artificial Intelligence" when he explored how a deep learning model differentiated wolves from dogs in photographs. Wolves were recognised by the snow in the background rather than any intrinsic characteristic (TEDx, 2017). This effect arose from bias in the training data, where wolves were often on snowy background, while dogs appeared in a variety of other backgrounds. In our case, when the first model using the entire training dataset underperformed our expectations, we sought to address the problem by retraining the model using more prominent mounds that hopefully stood out better from background noise. This correction, however, failed to improve performance – indeed, it deteriorated further.

Overall, training the model with a set of highly variable features with even more varied and complex backgrounds may have misled it regarding the target of detection. This confusion likely arose from the variety of land cover on the surface of mounds and surrounding the mounds, which may or may not provide a contrast (or may provide different kinds of contrast).

Building a better model

Focus the positive training data. To counter the tendency of the model to detect extraneous background features, we could better focus the training data. Confronted with small and varied features in noisy environments, Verschoof-van der Vaart *et al.* (2020, pp. 7–8), for example, invested some effort adjusting model parameters to account for target feature size. In our case, to steer the CNN away from "reading the background", we would likely need to resize each tile according to the recorded diameter of the mound. We could, for example, specify a particular ratio (determined through experimentation) between mound and tile size, thereby controlling the target to background pixel ratio. That way we would end up with stamps of variable size containing minimal background noise. While this could be time-consuming for features of variable shapes, our dataset contains consistently circular mounds whose diameter was measured on-site. The only complication is that the measured diameter often varies from the diameter visible in the image due to indistinct mound boundaries.

Segment the training data. To address the problem of land-cover variability, we could create different training dataset for each class of surrounding land cover, each class of over-mound cover, or perhaps for various combinations of surrounding and over-mound cover. Segmentation of training data into runs of bare mounds in ploughed fields, scrub-covered mounds in ploughed fields, bare mounds in meadows, scrub-covered mounds in meadows, etc., would offer more consistency and might improve success rates. With all these possibilities, however, we reduce the amount of training data for each run. We also risk returning to the templating approaches of Kvamme (2013) and Trier *et al.* (2009), which would defeat the purpose of deep neural networks, which are supposed to learn their way towards recognising the target features more autonomously.

Emphasise negative training data. It may be counterintuitive, but care in creating the NOT MOUND tiles is as – if not more – important than the MOUND tiles. One lesson we learnt is that CNN cannot "ignore" without extensive training on negative examples (Gao *et al.*, 2019; Tang *et al.*, 2017). We can tell a person to ignore roads and forests, and she or he will instantly recognise and ignore them. For a CNN to action such a request we need to expose the model to them using negative training data. Especially when target features are small or occur in noisy backgrounds, care must be taken to provide a CNN with more negative examples (tiles without archaeological features) than positive. Our experience suggests that even a 1:2 (positive:negative) ratio was not sufficient. For comparison, Verschoof-van der Vaart *et al.* (2020) suggest a ratio of 1:1.6 in training data.

Employ alternative discovery strategies. During discovery, iterating the CNN over the image using overlapping tiles of different sizes would give each mound a higher likelihood of falling completely inside a tile. Using a moving window would also avoid missing mounds at tile edges (Tang *et al.*, 2017). Employing a Regional CNN, which can detect multiple, even overlapping features in an image could also improve results, especially if combined with iterating over variations of tiles (Maxwell *et al.*, 2020; Verschoof-van der Vaart *et al.*, 2020). Overlapping tiles or moving windows, however, would impose a substantial computational overhead as they would exponentially increase the imagery needing analysis. While upscaling the analysis is possible in a virtual environment, it would not be runnable on a typical personal computer, making such analysis infeasible for many cultural heritage specialists with modest resources.

Is it worth it?

Our outcomes compare poorly to the other applications of ML to archaeological prospection cited above (Caspari and Crespo, 2019; Ekim *et al.*, 2021). Based on our experience and the available literature, solutions exist that can address the problems we encountered. But is the extra effort worthwhile? Although time-on-task data regarding ML approaches in

archaeology is limited (Sobotkova *et al.*, 2023), such activities can absorb a lot of resources. Can *et al.* (2021), for example, report spending 1,250 h manually creating a training dataset of road features and a further seven days on testing and training a CNN model to detect roads in Austro-Hungarian imperial maps. While their CNN outperformed other models, the magnitude of manual data preparation, expertise required, and the computational infrastructure needed for training is within reach of only a few well-funded projects.

