

RESEARCH ARTICLE

WILEY

Automated large-scale mapping and analysis of relict charcoal hearths in Connecticut (USA) using a Deep Learning YOLOv4 framework

Wouter Verschoof-van der Vaart¹  | Alexander Bonhage² | Anna Schneider² | William Ouimet³ | Thomas Raab²

¹Leiden University, Leiden, The Netherlands

²Brandenburgische Technische Universität
Cottbus-Senftenberg, Cottbus, Germany

³University of Connecticut, Mansfield,
Connecticut, USA

Correspondence

Wouter Baernd Verschoof-van der Vaart,
Leiden University, Leiden, The Netherlands.
Email: w.b.verschoof@arch.leidenuniv.nl

Funding information

German Research Foundation, Grant/Award
Number: RA 931/8-1

Abstract

In the past decade, numerous studies have successfully mapped thousands of former charcoal production sites (also called relict charcoal hearths) manually using digital elevation model (DEM) data from various forested areas in Europe and the north-eastern USA. The presence of these sites causes significant changes in the soil physical and chemical properties, referred to as legacy effects, due to high amounts of charcoal that remain in the soils. The overwhelming amount of charcoal hearths found in landscapes necessitates the use of automated methods to map and analyse these landforms. We present a novel approach based on open source data and software, to automatically detect relict charcoal hearths in large-scale LiDAR datasets (visualized with Simple Local Relief Model). In addition, the approach simultaneously provides both general as well as domain-specific information, which can be used to further study legacy effects. Different versions of the methodology were fine-tuned on data from north-western Connecticut and subsequently tested on two different areas in Connecticut. The results show that these perform adequate, with F1-scores ranging between 0.21 and 0.76, although additional post-processing was needed to deal with variations in LiDAR quality. After testing, the best performing version of the prediction model (with an average F1-score of 0.56) was applied on the entire state of Connecticut. The results show a clear overlap with the known distribution of charcoal hearths in the state, while new concentrations were found as well. This shows the usability of the approach on large-scale datasets, even when the terrain and LiDAR quality varies.

KEYWORDS

anthropogenic landforms, GIS, historical charcoal production, legacy effects, machine learning

This research was funded by the German Research Foundation (DFG) under grant number RA 931/8-1.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Archaeological Prospection* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Recent archaeological and pedological research has shown the widespread presence of former charcoal production sites, also called relict charcoal hearths (RCHs), across forested areas in Europe and the north-eastern USA (for an extensive overview see Hirsch et al., 2020). These landforms are mainly of pre-industrial age and persist in the present-day landscape as earthworks, that is, circular-shaped, elevated platforms of on average 11 m in diameter (generally between 8 and 15 m diameter) surrounded by a shallow ditch or multiple pits. LiDAR-based digital elevation model (DEM) analysis has proven an effective tool to find, map and investigate them (Figure 1).

One of the reasons for investigating RCHs is the value of these objects for geoarchaeological and pedological research. Recent larger scale studies have analysed the spatial association between RCH and historic industrial sites (Schneider et al., 2020), emphasizing the scale dependent heterogeneity of site densities (Schneider et al., 2022). RCH sites are often studied for past-landscape and historic reconstruction efforts (e.g., Dupin et al., 2019; Tolsdorf et al., 2015). The charcoal remaining in these RCHs is of particular value to reconstruct historic forest compositions and abundances (Deforce et al., 2013; Gocel-Chalté et al., 2020). As a result of high amounts of charcoal remaining in the soils, these sites (categorized as Spolic Technosols; IUSS Working Group World Reference Base, WRB, 2014) have a distinct enrichment of total organic and pyrogenic carbon (e.g., Borchard et al., 2014; Hirsch et al., 2017), resulting in significant changes of the soil physical and chemical properties (e.g., Donovan et al., 2020; Schneider et al., 2019). This enrichment is also affecting vegetation patterns and dynamics as well as microbial growth and abundance (e.g., Raab et al., 2022). Efforts have been made to quantify the changes caused by historical charcoal burning on today's soil landscapes, referred to as legacy effects, on a larger-than-site specific scale. For instance, Bonhage, Hirsch, Schneider, et al. (2020) found a positive correlation between local slope, RCH site volume, and stratigraphy, that is, on steeper slopes the sites total volume increases and vice versa. By using a GIS-based modelling approach, it enables the

quantification of nutrient and carbon stocks in RCH sites based on their topographical position. Thereby, information about the local topography and the sites' surface area is of high interest next to the site's location itself. A uniform and seamless large scale mapping can significantly help to improve the assessment of legacy effects on soil landscapes caused by historic land-use, whether it be in terms of morphology (e.g., how much soil substrate was redistributed) or soil properties (e.g., carbon and element stocks), and it can help in the effort to identify large scale clusters of site occurrences.

In the past decade, various studies have successfully mapped thousands of RCHs by hand using DEM data from various regions (Carrari et al., 2017; Deforce et al., 2013; Hazell et al., 2017; Hesse, 2010; Johnson et al., 2015; Raab et al., 2019; Risbøl et al., 2013; Rutkiewicz et al., 2019). For instance, the manual analysis of DEM data from northwestern Connecticut, USA resulted in the discovery of over 20 000 RCHs (Johnson et al., 2015; Johnson & Quimet, 2021; Raab et al., 2017; see also Table 1).

This overwhelming number of RCHs, combined with an ever-increasing set of available, high-quality, remotely sensed data necessitates the use of computer-aided methods for the automatic detection of these objects, thereby alleviating the complications surrounding manual analysis, for example, biased and heterogeneous detection accuracy (Quintus et al., 2017; Risbøl et al., 2013; Sadr, 2016), and documentation (Bennett et al., 2014; Bevan, 2015). Furthermore, in order to model legacy effects using these enormous numbers of RCHs, automation in the calculation of domain specific information based on their topographical position (e.g., local slope) is needed as well.

Previous research in an effort to automatically detect RCHs has relied on various methods including Template Matching (Schneider et al., 2015; Trier & Pilø, 2012) and Geographic Object-Based Image Analysis (Witharana et al., 2018), whereas more recently, Machine Learning approaches are being developed and utilized (Anderson, 2019; Bonhage et al., 2021; Carter et al., 2021; Davis & Lundin, 2021; Kazimi et al., 2020, 2019; Oliveira et al., 2021; Suh et al., 2021; Trier et al., 2021, 2018; Verschoof-van der Vaart

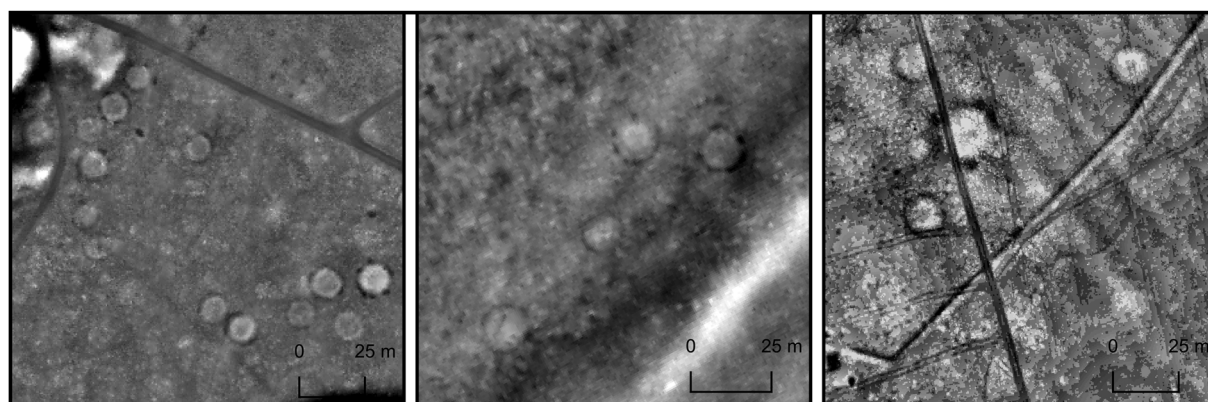


FIGURE 1 Excerpts of LiDAR data, visualized with Simple Local Relief Model (Hesse, 2010), showing examples of relict charcoal hearths in the Netherlands (left), Connecticut, USA (centre) and Germany (right).

TABLE 1 Overview of the research areas and the number and density of manually mapped relict charcoal hearts

Location	Type	Sq. km	Known RCH	Density (RCH/sq. km)	Forest cover	Ground point/sq. m
Huntsville, Litchfield County CT	Training area	40	3040	76	92%	2.44
Canaan, Litchfield County CT	Test area 1	11	314	29	77%	3.98
Ashford, Windham County CT	Test area 2	302	527	1.75	87%	2.13
State of Connecticut	Application	12 189	24 051	2	66%	-

Abbreviation: RCH, relict charcoal hearth.

et al., 2020). Deep Learning (LeCun et al., 2015), a subfield of Machine Learning, predominantly employs Convolutional Neural Networks (CNNs)—hierarchically structured algorithms that generally consist of a (image) feature extractor and classifier (Guo et al., 2016)—that learn to generalize from a large set of labelled examples, rather than relying on a human operator to set parameters or formulate rules. To date, these automated methods are mainly tested in an experimental setting but have yet to be applied in various contexts or on a large (e.g., regional or national) scale (Verschoof-van der Vaart et al., 2020; but see for instance Berganzo-Besga et al., 2021; Davis et al., 2021; Orenge et al., 2020), with this being the aim of previous initiatives (Trier et al., 2019). Questions remain concerning the usability of these approaches for large-scale surveys and the transferability of these methods when applied outside of the area where they were developed (Cowley et al., 2020; Kermit et al., 2018; Verschoof-van der Vaart & Landauer, 2022). Furthermore, archaeological automated detection can still be considered as being in a developmental stage (Opitz & Herrmann, 2018), and as such, studies generally focus on the training and testing of automated methods and the resulting detection rates, while the further analysis and use of the generated archaeological information are often not part of these studies (Davis, 2019). Leading to the understanding that, opportunities to automate the subsequent analysis of detections to gain domain-specific information are rarely explored.

In this paper, we present a novel approach to automatically detect RCHs in LiDAR-based DEM data. For this, a state-of-the-art Deep Learning object detection framework, YOLOv4, has been transfer-learned; that is, the model has been pre-trained on a generic image dataset and subsequently fine-tuned on our own dataset and combined with (GIS) processing algorithms with a focus on being effective and efficient to handle large amounts of spatial data (e.g., entire states). Simultaneously, it is designed to be able to provide both general information (e.g., the location), as well as domain-specific information, such as the area (i.e., the area covered by the RCH's platform and ditch) and average local slope (i.e., the slope in the direct vicinity of the RCH; see Bonhage, Hirsch, Schneider, et al., 2020; Johnson & Ouimet, 2021), of the objects of interest (see also Verschoof-van der Vaart, 2022). A further aim of this research is to utilize open-source data and software, to make this research more transparent, reproducible and more readily implementable by other researchers (Schmidt & Marwick, 2020).

2 | RESEARCH AREAS AND LIDAR DATA

In order to train, test and apply the developed method, four different, predominantly forested areas were first defined (see Table 1 and Figure 2). The Huntsville area (circa 40 km²) comprises the northern part of the Housatonic State Forest between Huntsville, Cornwall Hollow and West Cornwall in Litchfield County located in north-western Connecticut (Figure 2, Training area). This wider area is part of the Salisbury Iron district and has a very high density of known RCHs (Johnson et al., 2015; Johnson & Ouimet, 2021; Raab et al., 2017). RCH sites from the Huntsville area have been used for the training of our approach (see Section 3.1).

To test the developed method, two separate areas in Connecticut were selected. The Canaan area (approximately 11 km²), or Test area 1, lies circa 10 km north of the Huntsville area near the town of Canaan within the Salisbury Iron district. Due to the close proximity, this area closely resembles the training area in terms of landscape and to a lesser extent in known RCH density (Table 1). The discrepancy in forest cover (see <https://www.sciencebase.gov/catalog/item/5b6fa9d4e4b0f5d57878e707>) between these two areas is in large part caused by Washining Lake in the southwest of the Canaan area. The Ashford area (302 km²), or Test area 2, lies approximately 100 km to the east of the Huntsville area, in Windham County in eastern Connecticut. Although this area has a comparable topography and forest cover as the Huntsville area, the density of known RCHs is much lower and more comparable with the overall density of RCHs in Connecticut (Table 1; see also Suh et al., 2021). The difference in size and RCH density of the two test areas offers opportunities to see the influence of these parameters on the performance of the developed method (see Verschoof-van der Vaart, 2022).

After testing, the developed method has been applied on a large scale. The application area equals the entire state of Connecticut (approximately 12,542 km²), excluding the training and testing areas. As of this publication, about 24 000 RCHs are known from the entire state of Connecticut according to the publicly available Northeastern US Relict Charcoal Hearth ArcGIS Online Web Map (Ouimet, 2019).

Statewide LiDAR-based DEM data (the 2016 dataset with a raster cell size of 0.6 m) were acquired from the openly accessible repository of the Connecticut Department of Energy and Environmental Protection (<http://www.cteco.uconn.edu>). The global accuracy of this dataset is stated as a horizontal accuracy of ±1.0 m and a vertical accuracy

of ± 0.138 – 0.170 m. In addition, we assessed the quality of the LiDAR data for the three subregions by randomly selecting 10 point cloud .las tiles (760×760 m each) and then averaging their calculated ground point density, which varied between 2.1 and 4 (Table 1).

The downloaded DEM tiles were merged into a single file, which was subsequently split into 25 tiles (around 800 km^2 each) to ease further computational requirements. This procedure was a compromise of reducing the total number of DEM tiles and keeping the files sizes manageable. The 25 DEM tiles were visualized, using the RVT toolbox (see Kokalj & Hesse, 2017), with Simple Local Relief Model (SLRM; Hesse, 2010) and subsequently merged. The choice to use SLRM was based on several factors such as the fact that this visualization most clearly shows the RCHs for the purpose of human observation. Also, in prior automated detection research on RCHs the use of SLRM had led to satisfactory results (Trier et al., 2021; Verschoof-van der Vaart, 2022). We are aware that multiscale visualization techniques can result in better detection results when using Deep Learning (Guyot et al., 2021). Even though, SLRM produces relatively high results in a recent study comparing visualizations (Guyot et al., 2021) as well as in less recent publications (Gallwey et al., 2019; Kazimi et al., 2020). More importantly, we choose efficiency over total accuracy as we are dealing with (very) large-scale datasets. Compared with

other visualizations and especially blends and multiscale visualizations, SLRM is more easy to implement and can be utilized using open source tools that allow rapid batch processing of large datasets.

3 | METHODOLOGY

The aim of this study is to create an effective and efficient approach to automatically detect RCHs in large amounts of DEM data, while simultaneously providing both general (e.g., location) as well as domain-specific information (e.g., area and local slope). The developed approach (Figure 3), called ARCHMAGE (Automated Relict Charcoal Hearth Mapping And Geospatial Exploration), combines the state-of-the-art object detection framework, YOLOv4 (Bochkovskiy et al., 2020), with different (GIS) processing algorithms. The workflow can be divided into three parts: a preprocessing part that converts the DEM data into input images (subtiles), based on geospatial information about known RCHs (Section 3.1); a part concerning the Deep Learning-based object detection (Section 3.2); and a post-processing part where the results of the object detection are converted back into geospatial data (Section 3.3). In addition, two domain-specific parameters, that is, area and local slope, are calculated.

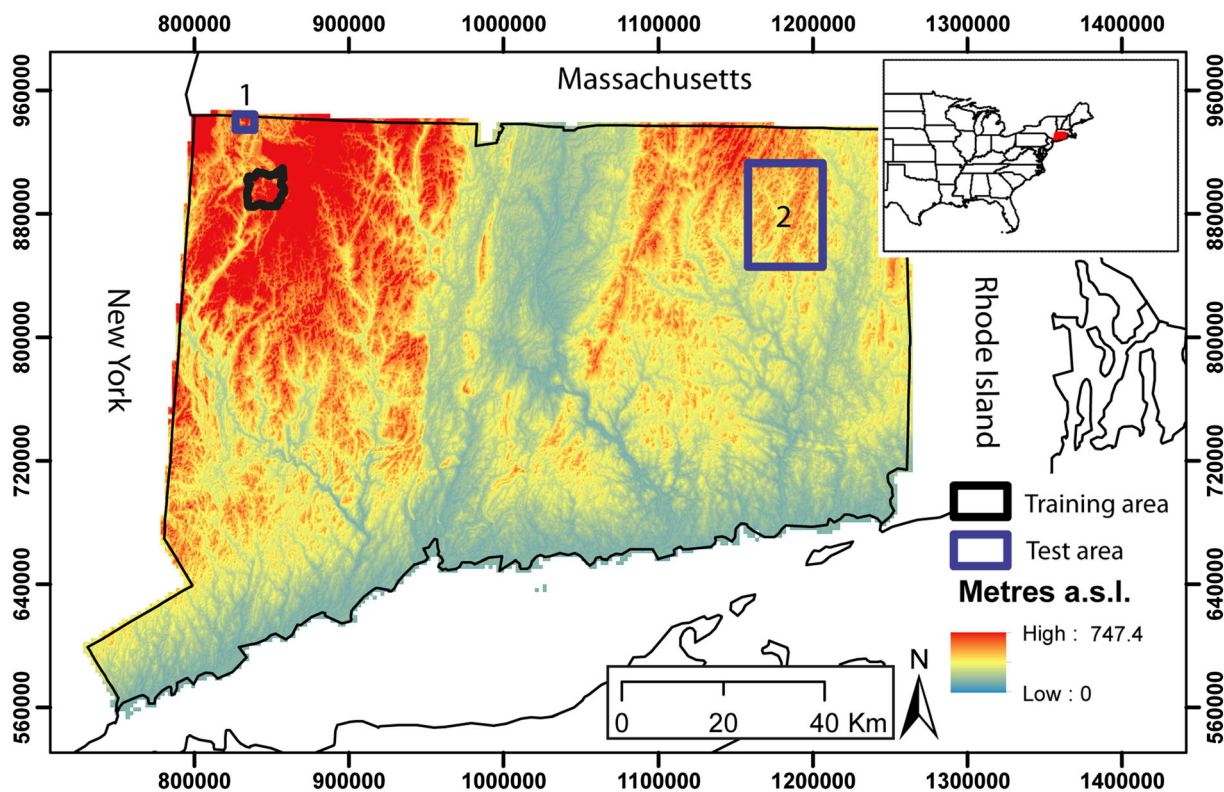


FIGURE 2 The research areas on an elevation map of Connecticut, USA [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/arp.1889)]

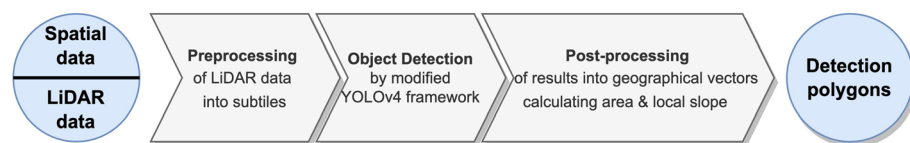


FIGURE 3 Simplified representation of the ARCHMAGE workflow [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/arp.1889)]

3.1 | Preprocessing of DEM data into subtiles

To generate a training dataset for the object detection model, all RCHs in the Huntsville area (approximately 40 km²; see Table 1) were manually mapped. The results of a prior analysis of the 2011 LiDAR data of Litchfield County (Johnson et al., 2015) were used as a starting point. These data on the location of RCHs originally consisted of point data, which we converted into polygons, using the outer edge of RCH sites as a maximum extent. Subsequently, two experts—both with ample experience in the analysis of DEM data—independently marked all other visible RCHs in the DEM data. The annotations were combined and compared in QGIS (QGIS Development Team, 2017). This resulted in a dataset containing 3040 unique RCH sites, of which 2350 were known from the 2015 analysis (see Johnson et al., 2015). The latter shows the problem of heterogeneous detection accuracy between operators in manual analysis, which is further enhanced by the use of DEM data of different date, quality and different visualizations (see Quintus et al., 2017; Risbøl et al., 2013; Sadr, 2016).

To make input subtiles for the YOLOv4 model, a modified version of the dataset generation method developed by Olivier and Verschoof-van der Vaart (2021) was used. This python script uses a spatial dataset of objects (in this case known RCH sites) to crop LiDAR data into smaller images (subtiles) with set dimensions (in this case 500 × 500 pixels or approximately 305 × 305 m). For every individual object in the dataset, a subtitle is generated, centred on the location of the object. To avoid bias, a small shift (or jitter) is added so that the RCH is not in the exact centre of the subtitle. However, as RCHs are often spatially clustered, these cropped subtiles generally contain multiple RCHs, and therefore, a RCH site can appear on multiple subtiles. This ‘re-use’ of objects has proven beneficial, as it not only increases the number of subtiles in the training dataset but also increases variability (Olivier & Verschoof-van der Vaart, 2021).