Developing our CNN model required approximately 135 person-hours from conceptualisation and experiments to validation and documentation. Our team included two digital archaeologists, a machine-learning specialist with experience applying ML approaches to cultural heritage data, and a junior developer who wrote much of the code used to implement these models. The approach used minimally processed datasets from previous fieldwork. Our approach to training was driven by expectations that the pre-trained CNN could tolerate a great deal of variation in training images, such as of mound size within a tile or the presence of confounding features, and that training data would not require extensive curation. We also brought with us from manual inspection of satellite imagery the preconception that burial mounds were relatively easy to recognise in imagery, at least compared to latent, "flat" sites. Our outcomes, however, indicate that even an advanced pre-trained model (ours is the same as that used by Can and Kabadayi, and perhaps also Crespo and Caspari) is likely to fail when asked to detect somewhat variable features against very heterogeneous backgrounds with only a low-touch approach to training. Improving the performance of our model is technically possible, mostly through more nuanced training approaches (see above). For our project, however, the return on additional time spent developing the ML approach diminishes rapidly, since we are already approaching a threshold where training student volunteers to identify mounds would be more efficient.

Nor is success guaranteed. To take one well-explained example, Verschoof-van der Vaart *et al.* (2020) implemented a sophisticated approach to LiDAR imagery using a region-based CNN. This approach is designed to classify multiple, adjacent, or overlapping features in a single image – potentially addressing some of the problems presented here. They also took care to provide adequate negative training data. Nevertheless, to reduce false positives, the project had to introduce a manual location-based ranking step, where experts identified, ranked, and mapped landscape characteristics (largely based on post-depositional processes) that might affect the survival or visibility of archaeological remains (Verschoof-van der Vaart *et al.*, 2020, pp. 9-10). Despite these manual interventions, and noting that performance evaluation involved realistic but artificial test data, the model still produced many false positives from natural or anthropogenic geometric shapes (not unlike our experience). In the end, Verschoof-van der Vaart *et al.* report that their models never reached the performance of crowdsourcing using volunteers (again, mostly due to its false positives), and the authors observe that the cost of ground truthing would be high.

Maxwell *et al.* (2020) achieve better results applying an R-CNN to LiDAR data for more prominent features (fill-valley faces produced by coal mining in the Appalachian Mountains), but the approach did not transfer to photogrammetric datasets. Even many successful ML applications require significant human intervention either in the form of crowdsourcing or specialist work (see, e.g. Casini *et al.*, 2023; Doyle *et al.*, 2023; Verschoof-van der Vaart and Lambers, 2019). As a result of challenges like these, Casana (2020), goes so far as to reject ML approaches entirely, instead favouring "brute force" manual inspection of satellite imagery by experts. Two of the authors on this paper themselves have had excellent large-scale (ca. 10,000 features) results from crowdsourcing (Sobotkova *et al.*, 2023). We reiterate the call in that paper for researchers to publish the challenges, costs, and resourcing requirements alongside results – including for serious but failed attempts – so that researchers, especially potential adopters of ML, can make informed decisions about what approach best fits their circumstances.

Conclusion

We set out to test the efficacy of a Machine Learning approach to detecting burial mounds situated in varied terrain and vegetation in the Kazanlak Valley of Bulgaria. We had at our disposal high-resolution, multispectral satellite imagery and a dataset of some 773 mounds registered during pedestrian survey. We employed a pre-trained CNN model with a low-touch approach to additional training, running the model twice using different collections of target features for training. We first used an indiscriminate collection of burial mounds, regardless of their visibility in the satellite imagery. We next used only those mounds that we deemed easily discernible in the satellite imagery. Unlike many projects, we had enough training data, but we did not devote a lot of time to prepare the training cutouts, e.g. by resizing them according to mound size or excluding background features. We believed that the volume of training data would offset other shortcomings. After this training, we applied the models to our imagery and validated results using our field data.

Although the first model reported a good fit (F1 = 0.87), validation with field data showed the model was confusing our target features with other phenomena. Tiles where the model predicted a mound with >60% probability contained a mound only 12.8% of the time – yielding a high rate of false positives (87.1%). At the same time, the model found only 38 of 773 known mounds (4.9%). Despite increased effort in the selection of training data, the second run of the model reported a lower F1 score (0.62) and at >60% probability produced even more false positives (94.8%) and false negatives (96.2%). This run found only 21 mounds (2.7%).

The high number of both false negatives and false positives demonstrated that our use of a low-touch approach – a pre-trained model plus sufficient but minimally curated training data – did not work. To overcome these challenges we would need to invest much more time, effort, and computing resources into additional steps such as: resize each training cutout based on the size of the mound to eliminate clipping and reduce noise from extraneous background features; segment the training data into different mound covers versus surrounding land covers; increase the ratio of negative training data; and utilise overlapping tiles, a moving window, and/or a Regional CNN.

A computationally manageable, low-touch model capable of discovering archaeological features in widely available satellite imagery with modest training data would be a valuable tool for cultural heritage managers who need to register such features and monitor them for change. Such managers often lack access to cloud computing infrastructure, or to technologists with the necessary geospatial analysis, data-wrangling, or machine-learning skills. It is not clear, however, that an ML approach remains cost effective if these additional steps are taken. Our work consumed about 135 person-hours and required significant domain and technical expertise, although it required only modest computing resources. The proposed modifications would certainly increase the time required, and potentially the expertise and computational resources. Meanwhile, our first intervention to improve the results added 20 h to the total time, but led to a poorer outcome. In our case, given the size of our study area and the number of features it likely contains, manual approaches like visual inspection by experts or crowdsourcing to volunteers would probably be more efficient.