A selection of 2427 RCHs from the 3040 RCHs found in the manually mapped dataset were used to crop the SLRM visualized DEM data of the Huntsville training area. The remaining 613 RCHs were omitted as these were mapped in the prior analysis of Johnson et al. (2015) but could not be verified in the LiDAR data used in this research. Subsequently, all subtiles were randomly split 80/20 into a train and validation dataset (Table 2), as is common practice in the training of Deep Learning algorithms (Goodfellow et al., 2016).

As the testing and application areas simply need to be split into equal parts, not based on the location of known RCHs, an additional

python script was used that cuts the DEM data into subtiles of 500 × 500 pixels (approximately 305 × 305 m) with a 25 pixel (approximately 15 m) overlap to all sides. The latter is done to avoid the dissecting of potential RCHs at the edge of the subtiles. A drawback of this overlap is the occurrence of multiple overlapping predictions (i.e., bounding boxes) on RCHs at the edge of multiple subtiles.

This resulted in two test datasets of, respectively, 143 and 3672 subtiles, whereas the application dataset consists of 214 620 subtiles (Table 2).

3.2 | Object detection with the YOLOv4 framework

The object detection portion of the ARCHMAGE workflow consists of the YOLOv4 detection framework¹ (Bochkovskiy et al., 2020), a recent implementation of the YOLO (‘You Only Look Once’) framework (Redmon et al., 2016). YOLOv4 is a so-called ‘one-stage’ detector, which, contrary to ‘two-stage’ detectors such as Faster R-CNN (Ren et al., 2017), combines the two parts of object detection (i.e., object localization and classification) in one process and approaches this as one would a regression problem. More specifically, as part of the input of YOLOv4, subtiles are downsampled to a fixed resolution and subsequently divided into an equally spaced grid. For every cell within this grid, a set amount of bounding boxes with confidence scores is predicted as well as a class probability (Bochkovskiy et al., 2020). This approach dramatically decreases the inference time (i.e., the speed of detections), generally without a loss of accuracy. For example, a comparison of the testing time between YOLOv4 and Faster R-CNN on DEM data showed that the former was about 18 times faster than the latter (see Olivier & Verschoof-van der Vaart, 2021). Although speed is normally not the focus of archaeological automated detection research, the aim to develop an effective and efficient approach for large-scale mapping—which involves massive amount of data up to several hundred gigabytes—necessitates a fast inference time without a loss in performance, especially on small objects which the YOLOv4 framework provides (Bochkovskiy et al., 2020; Carranza-García et al., 2021).

3.2.1 | Adjusting YOLOv4 for archaeological object detection

Implementing an off-the-shelf version of an object detection framework, pre-trained on a general-purpose dataset, for example, Microsoft COCO (Lin et al., 2014), on the task of detecting (archaeological) objects in remotely sensed data generally results in an unsatisfactory performance (Verschoof-van der Vaart & Lambers, 2019). This is directly related to the differences between the (objects in) images in general-purpose datasets and DEM data (Olivier & Verschoof-van der

TABLE 2 The number of subtiles per research area, resulting from the dataset generation

Location	Number of subtiles
Huntsville, Litchfield County CT (Training area)	1942 (training)
	485 (validation)
Canaan, Litchfield County CT (Test area 1)	143
Ashford, Windham County CT (Test area 2)	3672
State of Connecticut (Application area)	214 620

¹In this research, the code provided by Technizou (<https://medium.com/analytics-vidhya/train-a-custom-yolov4-object-detector-using-google-colab-61a659d4868#4be1>) was used to construct our model.

Vaart, 2021; Verschoof-van der Vaart, 2022). The main problem is that general-purpose datasets contain ‘natural images’ (i.e., photographs of scenes seen in normal, every-day settings) in which objects are generally large and prominent and occupy a major portion of the image. Object detection methods normally take advantage of this by downscaling (and pooling) the images, when they pass through the CNN to greatly reduce the computational cost (Guo et al., 2016). For instance, in the YOLOv4 framework input, images are downscaled to a fixed size, normally 416×416 pixels. However, within DEM data (and remotely-sensed imagery in general), the objects of interest are generally very small, especially in relation to the size of the images. Consequently, downscaling on these images will result in information loss and the removal of small objects, which makes it impossible to detect them (Olivier & Verschoof-van der Vaart, 2021). Therefore, in our versions of YOLOv4, the input size was set to 512×512 pixels—while the actual size of the subtile is 500×500 pixels—to prevent the downsampling of the subtile in the CNN (Table 3).

Another common problem remains in the fact that the objects of interest in remotely sensed imagery are often densely clustered but scarcely distributed. This means that while many subtile in the test and application datasets will be empty, some will contain a large number of RCHs. Traditional object detection methods can generally only detect a select number of objects within a single subtile, although this can be adjusted. Therefore, the maximum number of detections per subtile was increased to 200 in the YOLOv4 framework used in this study (Table 3). Finally, several data augmentation techniques (Goodfellow et al., 2016) were implemented to improve the robustness of the model to deal with occlusion and fragmentation of RCHs in the data (for an extensive explanation of augmentation techniques, see Bochkovskiy et al., 2020; Redmon et al., 2016).

TABLE 3 Overview of the general (hyper)parameters for the YOLOv4 framework (before specific modifications) used in this research

(Hyper)parameter	Value/type
backbone CNN	CSPDarknet53
input size	512×512
batchsize	64
learning rate	Initially 0.001
epochs	4000
data augmentation	Photometric distortions: brightness, contrast, hue, saturation and noise; geometric distortion: random scaling, cropping, flipping, and rotating; mosaic
regularization	Dropblock
loss function	CloU
non-maximum suppression	greedyNMS

In addition to the general adjustments to the YOLOv4 framework, we experimented with different upsample strides, loss functions and accuracy metrics. To measure the influence of these, six different versions of the YOLOv4 model were created, transfer-learned, and tested on the two test areas (Table 4).

Upsample stride

To further enhance YOLOv4's performance on small objects, that is, smaller than 19×19 pixels (the average size of RCH in the dataset is approximately 17×17 pixels or 11 by 11 m), the upsampling stride in YOLOv4 can be increased. At two points in the framework the image is upsampled, that is, if the image is upsampled with a stride 2, one pixel is transformed into four pixels in a 2 by 2 area. By increasing the stride parameter from 2 to 4, small objects, which constitute a few pixels, will appear larger in the upsampled image. A downside can be that the element of scale becomes more difficult.

Loss function

During the training of the CNN, the loss function—a function that calculates the penalties of incorrect classifications into a single number (Goodfellow et al., 2016)—is optimized. A low loss value is generally regarded as an indication for a well-trained approach and therefore high performance (Guo et al., 2016). In the case of object detection, the Intersection Over Union (IoU) is often used as loss function, which gives a measure for the overlap between the predicted bounding box and the ground truth. While this metric gives a good indication for bounding box quality, it completely disregards the positional relation between the predicted bounding box and the ground truth (Rezatofighi et al., 2019). Therefore, YOLOv4 employs an improved version of the IoU, the Complete IoU (CloU), which uses three parameters, that is, the overlap between predicted bounding box and ground truth, the distance between the centre point of the predicted bounding box and ground truth and the aspect ratio between the two (Zheng et al., 2021). However, other loss functions, which also take the position of the bounding box in regard to the ground truth into account, might prove more beneficial. Therefore, several versions (3–6) have been outfitted with the Generalized IoU (GloU) (Rezatofighi et al., 2019). The GloU takes into account the distance between the predicted bounding box and ground truth, as well as their overlap by using the size of a box enclosing the prediction and the ground truth.

TABLE 4 Overview of the different versions of the YOLOv4 framework used in this research

Model	Upsample stride	Loss function	Accuracy metric
1	2	CloU	mAP@0.5
2	4	CloU	mAP@0.5
3	2	GloU	mAP@0.5
4	4	GloU	mAP@0.5
5	2	GloU	mAP@0.9
6	4	GloU	mAP@0.9

Accuracy metric

The last modification involves using GloU (see above) and changing the metric used to evaluate the quality of detections during training, that is, the mean average precision, or mAP. This metric is changed from mAP@0.5 to mAP@0.9, which means that only predictions whose overlap with the ground truth is over 0.9 (instead of 0.5) are regarded as True Positives when determining mAP (Everingham et al., 2010). This modification could make the detected bounding boxes more accurate and consequently improve the overall performance of the detection framework.

3.3 | Post-processing

The initial post-processing part of the ARCHMAGE workflow consists of three steps: (1) Converting predicted bounding boxes to geospatial polygons and calculate their area; (2) Filter detections by land-use; and (3) Calculate the average slope in the vicinity of the detections. The output of the YOLOv4 model is one text file per subtitle, with a list of detected RCHs comprising pixel coordinates for the bounding box and a confidence score (range 0–100).

The first post-processing step involves converting these pixel coordinates into a 'real-world' coordinate system, so that these can be managed and analysed in a GIS environment (see also Verschoof-van der Vaart & Lambers, 2019). Therefore, a python script, based on the *detectionsToCSV* script by Olivier and Verschoof-van der Vaart (2021), was used to convert the content of the txt file into a CSV file by connecting every subtitle to the DEM tile from which it originally derived. The real-world coordinates from the DEM tile are used to compute the real-world coordinates of the bounding boxes and add them to the CSV file. Then, the area (i.e., the area covered by the potential RCH's platform and ditch) is calculated using the formula for the area of an oval, based on the size of the bounding box and the resolution, that is, cell size, of the DEM tile (Equation 1). However, this initially produced unsatisfactory results, as the predicted bounding boxes are generally larger than the actual RCH (Figure 4). Further analysis of this problem showed that most ground truths are also larger than the actual RCH, which might have led to larger predicted bounding boxes. To cope with this problem, a negative constant (α and β ; in this research, 5 and 6 pixel respectively) was added to the formula. This resulted in areas much closer to the manually calculated analogs.

$$\text{Area} = \pi * ((\text{width} - \alpha) * \text{resolution}) / 2 * ((\text{height} - \beta) * \text{resolution}) / 2. \quad (1)$$

In the second step of the post-processing, the detections are loaded into a GIS environment, and the location of every potential RCH is compared with a spatial layer containing forested areas (see <https://www.sciencebase.gov/catalog/item/5b6fa9d4e4b0f5d57878e707>). All detections situated outside of woodland are discarded, as research has shown that over 95% of all

known RCHs are situated in forested areas (deciduous and conifer forest, Johnson et al., 2015). This aids in filtering out some False Positives created by landscape objects such as swimming pools, roundabouts and so on.