Each project must make a decision regarding the trade-offs of different approaches to archaeological prospection using remotely sensed data, but the relevant literature underreports failures, challenges, and limitations of ML when used for this application. Very few projects, furthermore, report the time and resources invested in their approach, whether manual or automated. As ML approaches become more popular, researchers need more negative examples, discussion of problems, and resourcing information to make informed decisions about how to approach feature extraction from large remote-sensing datasets.

References

Akata, Z., Reed, S., Walter, D., Lee, H. and Schiele, B. (2015), "Evaluation of output embeddings for fine-grained image classification", *Presented at the Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2927-2936, doi: 10.1109/cvpr.2015.7298911.

Albrecht, C.M., Fisher, C., Freitag, M., Hamann, H.F., Pankanti, S., Pezzutti, F. and Rossi, F. (2019), "Learning and recognizing Archeological features from LiDAR data", in Baru, C., Huan, J., Khan, L., Hu, X.H., Ak, R., Tian, Y., Barga, R., *et al.* (Eds), pp. 5630-5636, doi: 10.1109/BigData47090.2019.9005548.

Argyrou, A. and Agapiou, A. (2022), "A review of artificial intelligence and remote sensing for archaeological research", *Remote Sensing*, Vol. 14 No. 23, p. 6000, doi: 10.3390/rs14236000.

Bailey, D.W. (1998), "Bulgarian archaeology; ideology, sociopolitics and the exotic", in Meskell, L. (Ed.), *Archaeology under Fire: Nationalism, Politics and Heritage in the Eastern Mediterranean and Middle East*, Routledge, London, New York, pp. 87-110.

Bennett, R., Cowley, D. and De Laet, V. (2014), "The data explosion: tackling the taboo of automatic feature recognition in airborne survey data", *Antiquity*, Vol. 88 No. 341, pp. 896-905, doi: 10.1017/S0003598X00050766.

Berganzo-Besga, I., Orengo, H.A., Lumbreras, F., Alam, A., Campbell, R., Gerrits, P.J., de Souza, J.G., Khan, A., Suárez-Moreno, M., Tomaney, J., Roberts, R.C. and Petrie, C.A. (2023), "Curriculum learning-based strategy for low-density archaeological mound detection from historical maps in India and Pakistan", *Scientific Reports*, Vol. 13 No. 1, p. 11257, doi: 10.1038/s41598-023-38190-x.

Bowen, E.F.W., Tofel, B.B., Parcak, S. and Granger, R. (2017), "Algorithmic identification of looted archaeological sites from space", *Frontiers in ICT*, Vol. 4, doi: 10.3389/fict.2017.00004.

Brophy, K. and Cowley, D. (2005), in Brophy, K. and Cowley, D. (Eds), *From the Air: Understanding Aerial Archaeology*, Tempus, Stroud, p. 190, available at: https://eprints.gla.ac.uk/64330/

Brown, A.W., Mehta, T.S. and Allison, D.B. (2017), "Publication bias in science", in Jamieson, K.H., Kahan, D.M. and Scheufele, D.A. (Eds), *Oxford Handbooks Online*, Oxford University Press, Vol. 1, doi: 10.1093/oxfordhb/9780190497620.013.10.

Can, Y.S., Petrus, J.G. and Kabadayi, M.E. (2021), "Automatic detection of road types from the third military mapping survey of Austria-Hungary historical map series with deep convolutional neural networks", *IEEE Access*, Vol. 9, pp. 62847-56, doi: 10.1109/ACCESS.2021.3074897.

Canedo, D., Fonte, J., Seco, L.G., Vazquez, M., Dias, R., Do Pereiro, T., Hipolito, J., Menéndez-Marsh, F., Georgieva, P. and Neves, A.J.R. (2023), "Uncovering archaeological sites in airborne LiDAR data with data-centric artificial intelligence", *IEEE Access*, Vol. 11, pp. 65608-65619, doi: 10.1109/ACCESS.2023.3290305.

Casana, J. (2020), "Global-scale archaeological prospection using CORONA satellite imagery: automated, crowd-sourced, and expert-led approaches", *Journal of Field Archaeology*, Vol. 45 Supp 1, pp. S89-S100, doi: 10.1080/00934690.2020.1713285.

Casini, L., Marchetti, N., Montanucci, A., Orrù, V. and Roccetti, M. (2023), "A human–AI collaboration workflow for archaeological sites detection", *Scientific Reports*, Vol. 13 No. 1, p. 8699, doi: 10.1038/s41598-023-36015-5.

Casini, L., Orru, V., Roccetti, M. and Marchetti, N. (2022), "When machines find sites for the archaeologists: a preliminary study with semantic segmentation applied on satellite imagery of the mesopotamian floodplain", *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, doi: 10.1145/3524458.3547121.