Finally, the average slope in the vicinity of the potential RCHs has been automatically calculated using a combination of different GIS processing algorithms (see Algorithm 1). This process is comparable with how the average slope surrounding RCHs is manually calculated usually (see for instance Johnson & Ouimet, 2021). For every

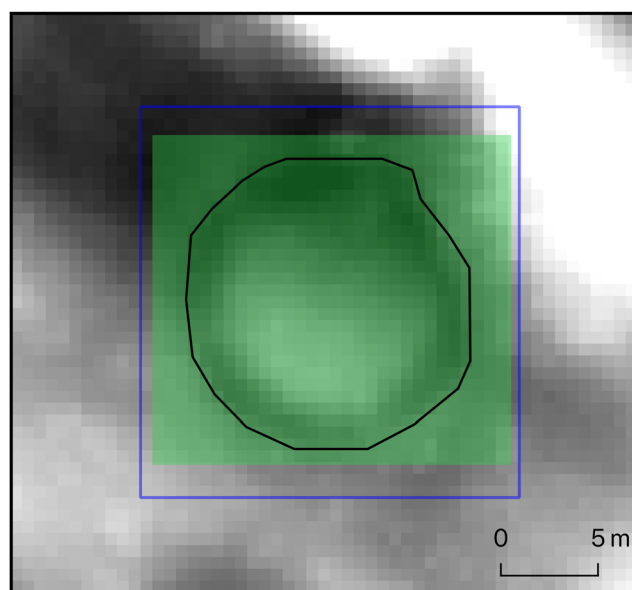


FIGURE 4 Excerpt of LiDAR data, visualized with Simple Local Relief Model (Hesse, 2010), showing the outline of a RCH (black), the ground truth (green) and the predicted bounding box (blue) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

Algorithm 1 Calculation of the average slope in the vicinity of RCH detections.

Algorithm 1: Average Slope

Input: detection polygons

Output: average slope value

```

1 for each polygon do
2   buffered_bbox ← compute buffer (20m radius) of
   polygon
3   surrounding_bbox ← compute difference
   between buffered_bbox and polygon
4   average_slope ← median of surrounding_bbox on
   slope map (5m resolution)
5   slope_class ← classify average_slope in ≤4 /
   4–8 / 8–12 / 12–16 / >16
6 end

```

bounding box, a buffer with 20 m radius is generated. The original bounding box is removed from this buffer to reduce distortions in the average slope value due to slope variation within the RCH. The resulting polygon is compared with a slope map (5 m cell size). Empirically, we found that the median of slope values inside the polygon gave the best results as compared with manually generated average slope values and was less affected by outliers. In addition, the detections are divided into five classes ($<4^{\circ}/4^{\circ}-8^{\circ}/8^{\circ}-12^{\circ}/12^{\circ}-16^{\circ}/>16^{\circ}$), based on the calculated average slope value. The latter is done to make an informed distinction between RCHs on (relatively) flat terrain and more steep terrain.

3.4 | Implementation details

We used the six versions of the YOLOv4 framework with CSPDarknet53 as the backbone CNN. The CNN was pre-trained on the Microsoft COCO dataset (Lin et al., 2014) and fine-tuned for 4000 epochs with an initial learning rate of 0.001 on our own training dataset (see Section 2). During and directly after training, the performance of the six versions of YOLOv4 was determined on the validation dataset. Subsequently, the models were used to detect RCHs in the test datasets (see Tables 1 and 2), and the results were post-processed (see Section 3.3) and evaluated.

The training and testing of the frameworks was implemented on the browser-based Colaboratory platform (Colab Pro+) from Google Research (Google, 2022). Google Colab is a specialized version of Jupyter Notebook, which is cloud-based and offers free computing resources (e.g., GPUs). This platform is connected to Google Drive, where the necessary data (e.g., training and testing dataset and model weights) are stored. Therefore, Colab provides straightforward implementation of the developed models, allows multiple researchers to access and use the same data and code and makes dissemination of the developed approach simple.

3.5 | Performance metrics

To evaluate and compare the performance of our Deep Learning model, the common metrics Recall (R; Equation 2), Precision (P; Equation 3) and F1-score (F1; Equation 4) were calculated (Chicco & Jurman, 2020; Verschoof-van der Vaart, 2022) by determining the number of True Positives (TP), False Positives (FP) and False Negatives (FN). Recall gives a measure of how many relevant objects are selected, whereas Precision measures how many of the selected items are relevant. The F1-score is the harmonic mean of the Precision and Recall and a single metric of the model's overall performance (Sammut & Webb, 2010). These metrics range between 0 and 1, where higher values indicate better performance.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}). \quad (2)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}). \quad (3)$$

$$\text{F1} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}). \quad (4)$$

To obtain the highest F1-scores, we employed threshold moving (Zou et al., 2016). By default, a Deep Learning model uses a certain confidence threshold: Detections with a confidence score that equals or exceeds this threshold are included in the results, whereas detections with a lower confidence score are discarded. This confidence threshold is generally set to an arbitrary number, typically 0.5. However, by changing the threshold and recalculating the performance metric, an optimal trade-off between Recall and Precision can be found, resulting in the highest F1-score (Zou et al., 2016). Therefore, we empirically calculated the optimal confidence threshold for the validation and test datasets and used these (see Table 5). Although this might complicate the comparison of the performance between datasets, it better shows the capability and maximum performance of the model on that particular dataset.

TABLE 5 The performance (Recall, Precision, F1) of the different versions of the YOLOv4 framework on the validation and test datasets. Notice the difference in confidence threshold between the different datasets.

Model	Upsample stride	Loss function	Accuracy metric	Validation (conf thresh 0.25)			Test area 1 (conf thresh 0.5)			Test area 2 (conf thresh 0.9)		
				Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
1	2	CloU	mAP@0.5	0.90	0.85	0.88	0.65	0.89	0.75	0.38	0.15	0.21
2	4	CloU	mAP@0.5	0.90	0.87	0.89	0.61	0.89	0.73	0.36	0.24	0.28
3	2	GloU	mAP@0.5	0.91	0.86	0.88	0.71	0.69	0.70	0.52	0.09	0.15
4	4	GloU	mAP@0.5	0.90	0.87	0.88	0.66	0.88	0.754	0.40	0.18	0.24
5	2	GloU	mAP@0.9	0.90	0.85	0.87	0.62	0.87	0.72	0.30	0.25	0.27
6	4	GloU	mAP@0.9	0.91	0.86	0.88	0.60	0.92	0.72	0.25	0.37	0.30

4 | RESULTS

4.1 | General results

Table 5 shows the performance of the different versions of the YOLOv4 framework on the validation and test datasets. The performance (F1) on the validation dataset is high, on average 0.88, which demonstrates that the detection method is suitable—at least on a technical level—for this specific task (see Verschoof-van der Vaart, 2022). The performance on both test areas is lower. The performance (F1) on Test area 1 is on average 0.73. Although the average Precision (0.86) is very similar to the validation dataset (0.86), a clear drop in average Recall can be observed (0.90 to 0.64). This reduced performance between the validation dataset (or a small, non-random, selective dataset) and test datasets ('realistic' datasets representing the real-world situation) has been observed in various studies (Soroush et al., 2020; Verschoof-van der Vaart et al., 2020) and is related to several factors, such as the proportion of positive and negative subtiles (i.e., subtiles with or without archaeological objects), the total number of labelled objects in the area as compared with the size of the area (i.e., the density), the variety in the state of preservation of the archaeological objects (Verschoof-van der Vaart et al., 2020) and land cover conditions (Suh et al., 2021).

Another factor that can have a considerable impact on the performance of detection methods are changes in the LiDAR data quality (Bonhage et al., 2021; Suh et al., 2021; Verschoof-van der Vaart & Landauer, 2022). This is clearly illustrated by the large decline in performance between Test area 1 and Test area 2 (Table 5).

An inspection of the properties of the LiDAR data of both test areas shows that Test area 2 has a lower ground point density than Test area 1 (Table 1). This results in the appearance of interpolation errors (Căţeanu & Ciubotaru, 2020). These errors are particularly problematic for this research, as their appearance in the interpolated and visualized DEM data is often round and thereby very comparable with RCHs (Figure 5). This results in the severe drop in Precision from on average 0.86 (Test area 1) to on average 0.21 (Test area 2). As a solution to this problem, an additional post-processing step was developed to automatically filter out the majority of FPs resulting from these interpolation errors.

Further investigation proved that the slope within the interpolation errors varies little; that is, the standard deviation (stdev) of the slope within the errors is low. Therefore, a combination of different GIS processing algorithms (Algorithm 2) was developed. For every bounding box (detection polygon), the centroid is computed, and a buffer with 4 m radius is generated. The size of the buffer was empirically determined to make sure no edges of the RCHs or the interpolation artefacts were included, while still allowing the computation of the slope standard deviation by including enough raster cells. These buffer polygons are compared with a slope map of the original DEM resolution, which has been created with an added z-factor of 10 and then has been subsequently converted to integer values to allow the use of the Zonal Statistic Toolbox in ArcGIS. For each buffer polygon, the standard deviation (stdev) of the slope is calculated. Subsequently,

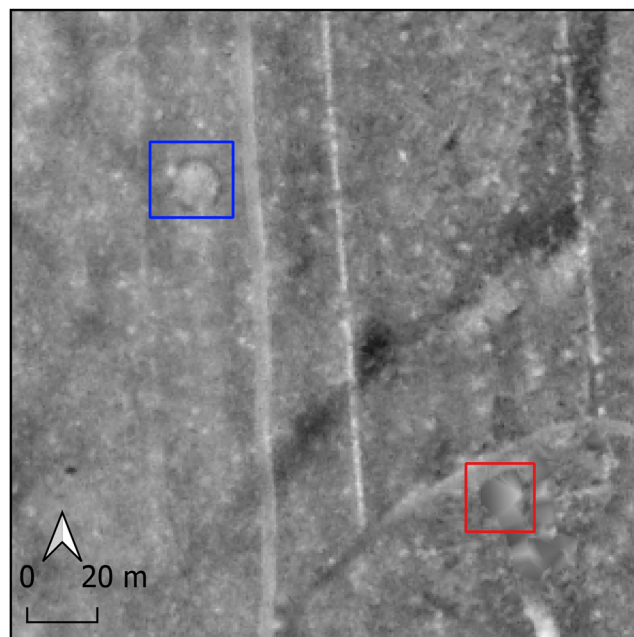


FIGURE 5 Excerpt of LiDAR data, in scale 1:2000, visualized with Simple Local Relief Model (Hesse, 2010), showing a relict charcoal hearth (RCH) in blue outline and an interpolation error in red outline [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/ar.1889)]

Algorithm 2 Removal of interpolation errors based on the standard deviation of the slope within detections.