Casini, L., Roccetti, M., Delnevo, G., Marchetti, N. and Orru, V. (2021), "The Barrier of meaning in archaeological data science", *SciFi-It'2020: Designing Your Future With Science Fiction*, Ostend, Ghent, Belgium, pp. 61-65, doi: 10.48550/arXiv.2102.06022.

Caspari, G. (2020), "Mapping and damage assessment of 'royal' burial mounds in the Siberian Valley of the kings", *Remote Sensing*, Vol. 12 No. 5, p. 773, doi: 10.3390/rs12050773.

Caspari, G. and Crespo, P. (2019), "Convolutional neural networks for archaeological site detection–Finding 'princely' tombs", *Journal of Archaeological Science*, Vol. 110, p. 104998, doi: 10.1016/j.jas.2019.104998.

Caspari, G., Balz, T., Gang, L., Wang, X. and Liao, M. (2014), "Application of Hough forests for the detection of grave mounds in high-resolution satellite imagery", *2014 IEEE Geoscience and Remote Sensing Symposium*, pp. 906-909, doi: 10.1109/IGARSS.2014.6946572.

Cerrillo-Cuenca, E. (2017), "An approach to the automatic surveying of prehistoric barrows through LiDAR", *Quaternary International: The Journal of the International Union for Quaternary Research*, Vol. 435, pp. 135-145, doi: 10.1016/j.quaint.2015.12.099.

Character, L., Ortiz, A. Jr, Beach, T. and Luzzadder-Beach, S. (2021), "Archaeologic machine learning for shipwreck detection using lidar and sonar", *Remote Sensing*, Vol. 13 No. 9, p. 1759, doi: 10.3390/rs13091759.

Cholakov, I.D. and Chukalev, K. (2008), "Statistical data on the archaeological field activities in Bulgaria, season 2007", *Archaeologica Bulgarica*, Vol. XII No. 3, pp. 89-100.

Contreras, D.A. (2010), "Huaqueros and remote sensing imagery: assessing looting damage in the viru valley, Peru", *Antiquity*, Vol. 84 No. 324, pp. 544-555, doi: 10.1017/s0003598x0006676x.

Cowley, D.C. (2012), "In with the new, out with the old? Auto-extraction for remote sensing archaeology", *Presented at the Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2012*, International Society for Optics and Photonics, Vol. 8532, 853206, doi: 10.1117/12.981758.

Crawford, O.G.S. (1929), *Air Photography for Archaeologists*, H.M. Stationery Office, London.

Davis, D.S., Sanger, M.C. and Lipo, C.P. (2019), "Automated mound detection using lidar and object-based image analysis in Beaufort County, South Carolina", *Southeastern Archaeology*, Vol. 38 No. 1, pp. 23-37, doi: 10.1080/0734578X.2018.1482186.

De Laet, V. and Lambers, K. (2009), "Archaeological prospecting using high-resolution digital satellite imagery : recent advances and future prospects; a session held at the computer applications and quantitative methods in archaeology (CAA) conference, Williamsburg, USA, March 2009", *AARGnews - The Newsletter of the Aerial Archaeology Research Group*, Vol. 39, pp. 9-17.

De Laet, V., Paulissen, E. and Waelkens, M. (2007), "Methods for the extraction of archaeological features from very high-resoltuion Ikonos-2 remote sensing imagery, Hisar (Southwest Turkey)", *Journal of Archaeological Science*, Vol. 34 No. 5, pp. 830-841, doi: 10.1016/j.jas.2006.09.013.

Deng, J., Dong, W., Socher, R., Li, L.J. and Li, K. (2009), "Imagenet: a large-scale hierarchical image database", *Conference on Computer Vision and Pattern Recognition., presented at the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, IEEE.

Dickersin, K., Chan, S., Chalmersx, T.C., Sacks, H.S. and Smith, H. (1987), "Publication bias and clinical trials", *Controlled Clinical Trials*, Vol. 8 No. 4, pp. 343-353, doi: 10.1016/0197-2456(87)90155-3.

Dimitrova, D. (2006), "Ruler's tumular burials from the kazanluk region in Bulgaria", *Pratiques Funeraires et Manifestations de L'identite*, academia.edu.

Doyle, C., Luzzadder-Beach, S. and Beach, T. (2023), "Advances in remote sensing of the early Anthropocene in tropical wetlands: from biplanes to lidar and machine learning", *Progress in Physical Geography-Earth and Environment*, Vol. 47 No. 2, pp. 293-312, doi: 10.1177/03091333221134185.

Duckers, G.L. (2013), "Bridging the 'geospatial divide' in archaeology: community based interpretation of LIDAR data", *Internet Archaeology*, Vol. 35, doi: 10.11141/ia.35.10.

D'Orazio, T., Palumbo, F. and Guaragnella, C. (2012), "Archaeological trace extraction by a local directional active contour approach", *Pattern Recognition*, Vol. 45 No. 9, pp. 3427-3438, doi: 10.1016/j.patcog.2012.03.003.