Algorithm 2: Interpolation Error Removal

Input: detection polygons

Output: detections without interpolation error FPs

```

1 for each polygon do
2   centroid ← compute centroid of
   detection_polygon
3   buffered_point ← compute buffer (2m radius)
   of centroid
4   stdev_slope = stdev of buffered_point on
   slope map (0.6m resolution)
5   detections = join on location buffered_points
   and detection_polygons
6   Filtered_detections = remove detection if
   stdev_slope ≤ 4.5
7 end

```

the buffer polygon feature tables are joined with the detection polygons based on their location. Every detection polygon with a slope standard deviation lower or equal than the threshold of 4.5 is removed. This threshold was determined by comparing the slope standard deviation of a selection of the interpolation errors.

To evaluate the effectiveness of this post-processing step, the models' performance on Test area 1 and Test area 2 was recalculated (Table 6). Based on the results, interpolation errors account for 25–41% of the FPs in Test area 1, whereas in Test area 2, the percentage lies between 43% and 59%. The remaining FPs are mainly caused by a variety of natural landforms that 'mimic' RCHs. For Test area 1, a slight decrease in Recall (0–1 point) and an increase in Precision (2–8 points), with an overall increase in performance (F1) between 0 and 3 points, can be observed. The impact on the performance of Test area 2 is more notable. Whereas the Recall is only slightly decreasing, the Precision increases by 5–20 points; that is, the procedure is effective and efficient in predominantly deleting false positive sites. In summation, false positives are generally caused by areas with low ground point density (resulting in interpolation errors) but with a central elevation value higher than the surrounding values. Therefore, these areas can be automatically selected as being detections with low standard deviation of the elevation values.

4.2 | Results of the implementation of specific adjustments

The results (Tables 5 and 6) display that the specific adjustments are of influence on the performance. More specifically, the increase of the upsampling stride seems to have a negative influence on Recall, while improving Precision. Changing the loss function from CloU to GloU improves Recall but decreases Precision. Finally, changing the mAP results in an increase in Precision but a decrease in Recall. The version of YOLOv4 with the best Recall is model 3, which concerns the base model with GloU. However, this model also has the lowest Precision of all versions (see Tables 5 and 6). The opposite is true for model

6. This version reached the highest Precision but has the lowest Recall of all versions (Tables 5 and 6). The models with the overall best performance (F1) in both test areas are models 2 and 4, of which the former uses an increased upsampling stride and the latter an increased upsampling stride and GloU.

4.3 | Results of the post-processing

To evaluate the automatically computed area (see Equation 1), the results of 25 randomly selected RCHs in Test area 1 were compared with manually determined areas. The results (Figure 6) show that on

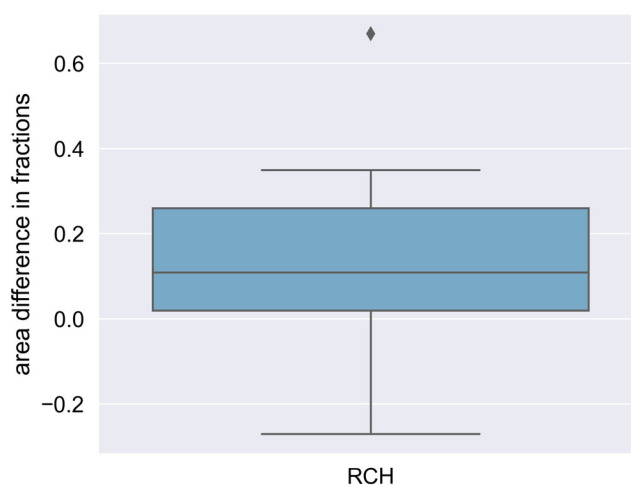


FIGURE 6 Boxplot graph showing the difference (in fractions) between the automatically and manually determined area of 25 relict charcoal hearths (RCHs) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/arp.1889)]

TABLE 6 The original and recalculated performance (Recall, Precision, F1) of the different versions of the YOLOv4 framework on the test datasets. Notice the difference in confidence threshold between Test areas 1 and 2.

Model	Upsample stride	Loss function	Accuracy metric	Test area 1 (conf thresh 0.5)Original			Test area 1 (conf thresh 0.5)Recalculated		
				Recall	Precision	F1	Recall	Precision	F1
1	2	CloU	mAP@0.5	0.65	0.89	0.749	0.64	0.96	0.757
2	4	CloU	mAP@0.5	0.61	0.89	0.73	0.61	0.93	0.73
3	2	GloU	mAP@0.5	0.71	0.69	0.70	0.70	0.76	0.73
4	4	GloU	mAP@0.5	0.66	0.88	0.754	0.65	0.91	0.759
5	2	GloU	mAP@0.9	0.62	0.87	0.71	0.62	0.92	0.74
6	4	GloU	mAP@0.9	0.60	0.92	0.725	0.59	0.94	0.72
Model	Upsample stride	Loss function	Accuracy metric	Test area 2 (conf thresh 0.9)Original			Test area 2 (conf thresh 0.9)Recalculated		
				Recall	Precision	F1	Recall	Precision	F1
1	2	CloU	mAP@0.5	0.38	0.15	0.21	0.37	0.29	0.32
2	4	CloU	mAP@0.5	0.36	0.24	0.28	0.34	0.42	0.38
3	2	GloU	mAP@0.5	0.52	0.09	0.15	0.5	0.14	0.21
4	4	GloU	mAP@0.5	0.40	0.18	0.24	0.39	0.32	0.35
5	2	GloU	mAP@0.9	0.30	0.25	0.27	0.3	0.43	0.35
6	4	GloU	mAP@0.9	0.25	0.37	0.30	0.25	0.57	0.34

average, the automatically determined area is still an overestimation of the actual area. Even though the overestimation is less than 20 m² on average, which we deem an acceptable error.

To evaluate the performance of the average slope algorithm (see Algorithm 1), the results for 50 RCH locations in the training area were compared with manually determined slope values (see Bonhage, Hirsch, Raab, et al., 2020). The results show that 46% (23 out of 50) of the automatically calculated slope values are within 1 degree of the manually determined slopes, while 94% (47 out of 50) are within 4 degrees. Three locations yield widely different values (>4 degree difference). A manual investigation of these outliers shows that two are located on a ridge surrounded by low-lying areas (e.g., gullies; Figure 7a,b). For the other outlier (Figure 7c), no clear cause for the discrepancy can be found, although it should be noted that the difference between the two slope values for this location is reasonably low (4.1 degrees) in comparison with the other outliers (6.1 and 9.5 degrees, respectively).

4.4 | Application and transferability of the ARCHMAGE workflow

Based on the results (Section 4.3), model 2 was regarded as the most suitable to be used for large scale mapping of RCHs in the entire state of Connecticut. The testing time on approximately 214 500 subtiles took about 40 h on Colab, using a Nvidia T4, P100 or V100 GPU. The results were post-processed (see Sections 3.3 and 4.3), which resulted in 43 197 detections. The distribution of these detections shows a clear concentration in the northwestern part of Connecticut, with up to 194 RCHs per sq. km (see Figure 8). The highest concentration can be found in the western part of the Salisbury district, in Sharon County. To get a first impression of the results, a comparison was made to the distribution of known RCHs, obtained through manually mapping, by Anderson (2019). Note that the latter only shows the presence or absence of RCHs, and our results also show the number

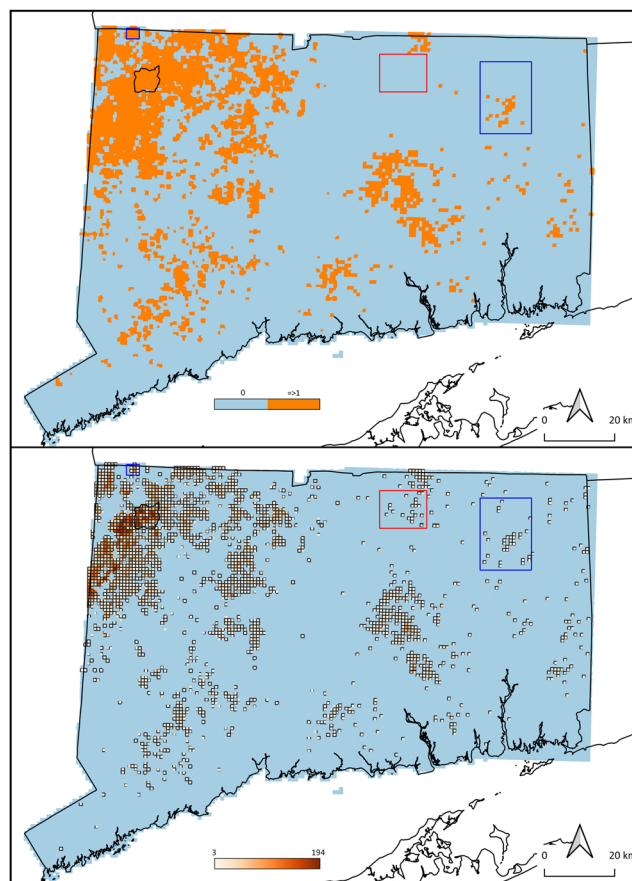


FIGURE 8 Overview of the distribution of relict charcoal hearths (RCHs) (per sq. km) in the state of Connecticut, based on prior research by Anderson (2019; top) and current research (bottom). The top figure shows the presence (orange) or absence (light blue) of RCHs. The bottom figure shows the numbers of RCHs per sq. km (in shades of orange) or absence (light blue). Note that cells with less than 3 RCHs are omitted. The Training area is shown in black, the Test areas in blue. The red outline shows the newly discovered concentration of RCHs near West Stafford. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