D'Orazio, T., Da Pelo, P., Marani, R. and Guaragnella, C. (2015), "Automated extraction of archaeological traces by a modified variance analysis", *Remote Sensing*, Vol. 7 No. 4, pp. 3565-3587, doi: 10.3390/rs70403565.

Eftimoski, M., Ross, S.A. and Sobotkova, A. (2017), "The impact of land use and depopulation on burial mounds in the Kazanlak Valley, Bulgaria: an ordered logit predictive model", *Journal of Cultural Heritage*, Vol. 23, pp. 1-10, doi: 10.1016/j.culher.2016.10.002.

Ekim, B., Sertel, E. and Kabadayı, M.E. (2021), "Automatic road extraction from historical maps using deep learning techniques: a regional case study of Turkey in a German world war II map", *ISPRS International Journal of Geo-Information*, Vol. 10 No. 8, p. 492, doi: 10.3390/ijgi10080492.

Fuentes-Carbajal, J.A., Carrasco-Ochoa, J.A., Martinez-Trinidad, J.F. and Flores-Lopez, J.A. (2023), "Machine learning and image-processing-based method for the detection of archaeological structures in areas with large amounts of vegetation using satellite images", *Applied Sciences-Basel*, Vol. 13 No. 11, p. 6663, doi: 10.3390/app13116663.

Gallwey, J., Eyre, M., Tonkins, M. and Coggan, J. (2019), "Bringing lunar LiDAR back down to earth: mapping our industrial heritage through deep transfer learning", *Remote Sensing*, Vol. 11 No. 17, p. 1994, doi: 10.3390/rs11171994.

Gao, L., He, Y., Sun, X., Jia, X. and Zhang, B. (2019), "Incorporating negative sample training for ship detection based on deep learning", *Sensors*, Vol. 19 No. 3, p. 684, doi: 10.3390/s19030684.

GlobalXplorer (2016), available at: https://www.globalxplorer.org/ (accessed 26 April 2022).

Groom, G., Levin, G., Svenningsen, S.R. and Perner, M.L. (2021), "Dune Sand–Object based image analysis for vectorization of a dotted signature in Danish late 1800s maps", *E-Perimetron*, Vol. 16 No. 4, pp. 156-165.

Guyot, A., Hubert-Moy, L. and Lorho, T. (2018), "Detecting neolithic burial mounds from LiDAR-derived elevation data using a multi-scale approach and machine learning techniques", *Remote Sensing*, Multidisciplinary Digital Publishing Institute, Vol. 10 No. 2, p. 225, doi: 10.3390/rs10020225.

Harrison, J.S., Banks, G.C., Pollack, J.M., O'Boyle, E.H. and Short, J. (2017), "Publication bias in strategic management research", *Journal of Management*, Vol. 43 No. 2, pp. 400-425, doi: 10.1177/0149206314535438.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770-778, doi: 10.1109/cvpr.2016.90.

Ioannidis, J.P.A. (2005), "Why most published research findings are false", *PLoS Medicine*, Vol. 2 No. 8, p. e124, doi: 10.1371/journal.pmed.0020124.

Jessop, M. (2007), "The inhibition of geographical information in digital humanities scholarship", *Literary and Linguistic Computing*, Vol. 23 No. 1, pp. 39-50, doi: 10.1093/llc/fqm041.

Kadhim, I. and Abed, F.M.M. (2023), "A critical review of remote sensing approaches and deep learning techniques in archaeology", *Sensors*, Vol. 23 No. 6, p. 2918, doi: 10.3390/s23062918.

Karamitrou, A., Sturt, F. and Bogiatzis, P. (2023), "Identification of black reef shipwreck sites using AI and satellite multispectral imagery", *Remote Sensing*, Vol. 15 No. 8, p. 2030, doi: 10.3390/rs15082030.

Karamitrou, A., Sturt, F., Bogiatzis, P. and Beresford-Jones, D. (2022), "Towards the use of artificial intelligence deep learning networks for detection of archaeological sites", *Surface Topography-Metrology and Properties*, Vol. 10 No. 4, p. 044001, doi: 10.1088/2051-672X/ac9492.

Kitov, G. (1993), "Trakiyskite mogili", *Thracia*, Vol. 10, pp. 39-80.

Kitov, G. (1999), "Royal insignia, tombs and temples in the valley of the thracian rules", *Archaeologia Bulgarica*, Vol. 3, pp. 1-20.

Kristensen-McLachlan, R. and Mallon, O. (2021), "Mound detection", *Github*, available at: https://github.com/centre-for-humanities-computing/MoundDetection

Kristensen-McLachlan, R. and Mallon, O. (2022), "Burial mounds", *Github*, available at: https://github.com/centre-for-humanities-computing/burial-mounds

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012), "ImageNet classification with deep convolutional neural networks", in Pereira, F., Burges, C.J., Bottou, L. and Weinberger, K.Q. (Eds), *Advances in Neural Information Processing Systems*, Curran Associates, Vol. 25.