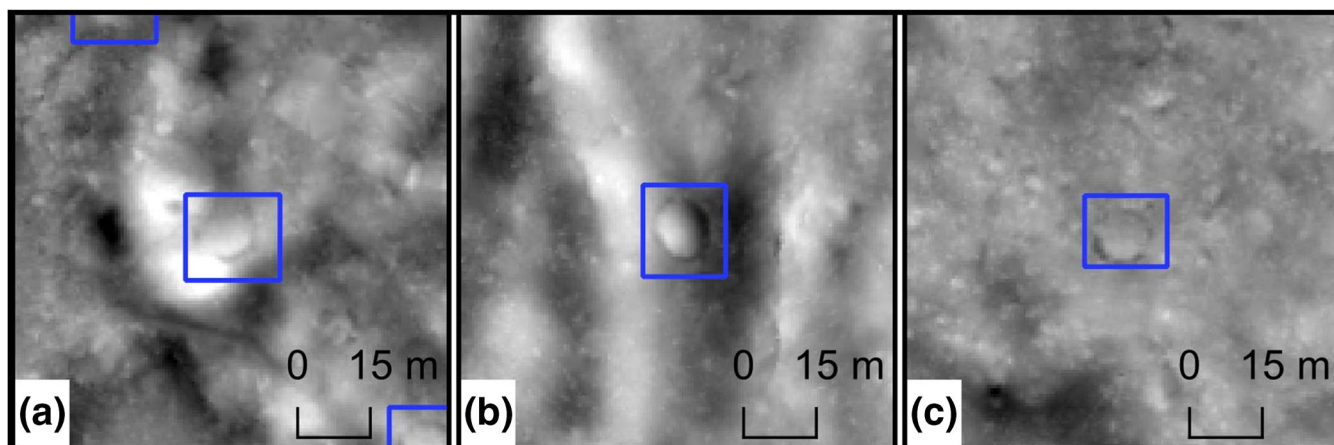


FIGURE 7 Excerpts of LiDAR data, in scale 1:2000, visualized with Simple Local Relief Model (Hesse, 2010), showing the three predicted relict charcoal hearth (RCH) locations in blue outline [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

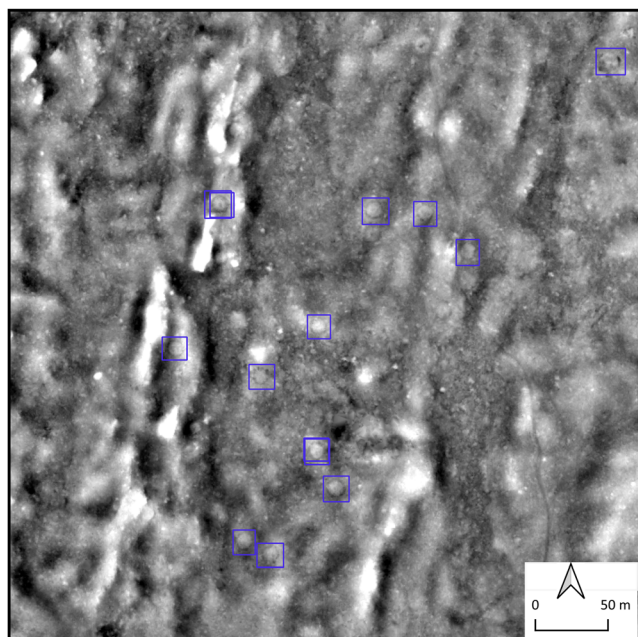


FIGURE 9 Excerpt of LiDAR data, visualized with Simple Local Relief Model (Hesse, 2010), showing part of the newly discovered cluster of relict charcoal hearths (RCHs) in Shenipsit State Forest near West Stafford [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/ar.1889)]

of RCHs per sq. km (Figure 8). A visual inspection reveals a clear overlap between both distributions, with minor differences between the extents of individual concentrations. A major deviation can be observed in the northeastern part of the state (Figure 8, red outline). This was confirmed to be a concentration of previously unknown RCHs in Shenipsit State Forest near the town of West Stafford (Figure 9). This demonstrates that ARCHMAGE is able to successfully detect RCHs over the entire state of Connecticut, even when the terrain and LiDAR quality varies (9).

5 | DISCUSSION

With the ever-increasing reliance of archaeologists on remotely sensed data (Opitz & Herrmann, 2018)—with some even advocating for remote sensing as the primary source for archaeological prospection of large areas (Banaszek et al., 2018)—the usability of automated mapping approaches for large-scale archaeological survey becomes ever more important and necessitates investigation (Lambers et al., 2019). However, up-to-now, the application of automated methods is generally limited to relatively small test areas, although a trend towards covering larger areas can be observed (e.g., Berganzo-Besga et al., 2021; Carter et al., 2021; Suh et al., 2021). Even though, questions concerning the reliability and transferability of these methods for large spatial scales remain (Cowley et al., 2020; Kermit et al., 2018; Verschoof-van der Vaart & Landauer, 2022). Presumably, the application on large areas is still mostly limited by DEM data availability, problems in handling and processing large amounts of spatial

data and, most pressing, the uncertainty caused by changes in LiDAR data quality over larger areas. The usability of a method on areas with different topography, land-use and/or LiDAR data quality is challenging. As shown in this research, variations in LiDAR data quality presumably caused by differences in data acquisition and processing (Opitz & Cowley, 2013) and/or varying land cover conditions (Bonhage et al., 2021; Cowley et al., 2020; Suh et al., 2021) can result in reduced and heterogeneous ground points densities and interpolation errors (Căteanu & Ciubotaru, 2020) that can greatly influence detection performance. As shown in this study (Tables 5 and 6), increasing the spatial scale of the investigated area results in a drastic decrease in performance (see also Trier et al., 2021). Unsurprisingly, the efficiency of automated mapping approaches seems scale dependent (Bonhage et al., 2021). The effect of differences in LiDAR quality has also been noted in other research on automated (Banasiak et al., 2022; Dolejš et al., 2020; Trier & Pilø, 2012) as well as manual mappings (Risbøl et al., 2013) of archaeological objects in LiDAR data. We showed that these variations in LiDAR data quality could be addressed by an additional post-processing step, without decreasing the information value of the LiDAR data (e.g., by increasing the resolution of the DEM although this could be a valid solution as well). The results of the application (Section 4.4) show that with this additional step, ARCHMAGE is able to efficiently and effectively detect RCHs on very large spatial scales.

Another obstacle appears to be the general low density of objects (i.e., the number of objects per sq km) within the test areas. This is a problem inherent to archaeological automated detection, as archaeological objects are more often absent than present in the landscape (Trier et al., 2021), and consequently, the majority of subtiles in a test datasets will not contain an object of interest. Identifying objects in such low-density datasets is a challenging task, not only for automated methods but also for domain experts, with a decrease in object density having a negative influence on performance (Soroush et al., 2020; Verschoof-van der Vaart, 2022). This is, apart from the problem with LiDAR data quality outlined above, the most probably cause of the decrease in performance of our method between Test area 1 (high-density: 29 RCHs/sq. km) and Test area 2 (low-density: 1.75 RCHs/sq. km; see also Table 1).

In addition, the objects within our test datasets are generally small and scarcely distributed, especially compared with objects in more general image datasets, such as the Microsoft COCO dataset (Lin et al., 2014), which are often large and predominately present (see for instance Verschoof-van der Vaart, 2022, fig. 7.2). Deep Learning methods take advantage of the large size of objects to reduce computational cost by downscaling images when they pass through a CNN. However, this also removes small objects, rendering them impossible to detect (Olivier & Verschoof-van der Vaart, 2021). Small and scarcely distributed objects lead to the problem of foreground-background class imbalance in object detection (Oksuz et al., 2019), where one class is over-represented, in this case the background class, whereas the other class (foreground, i.e., the archaeological objects) is under-represented (Luque et al., 2019). This imbalance can have a major impact on the classification and generalization capacity of

CNNs, leading to bias and low performance, as these are generally geared towards detecting large, abundantly present objects.

A final challenge when using automated detection on a large scale is the matter of validating the results, as using a singular methodology as the primary source of information, without verifying the results, seems neither scientifically sound, nor desirable (Bennett et al., 2014). However, manually verifying all detections made on such a large spatial scale runs the risk of moving the professional bottleneck (Smith, 2014) from manual analysis to manual verification, especially when automated approaches with middling performance are used. Therefore, either an approach with very high performance is required or alternative strategies need to be explored to verify the results (Trier et al., 2021; Verschoof-van der Vaart et al., 2020), such as the use of citizen science (Herfort et al., 2019; Lambers et al., 2019). Notwithstanding, the need for detailed verification is very dependent of how the method, and its results are embedded within the wider research framework (Banaszek et al., 2018; Cowley et al., 2020; Lambers et al., 2019; Opitz & Herrmann, 2018). In the case of this research, the aim was to develop a method that can be used in subsequent research on legacy effects on a landscape scale (see Section 1). As such, the validation and correctness of every single detection might not be necessary (see also Soroush et al., 2020). Alternatively, a shift may become necessary from a fixation on individual detections being correct, to the overall patterns being descriptive (Cowley, 2012; Sadr, 2016). While individual detections might not always be correct; the overall patterns in the landscape might be correctly reproduced by the automated method (Gallwey et al., 2019). For instance, the comparison in distribution of RCHs in Connecticut, as shown in Figure 8, can inform us on of the methods 'pattern descriptiveness'. Although, the true value of using automated detection would lie in the discovery of RCHs in areas priory devoid of these objects of interest, that is, deviations from the prevailing pattern.

6 | CONCLUSION

In this research, we presented a novel approach (ARCHMAGE) based on open source data and software to automatically detect RCHs in large-scale LiDAR datasets and simultaneously provide general as well as domain-specific information. The workflow consists of three steps: a preprocessing step that converts the DEM data into input images (subtiles), based on geospatial information about known RCHs; a step concerning the Deep Learning-based object detection; and a post-processing step where the results of the object detection are converted back into geospatial data. In addition, two domain-specific parameters, that is, area and local slope are calculated.

Our study shows that ARCHMAGE is able to effectively detect RCHs in Connecticut, with F1-scores ranging between 0.21 and 0.76, although an additional post-processing was needed to deal with variations in LiDAR quality between different test areas. The results of applying best performing version of the prediction model (with an average F1-score of 0.56) on the entire state of Connecticut show a clear overlap with the known distribution of RCHs in the region, and

new site clusters were found as well, showing the usability of the approach on large-scale datasets even when the terrain and LiDAR quality varies.

Future research will focus on the improvement of the overall workflow and especially the generation of domain specific information. For instance, preprocessing of the LiDAR data to remove interpolation errors would enhance the performance of the workflow. Additional post-processing is envisioned to deal with overlapping bounding boxes, caused by the overlap in the subtiles (which is estimated to occur in approximately 4% of all detections). Another possible angle of research would be to use semantic segmentation (e.g., Guyot et al., 2021) to improve the calculation of domain specific information. Furthermore, the results of the statewide detection and analysis will be used in subsequent research to analyse site distributions as well as morphological and pedological legacy effects of historic charcoal burning on a landscape scale. The additional information of the sites local slope and surface area will be used to calculate specific site volumes and subsequently how much soil substrate was moved and what the sites element stocks are. The datasets and methods created in this research will be made freely available in the near future.