Kühberger, A., Fritz, A. and Scherndl, T. (2014), "Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size", in Fanelli, D. (Ed.), *PloS One*, Vol. 9 No. 9, p. e105825, doi: 10.1371/journal.pone.0105825.

Kvamme, K.L. (2013), "An examination of automated archaeological feature recognition in remotely sensed imagery", in Bevan, A. and Lake, M. (Eds), *Computational Approaches to Archaeological Spaces*, Left Coast Press, pp. 53-68.

Lambers, K., Verschoof-van der Vaart, W.B. and Bourgeois, Q.P.J. (2019), "Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection", *Remote Sensing*, Vol. 11 No. 7, p. 794, doi: 10.3390/rs11070794.

Lazar, A. (2015), "Illicit trafficking in cultural goods in south east Europe: Fiat Lux!", in Desmarais, F. (Ed.), *Countering Illicit Traffic in Cultural Goods the Global Challenge of Protecting the World's Heritage, ICOM International Observatory on Illicit Traffic in Cultural Goods*, ICOM, pp. 107-124.

Lin, A.Y.M., Huynh, A., Lanckriet, G. and Barrington, L. (2014), "Crowdsourcing the unknown: the satellite search for Genghis Khan", *PloS One*, Vol. 9 No. 12, p. e114046, doi: 10.1371/journal.pone.0114046.

Loulanski, T. and Loulanski, V. (2017), "Thracian mounds in Bulgaria: heritage at risk", *The Historic Environment: Policy and Practice*, Vol. 8 No. 3, pp. 246-277, doi: 10.1080/17567505.2017.1359918.

Ma, J., Sefer, A. and Kabadayı, M.E. (2021), "Geolocating Ottoman settlements: the use of historical maps for digital humanities", *Proceedings of the ICA*, Vol. 3, pp. 1-8, doi: 10.5194/ica-proc-3-10-2021.

Mateva, B. (2011), "Problemi na ohranata i ispolzvaneto na pametnitsite na kulturata za turisticheski celi", *Bulgarian E-Journal of Archaeology*, Vol. 1, pp. 123-126.

Maxwell, A.E., Pourmohammadi, P. and Poyner, J.D. (2020), "Mapping the topographic features of mining-related valley fills using mask R-CNN deep learning and digital elevation data", *Remote Sensing*, Vol. 12 No. 3, p. 547, doi: 10.3390/rs12030547.

Menze, B.H. and Ur, J.A. (2012), "Mapping patterns of long-term settlement in Northern Mesopotamia at a large scale", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109 No. 14, pp. E778-E787, doi: 10.1073/pnas.1115472109.

Møller, A.P. and Jennions, M.D. (2001), "Testing and adjusting for publication bias", *Trends in Ecology and Evolution*, Vol. 16 No. 10, pp. 580-586, doi: 10.1016/S0169-5347(01)02235-2.

Moore, G.A. (1991), *Crossing the Chasm: Marketing and Selling High-Tech Goods to Mainstream Customers*, HarperBusiness, New York.

Nekhrizov, G., Parvin, M. and Kecheva, N. (2013), "Prouchvane na nadgrobni mogili ot nekropola pri selata Yasenovo i Golyamo Dryanovo, obsht. Kazanlak", *Problemi I Izsledvaniya Na Trakiiskata Kultura*, Vol. VI, pp. 17-41.

Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., Briant, J., Millet, P., Reinhard, F., Parkan, M. and Joost, S. (2016), "Combining human computing and machine learning to make sense of big (aerial) data for disaster response", *Big Data*, Vol. 4 No. 1, pp. 47-59, doi: 10.1089/big.2014.0064.

Oltean, I. (2013), "Burial mounds and settlement patterns: a quantitative approach to their identification from the air and interpretation", *Antiquity*, Vol. 87 No. 335, pp. 202-219, doi: 10.1017/s0003598x00048729.

Pan, S.J. and Yang, Q. (2010), "A survey on transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22 No. 10, pp. 1345-1359, doi: 10.1109/TKDE.2009.191.

Parcak, S.H. (2009), *Satellite Remote Sensing for Archaeology*, Routledge, London, New York.

Quintus, S., Day, S.S. and Smith, N.J. (2017), "The efficacy and analytical importance of manual feature extraction using lidar datasets", *Advances in Archaeological Practice*, Vol. 5 No. 4, pp. 351-364, doi: 10.1017/aap.2017.13.

Riley, A.M. (2009), "Automated detection of prehistoric conical burial mounds from LiDAR bare-earth digital elevation models: a thesis presented to the department of geology and geography in candidacy for the degree of master of science", Northwest Missouri State University.

Roccetti, M., Casini, L., Delnevo, G., Orrù, V. and Marchetti, N. (2020), "Potential and limitations of designing a deep learning model for discovering new archaeological sites: a case with the mesopotamian floodplain", *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, New York, NY, USA, Association for Computing Machinery, pp. 216-221, doi: 10.1145/3411170.3411254.