ACKNOWLEDGEMENTS

We are thankful to Jin Wo Suh for facilitating helpful discussions. We also would like to thank Martin Olivier for his assistance on the development of the python scripts and Mark Grimm for proof-reading the manuscript. We are grateful to the reviewers for their constructive comments.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Wouter Verschoof-van der Vaart  <https://orcid.org/0000-0002-1053-3009>

REFERENCES

- Anderson, E. (2019). *Mapping Relict Charcoal Hearths in the Northeast US Using Deep Learning Convolutional Neural Networks and LIDAR data*. University of Connecticut.
- Banasiak, P. Z., Berezowski, P. L., Zapłata, R., Mielcarek, M., Duraj, K., & Stereńczak, K. (2022). Semantic segmentation (U-net) of archaeological features in airborne laser scanning—Example of the Białowieża Forest. *Remote Sensing*, 14(4), 995. <https://doi.org/10.3390/RS14040995>
- Banaszek, Ł., Cowley, D., & Middleton, M. (2018). Towards National Archaeological Mapping. Assessing source data and methodology—A case study from Scotland. *Geosciences*, 8, 272. <https://doi.org/10.3390/geosciences8080272>
- Bennett, R., Cowley, D., & De Laet, V. (2014). The data explosion: Tackling the taboo of automatic feature recognition in airborne survey data. *Antiquity*, 88(341), 896–905. <https://doi.org/10.1017/S0003598X00050766>

- Berganzo-Besga, I., Orenge, H. A., Lumbreras, F., Carrero-Pazos, M., Fonte, J., & Vilas-Estévez, B. (2021). Hybrid MSRM-based deep learning and multitemporal sentinel 2-based machine learning algorithm detects near 10k archaeological tumuli in North-Western Iberia. *Remote Sensing*, 13(20), 4181. <https://doi.org/10.3390/rs13204181>
- Bevan, A. (2015). The data deluge. *Antiquity*, 89(348), 1473–1484. <https://doi.org/10.15184/aqy.2015.102>
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *In arXiv*, 2004, 109.
- Bonhage, A., Hirsch, F., Raab, T., Schneider, A., Raab, A., & Ouimet, W. (2020). Characteristics of small anthropogenic landforms resulting from historical charcoal production in western Connecticut, USA. *Catena*, 195, 104896. <https://doi.org/10.1016/j.catena.2020.104896>
- Bonhage, A., Hirsch, F., Schneider, A., Raab, A., Raab, T., & Donovan, S. (2020). Long term anthropogenic enrichment of soil organic matter stocks in forest soils—Detecting a legacy of historical charcoal production. *Forest Ecology and Management*, 459, 117814. <https://doi.org/10.1016/j.foreco.2019.117814>
- Bonhage, A., Raab, A., Eltaher, M., Raab, T., Breuß, M., & Schneider, A. (2021). 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. *Archaeological Prospection*, 28(2), 177–186. <https://doi.org/10.1002/arp.1806>
- Borchard, N., Ladd, B., Eschemann, S., Hegenberg, D., Mösel, M., & Amelung, W. (2014). Black carbon and soil properties at historical charcoal production sites in Germany. *Geoderma*, 232–234, 236–242. <https://doi.org/10.1016/j.geoderma.2014.05.007>
- Carranza-García, M., Torres-Mateo, J., Lara-Benítez, P., & García-Gutiérrez, J. (2021). On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data. *Remote Sensing*, 13(1), 89. <https://doi.org/10.3390/rs13010089>
- Carrari, E., Ampoorter, E., Bottalico, F., Chirici, G., Coppi, A., Travaglini, D., Verheyen, K., & Selvi, F. (2017). The old charcoal kiln sites in central Italian forest landscapes. *Quaternary International*, 458, 214–223. <https://doi.org/10.1016/j.quaint.2016.10.027>
- Carter, B. P., Blackadar, J. H., & Conner, W. L. A. (2021). When computers dream of charcoal using deep learning, open tools, and open data to identify relict charcoal hearths in and around state game lands in Pennsylvania. *Advances in Archaeological Practice*, 1–15, 257–271. <https://doi.org/10.1017/aap.2021.17>
- Cațeanu, M., & Ciubotaru, A. (2020). Accuracy of ground surface interpolation from airborne laser scanning (ALS) data in dense forest cover. *ISPRS International Journal of Geo-Information*, 9(4), 224. <https://doi.org/10.3390/ijgi9040224>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Cowley, D. (2012). In with the new, out with the old? Auto-extraction for remote sensing archaeology. *Proceedings of SPIE*, 8532, 853206. <https://doi.org/10.1117/12.981758>
- Cowley, D., Banaszek, Ł., Geddes, G., Gannon, A., Middleton, M., & Millican, K. (2020). Making LiGHT work of large area survey? Developing approaches to rapid archaeological mapping and the creation of systematic National-scaled Heritage Data. *Journal of Computer Applications in Archaeology*, 3(1), 109–121. <https://doi.org/10.5334/jcaa.49>
- Davis, D. S. (2019). Object-based image analysis: A review of developments and future directions of automated feature detection in landscape archaeology. *Archaeological Prospection*, 26(2), 155–163. <https://doi.org/10.1002/arp.1730>
- Davis, D. S., Caspari, G., Lipo, C. P., & Sanger, M. C. (2021). Deep learning reveals extent of archaic native American Shell-ring building practices. *Journal of Archaeological Science*, 132, 105433. <https://doi.org/10.1016/j.jas.2021.105433>
- Davis, D. S., & Lundin, J. (2021). Locating charcoal production sites in Sweden using LiDAR, hydrological algorithms, and deep learning. *Remote Sensing*, 13(18), 3680. <https://doi.org/10.3390/rs13183680>
- Deforce, K., Boeren, I., Adriaenssens, S., Bastiaens, J., De Keersmaeker, L., Haneca, K., Tys, D., & Vandekerckhove, K. (2013). Selective woodland exploitation for charcoal production. A detailed analysis of charcoal kiln remains (ca. 1300–1900 AD) from Zoersel (northern Belgium). *Journal of Archaeological Science*, 40(1), 681–689. <https://doi.org/10.1016/j.jas.2012.07.009>
- Dolejš, M., Pacina, J., Veselý, M., & Brétt, D. (2020). Aerial bombing crater identification: Exploitation of precise digital terrain models. *ISPRS International Journal of Geo-Information*, 9(12), 713. <https://doi.org/10.3390/ijgi9120713>
- Donovan, S., Ignatiadis, M., Ouimet, W., Dethier, D., & Hren, M. (2020). Gradients of geochemical change in relic charcoal hearth soils. *North-western Connecticut, USA*, 197, 104991. <https://doi.org/10.1016/j.catena.2020.104991>
- Dupin, A., Sordoillet, D., Fréville, K., Girardclos, O., & Gauthier, E. (2019). The taphonomic characterization of a charcoal production platform. Contribution of an innovative pair of methods: Raman analysis and micromorphology. *Journal of Archaeological Science*, 107, 87–99. <https://doi.org/10.1016/j.jas.2019.05.003>
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 303–338. <https://doi.org/10.1007/s11263-009-0275-4>
- Gallwey, J., Eyre, M., Tonkins, M., & Coggan, J. (2019). Bringing lunar LiDAR back down to earth: Mapping our industrial heritage through deep transfer learning. *Remote Sensing*, 11(17), 1994. <https://doi.org/10.3390/rs11171994>
- Gocel-Chalté, D., Guerold, F., Knapp, H., & Robin, V. (2020). Anthracological analyses of charcoal production sites at a high spatial resolution: The role of topography in the historical distribution of tree taxa in the northern Vosges mountains, France. *Vegetation History and Archaeobotany*, 29(6), 641–655. <https://doi.org/10.1007/S00334-020-00769-Z>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.
- Google. (2022). Google Research Colaboratory. <https://colab.research.google.com/>
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187, 27–48. <https://doi.org/10.1016/j.neucom.2015.09.116>
- Guyot, A., Lennon, M., & Hubert-Moy, L. (2021). Objective comparison of relief visualization techniques with deep CNN for archaeology. *Journal of Archaeological Science: Reports*, 38, 103027. <https://doi.org/10.1016/j.jasrep.2021.103027>
- Hazell, Z., Crosby, V., Oakey, M., & Marshall, P. (2017). Archaeological investigation and charcoal analysis of charcoal burning platforms, Barbon, Cumbria, UK. *Quaternary International*, 458, 178–199. <https://doi.org/10.1016/j.quaint.2017.05.025>
- Herfort, B., Li, H., Fendrich, S., Lautenbach, S., & Zipf, A. (2019). Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning. *Remote Sensing*, 11(15), 1799. <https://doi.org/10.3390/rs11151799>
- Hesse, R. (2010). LiDAR-derived local relief models—A new tool for archaeological prospection. *Archaeological Prospection*, 17(2), 67–72. <https://doi.org/10.1002/arp.374>
- Hirsch, F., Raab, T., Ouimet, W., Dethier, D., Schneider, A., & Raab, A. (2017). Soils on historic charcoal hearths: Terminology and chemical properties. *Soil Science Society of America Journal*, 81(6), 1427–1435. <https://doi.org/10.2136/sssaj2017.02.0067>
- Hirsch, F., Schneider, A., Bonhage, A., Raab, A., Drohan, P. J., & Raab, T. (2020). An initiative for a morphologic-genetic catalog of relict charcoal hearths from Central Europe. *Geoarchaeology*, 1–10, 974–983. <https://doi.org/10.1002/gea.21799>