Rogers, E.M. (2003), *Diffusion of Innovations*, 5th ed., Free Press, New York.

Ross, S.A., Sobotkova, A., Connor, S. and Iliev, I. (2010), "An interdisciplinary pilot project in the environs of Kabyle, Bulgaria", *Archaeologica Bulgarica*, Vol. 14 No. 2, pp. 69-85.

Ross, S.A., Sobotkova, A., Tzvetkova, J., Nekhrizov, G. and Connor, S. (2018), *The Tundzha Regional Archaeology Project: Surface Survey, Palaeoecology, and Associated Studies in Central and Southeast Bulgaria, 2009-2015 Final Report*, Oxbow Books, Oxford.

Ruder, S., Peters, M.E., Swayamdipta, S. and Wolf, T. (2019), "Transfer learning in natural language processing", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, Minneapolis, Minnesota, Association for Computational Linguistics, pp. 15-18, doi: 10.18653/v1/N19-5004.

Sadr, K. (2016), "The impact of coder reliability on reconstructing archaeological settlement patterns from satellite imagery: a case study from South Africa", *Archaeological Prospection*, Vol. 23 No. 1, pp. 45-54, doi: 10.1002/arp.1515.

Sech, G., Soleni, P., Verschoof-van der Vaart, W.B., Kokalj, Ž., Traviglia, A. and Fiorucci, M. (2023), "Transfer learning of semantic segmentation methods for identifying buried archaeological structures on LiDAR data [arXiv]", *arXiv*. doi: 10.48550/arXiv.2307.03512.

Shkorpil, H. and Shkorpil, K. (1989), *Mogili*, Pchela press, Plovdiv.

Simonyan, K. and Zisserman, A. (2015), "Very deep convolutional networks for large-scale image recognition", CoRR abs/1409.1556.

Škorpil, K. (1925), *Megalitni Pametnitsi I Mogilishta (Starini V Chernomorskata Oblast - Chast 1)*, Drzhavna Pechatnitsa, Sofia.

Sobotkova, A. (2022), "CNN validation scripts (version 9852d5b)", *Github*, available at: https://github.com/adivea/cnn-testing/releases/tag/v1.0.0

Sobotkova, A. and Ross, S.A. (2010), "High-resolution, multi-spectral satellite imagery and extensive archaeological prospection: case studies from Apulia, Italy, and Kazanluk, Bulgaria", in Forte, M., Campana, S. and Liuzza, C. (Eds), *Space, Time, Place. Third International Conference on Remote Sensing in Archaeology*, Tiruchirrappalli, India, Vol. S2118, Archaeopress, pp. 25-28.

Sobotkova, A. and Ross, S.A. (2018), "Kazanlak survey results", in Ross, S.A., Sobotkova, A., Tzvetkova, J., Georgi, N. and Simon, C. (Eds), *The Tundzha Regional Archaeological Project: Surface Survey, Palaeoecology, and Associated Studies in Central and Southeast Bulgaria, 2009-2015 Final Report*, Oxbow Books, Oxford, pp. 66-81.

Sobotkova, A. and Weissova, B. (2019), "Locational analysis of burial mounds in the middle Tundzha river watershed. Combining historical maps with field survey and satellite image analysis data", *Vesti Na Yambolskiya Musei*, Vol. 6 No. 9, pp. 161-175.

Sobotkova, A. and Weissova, B. (2020), "Soviet topographic maps and burial mounds of the Yambol province: digital workflow for mortuary landscape verification", *Archaeological Prospection*, Vol. 33 No. 3, pp. 233-262, doi: 10.1002/arp.1769.

Sobotkova, A., Ross, S.A., Nassif-Haynes, C. and Ballsun-Stanton, B. (2023), "Creating large, high-quality geospatial datasets from historical maps using novice volunteers", *Applied Geography*, Vol. 155, p. 102967, doi: 10.1016/j.apgeog.2023.102967.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016), "Rethinking the inception architecture for computer vision", *Presented at the 2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

Tang, T., Zhou, S., Deng, Z., Zou, H. and Lei, L. (2017), "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining", *Sensors*, Vol. 17 No. 2, p. 336, doi: 10.3390/s17020336.

TEDx (2017), "The real reason to be afraid of artificial intelligence | peter haas | TEDxDirigo", *Youtube*, 15 December, available at: https://www.youtube.com/watch?v=TRzBk_KuIaM (accessed 27 April 2022).

Trier, Ø.D., Larsen, S.Ø. and Solberg, R. (2009), "Automatic detection of circular structures in high-resolution satellite images of agricultural land", *Archaeological Prospection*, Vol. 16 No. 1, pp. 1-15, doi: 10.1002/arp.339.