- Johnson, K. M., & Ouimet, W. B. (2021). Reconstructing historical forest cover and land use dynamics in the northeastern United States using geospatial analysis and airborne LiDAR. *Annals of the American Association of Geographers*, 111(6), 1656–1678. <https://doi.org/10.1080/24694452.2020.1856640>
- Johnson, K. M., Ouimet, W. B., & Raslan, Z. (2015). Geospatial and LiDAR-based analysis of 18th to early 20th century timber harvesting and charcoal production in southern New England. *Geological Society of America Abstracts with Programs*, 47, 65.
- Kazimi, B., Malek, K., Thiemann, F., & Sester, M. (2020). Effectiveness of DTM Derivatives for Object Detection Using Deep Learning. In *CHNT 25, 2020—Abstracts* (pp. 1–4). Propylaeum.
- Kazimi, B., Thiemann, F., & Sester, M. (2019). Semantic segmentation of manmade landscape structures in digital terrain models. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(2), 87–94. <https://doi.org/10.5194/isprs-annals-IV-2-W7-87-2019>
- Kermit, M., Reksten, J. H., & Trier, Ø. D. (2018). Towards a national infrastructure for semi-automatic mapping of cultural heritage in Norway. In M. Matsumoto & E. Uleberg (Eds.). (Red.) *Oceans of data. Proceedings of the 44th conference on computer applications and quantitative methods in archaeology* (pp. 159–172). Archaeopress.
- Kokalj, Ž., & Hesse, R. (2017). *Airborne laser scanning raster data visualization: A guide to good practice*. Založba ZRC. <https://doi.org/10.3986/9789612549848>
- Lambers, K., Verschoof-van der Vaart, W. B., & Bourgeois, Q. P. J. (2019). Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection. *Remote Sensing*, 11(7), 794. <https://doi.org/10.3390/rs11070794>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science*, 8693, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
- Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216–231. <https://doi.org/10.1016/j.patcog.2019.02.023>
- Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2019). Imbalance problems in object detection: A review. In *arXiv*. <https://arxiv.org/pdf/1909.00169>
- Oliveira, C., Aravecchia, S., Pradalier, C., Robin, V., & Devin, S. (2021). The use of remote sensing tools for accurate charcoal kilns' inventory and distribution analysis: Comparative assessment and prospective. *International Journal of Applied Earth Observation and Geoinformation*, 105, 102641. <https://doi.org/10.1016/j.jag.2021.102641>
- Olivier, M., & Verschoof-van der Vaart, W. B. (2021). Implementing state-of-the-art deep learning approaches for archaeological object detection in remotely-sensed data: The results of cross-domain collaboration. *Journal of Computer Applications in Archaeology*, 4(1), 274–289. <https://doi.org/10.5334/jcaa.78>
- Opitz, R., & Cowley, D. (2013). *Interpreting Archaeological Topography. Airborne Laser Scanning, 3D Data and Ground Observation*. Oxbow Books.
- Opitz, R., & Herrmann, J. (2018). Recent trends and long-standing problems in archaeological remote sensing. *Journal of Computer Applications in Archaeology*, 1(1), 19–41. <https://doi.org/10.5334/JCAA.11>
- Orengo, H. A., Conesa, F. C., Garcia-Molsosa, A., Lobo, A., Green, A. S., Madella, M., & Petrie, C. A. (2020). Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *PNAS*, 117(31), 18240–18250. <https://doi.org/10.1073/pnas.2005583117>
- Ouimet, W. (2019). Northeastern US relict charcoal hearth mapper. <https://connecticut.maps.arcgis.com/apps/webappviewer/index.html?id=102f6831a12843878ea8081aec41029d>
- QGIS Development Team. (2017). *QGIS Geographic Information System* (3.4). Open Source Geospatial Foundation Project. <http://qgis.org>
- Quintus, S., Day, S. S., & Smith, N. J. (2017). The efficacy and analytical importance of manual feature extraction using lidar datasets. *Advances in Archaeological Practice*, 5(4), 351–364. <https://doi.org/10.1017/aap.2017.13>
- Raab, A., Bonhage, A., Schneider, A., Raab, T., Rösler, H., Heußner, K. U., & Hirsch, F. (2019). Spatial distribution of relict charcoal hearths in the former royal forest district Tauer (SE Brandenburg, Germany). *Quaternary International*, 511, 153–165. <https://doi.org/10.1016/j.quaint.2017.07.022>
- Raab, T., Hirsch, F., Ouimet, W., Johnson, K. M., Dethier, D., & Raab, A. (2017). Architecture of relict charcoal hearths in northwestern Connecticut, USA. *Geoarchaeology*, 32(4), 502–510. <https://doi.org/10.1002/geo.21614>
- Raab, T., Raab, A., Bonhage, A., Schneider, A., Hirsch, F., Birkhofer, K., Drohan, P., Wilmking, M., Kreyling, J., Malik, I., Wistuba, M., van der Maaten, E., van der Maaten-Theunissen, M., & Urich, T. (2022). Do small landforms have large effects? A review on the legacies of pre-industrial charcoal burning. *Geomorphology*, 413, 108332. <https://doi.org/10.1016/j.geomorph.2022.108332>
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *arXiv preprint*, 1506.02640.
- Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- Rezatofghi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 658–666. <https://doi.org/10.1109/CVPR.2019.00075>
- Risbøl, O., Bollandsås, O. M., Nesbakken, A., Ørka, H. O., Naeset, E., & Gobakken, T. (2013). Interpreting cultural remains in airborne laser scanning generated digital terrain models: Effects of size and shape on detection success rates. *Journal of Archaeological Science*, 40(12), 4688–4700. <https://doi.org/10.1016/j.jas.2013.07.002>
- Rutkiewicz, P., Malik, I., Wistuba, M., & Osika, A. (2019). High concentration of charcoal hearth remains as legacy of historical ferrous metallurgy in southern Poland. *Quaternary International*, 512, 133–143. <https://doi.org/10.1016/j.quaint.2019.04.015>
- Sadr, K. (2016). The impact of coder reliability on reconstructing archaeological settlement patterns from satellite imagery: A case study from South Africa. *Archaeological Prospection*, 23, 45–54. <https://doi.org/10.1002/arp.1515>
- Sammut, C., & Webb, G. I. (2010). *Encyclopaedia of machine learning*. Springer. <https://doi.org/10.1007/978-0-387-30164-8>
- Schmidt, S. C., & Marwick, B. (2020). Tool-driven revolutions in archaeological science. *Journal of Computer Applications in Archaeology*, 3(1), 18–32. <https://doi.org/10.5334/jcaa.29>
- Schneider, A., Bonhage, A., Hirsch, F., Raab, A., & Raab, T. (2022). Hot spots and hot zones of soil organic matter in forests as a legacy of historical charcoal production. *Forest Ecology and Management*, 504, 119846. <https://doi.org/10.1016/j.foreco.2021.119846>
- Schneider, A., Bonhage, A., Raab, A., Hirsch, F., & Raab, T. (2020). Large-scale mapping of anthropogenic relief features—Legacies of past forest use in two historical charcoal production areas in Germany. *Geoarchaeology*, 35(4), 545–561. <https://doi.org/10.1002/GEA.21782>
- Schneider, A., Hirsch, F., Raab, A., & Raab, T. (2019). Temperature regime of a charcoal-enriched land use legacy soil. *Soil Science Society of America Journal*, 83(3), 565–574. <https://doi.org/10.2136/sssaj2018.12.0483>
- Schneider, A., Takla, M., Nicolay, A., Raab, A., & Raab, T. (2015). A template-matching approach combining morphometric variables for automated mapping of charcoal kiln sites. *Archaeological Prospection*, 22(1), 45–62. <https://doi.org/10.1002/arp.1497>

- Smith, M. L. (2014). Citizen science in archaeology. *American Antiquity*, 79(4), 749–762. <https://doi.org/10.7183/0002-7316.79.4.749>
- Soroush, M., Mehrtash, A., Khazraee, E., & Ur, J. A. (2020). Deep learning in archaeological remote sensing: Automated Qanat detection in the Kurdistan region of Iraq. *Remote Sensing*, 12(3), 500. <https://doi.org/10.3390/rs12030500>
- Suh, J. W., Anderson, E., Ouimet, W., Johnson, K. M., & Witharana, C. (2021). Mapping relict charcoal hearths in the northeast US using deep learning convolutional neural networks and LIDAR data. *Remote Sensing*, 13, 4630. <https://doi.org/10.3390/rs13224630>
- Tolksdorf, J. F., Elburg, R., Schröder, F., Knapp, H., Herbig, C., Westphal, T., Schneider, B., Fülling, A., & Hemker, C. (2015). Forest exploitation for charcoal production and timber since the 12th century in an intact medieval mining site in the Niederpöbel Valley (Erzgebirge, Eastern Germany). *Journal of Archaeological Science: Reports*, 4, 487–500. <https://doi.org/10.1016/J.JASREP.2015.10.018>
- Trier, Ø. D., Cowley, D., & 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*, 26(2), 165–175. <https://doi.org/10.1002/arp.1731>
- Trier, Ø. D., & Pilø, L. H. (2012). Automatic detection of pit structures in airborne laser scanning data. *Archaeological Prospection*, 19(2), 103–121. <https://doi.org/10.1002/arp.1421>
- Trier, Ø. D., Reksten, J. H., & Løseth, K. (2021). Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. *International Journal of Applied Earth Observation and Geoinformation*, 95, 102241. <https://doi.org/10.1016/j.jag.2020.102241>
- Trier, Ø. D., Salberg, A.-B., & Pilø, L. H. (2018). Semi automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In M. Matsumoto & E. Uleberg (Eds.). (Red.)CAA 2016: *Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology* (pp. 219–231). Archaeopress.
- Verschoof-van der Vaart, W. B. (2022). *Learning to Look at LiDAR. Combining CNN-Based Object Detection and GIS for Archaeological Prospection in Remotely-Sensed Data*. Leiden University.
- Verschoof-van der Vaart, W. B., & 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*, 2(1), 31–40. <https://doi.org/10.5334/jcaa.32>
- Verschoof-van der Vaart, W. B., Lambers, K., Kowalczyk, W., & 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*, 9(5), 293. <https://doi.org/10.3390/ijgi9050293>
- Verschoof-van der Vaart, W. B., & Landauer, J. (2022). Testing the transferability of CarcassonNet. A Case study to detect hollow roads in Germany and Slovenia. *Proceedings of the 25th Conference on Cultural Heritage and New Technologies, CHNT25*.
- Witharana, C., Ouimet, W. B., & Johnson, K. M. (2018). Using LiDAR and GEOBIA for automated extraction of eighteenth–late nineteenth century relict charcoal hearths in southern New England. *GIScience and Remote Sensing*, 55(2), 183–204. <https://doi.org/10.1080/15481603.2018.1431356>
- Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., & Zuo, W. (2021). Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Transactions on Cybernetics*, 1–13, 8574–8586. <https://doi.org/10.1109/TCYB.2021.3095305>
- Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5, 2–8. <https://doi.org/10.1016/j.bdr.2015.12.001>

How to cite this article: der Vaart, W. V., Bonhage, A., Schneider, A., Ouimet, W., & Raab, T. (2022). Automated large-scale mapping and analysis of relict charcoal hearths in Connecticut (USA) using a Deep Learning YOLOv4 framework. *Archaeological Prospection*, 1–16. <https://doi.org/10.1002/arp.1889>