Trier, Ø.D., Zortea, M. and Tonning, C. (2015), "Automatic detection of mound structures in airborne laser scanning data", *Journal of Archaeological Science: Reports*, Vol. 2, pp. 69-79, doi: 10.1016/j.jasrep.2015.01.005.

Trier, Ø.D., Cowley, D.C. and Waldeland, A.U. (2019), "Using deep neural networks on airborne laser scanning data: results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland", *Archaeological Prospection*, Vol. 26 No. 2, pp. 165-175, doi: 10.1002/arp.1731.

Tsetskhladze, G. (1998), "Who built the scythian and thracian royal and elite tombs?", *Oxford Journal of Archaeology*, Vol. 17 No. 1, pp. 55-92, doi: 10.1111/1468-0092.00051.

Valchev, T. and Sobotkova, A. (2019), "Monitoring burial mounds in the Yambol province: deploying mobile technology to improve cultural heritage protection", in Kyriakidis, P., Agapiou, A. and Lysandrou, V. (Eds), *Spreading Excellence in Computer Applications for Archaeology and Cultural Heritage. Proceedings of the 3rd Conference on Computer Applications and Quantitative Methods in Archaeology Greek Chapter (CAA-GR) Limassol*, Limassol-Cyprus, 18-20 June 2018, Cyprus University of Technology, pp. 19-23.

Vasileva, D. (2005), *The Thracian Tombs - Architectural-Metrical Study*, Vol. Suppl. III, Sofia University St. Kliment Ohridski, Sofia.

Verschoof-van der Vaart, W.B. and Lambers, K. (2019), "Learning to look at LiDAR: the use of R-CNN in the automated detection of archaeological objects in LiDAR data from The Netherlands", *Journal of Computer Applications in Archaeology*, Vol. 2 No. 1, pp. 31-40, doi: 10.5334/jcaa.32.

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

Verschoof-van der Vaart, W.B., Lambers, K., Kowalczyk, W. and Bourgeois, Q.P.J. (2020), "Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from The Netherlands", *Isprs International Journal of Geo-Information*, Vol. 9 No. 5, p. 293, doi: 10.3390/ijgi9050293.

Vinkers, C.H., Tijdink, J.K. and Otte, W.M. (2015), "Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis", *BMJ*, Vol. 351, p. h6467, doi: 10.1136/bmj.h6467.

Wald, D.M., Longo, J. and Dobell, A.R. (2016), "Design principles for engaging and retaining virtual citizen scientists", *Conservation Biology: The Journal of the Society for Conservation Biology*, Vol. 30 No. 3, pp. 562-570, doi: 10.1111/cobi.12627.

Wang, J. and Perez, L. (2017), "The effectiveness of data augmentation in image classification using deep learning", *Convolutional Neural Networks for Visual Recognition*, Vol. 11, pp. 1-8, available at: http://vision.stanford.edu/teaching/cs231n/reports/2017/pdfs/300.pdf

Weiss, K., Khoshgoftaar, T.M. and Wang, D. (2016), "A survey of transfer learning", *Journal of Big Data, SpringerOpen*, Vol. 3 No. 1, pp. 1-40, doi: 10.1186/s40537-016-0043-6.

West, S. and Pateman, R. (2016), "Recruiting and retaining participants in citizen science: what can be learned from the volunteering literature?", *Citizen Science Theory and Practice*, Vol. 1 No. 2, p. 15, doi: 10.5334/cstp.8.

Wheeler, M.A., Vylomova, E., McGrath, M.J. and Haslam, N. (2021), "More confident, less formal: stylistic changes in academic psychology writing from 1970 to 2016", *Scientometrics*, Vol. 126 No. 12, pp. 9603-9612, doi: 10.1007/s11192-021-04166-9.

Wildesen, L.E. (1982), "The study of impacts on archaeological sites", *Advances in Archaeological Method and Theory*, Vol. 5, pp. 51-96, doi: 10.1016/b978-0-12-003105-4.50007-8.

Woolf, T. (2018), "Deep convolutional neural networks for remote sensing investigation of looting of the archeological site of Al-Lisht, Egypt, Msc, University of Southern California, August.

Xiong, Y., Chen, Q., Zhu, M., Zhang, Y. and Huang, K. and IEEE (2020), "Accurate detection of historical buildings using aereal photographs and deep transfer learning", *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1592-1595, doi: 10.1109/IGARSS39084.2020.9323541.

Yao, M., Wei, Y. and Wang, H. (2023), "Promoting research by reducing uncertainty in academic writing: a large-scale diachronic case study on hedging in Science research articles across 25 years", *Scientometrics*, Vol. 128 No. 8, pp. 4541-4558, doi: 10.1007/s11192-023-04759-6.

Yuan, Z.-M. and Yao, M. (2022), "Is academic writing becoming more positive? A large-scale diachronic case study of Science research articles across 25 years", *Scientometrics*, Vol. 127 No. 11, pp. 6191-6207, doi: 10.1007/s11192-022-04515-2.

Corresponding author
Adela Sobotkova can be contacted at: adela@cas.au.dk