Research Thesis: Declaration of Authorship

Print name: Dr Iris Kramer

Title of thesis: Machine Learning for the Detection of Archaeological Sites from Remote Sensor Data

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. {Delete as appropriate} None of this work has been published before submission;

Signature:

Date: 14-08-2021

UNIVERSITY OF SOUTHAMPTON
Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

Machine Learning for the Detection of Archaeological Sites from Remote Sensor Data

by

Iris Kramer

A thesis submitted for the degree of
Doctor of Philosophy

January 2021

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

Machine Learning for the Detection of Archaeological Sites from Remote Sensor Data

by Iris Kramer

Deep learning for automated detection of archaeological sites (objects) on remote sensing data is a highly novel field. The key challenge of this field is in the inherent nature of the objects; they occur in small numbers, are sparsely located and feature a unique pattern on the different remote sensing data modalities. To this extent we identify three main contributions, (1) to include multi-sensor data, (2) to optimise Convolutional Neural Networks (CNNs) for small datasets and, (3) to optimise detection of the sparsely located objects. Our results demonstrate that deep learning can be successfully applied to detect archaeological sites on each of the individual remote sensing images, that our efforts to optimise CNNs for small datasets are successful, and that we have discovered new sites that were missed in a manual data analysis and field survey. We have optimised a workflow for the detection of new archaeological sites. We also share the first large-scale publicly available dataset archaeological image classification and object detection along with benchmarks of the most promising models that we applied in this thesis.

Contents

List of Figures

List of Tables

Acronyms

ALS	Airborne Laser Scanning
CNN	Convolutional Neural Network
DEM	Digital Elevation Model
DSM	Digital Surface Model
DTM	Digital Terrain Model
Faster R-CNN	Faster Region-CNN
FPN	Feature Pyramid Network
GIS	Geographical Information System
HER	Historic Environment Record
HES	Historic Environment Scotland
HLA	Historic Land-use Assessment
IoU	Intersection over Union
LiDAR	Light Detection and Ranging
MaP	Mean average Precision
NIR	Near InfraRed
NMS	Non-Maximum Suppression
NRHE	National Record of the Historic Environment
OBIA	Object Based Image Analysis
R-CNN	region-CNN
RELU	Rectified Linear Unit

RGB	Red-Green-Blue
RGBN	Red-Green-Blue-Near infrared
SLRM	Simplified Local Relief Model
SOTA	State-Of-The-Art
SSD	Single Shot Detection
SVM	Support Vector Machine
VAT	Visualization for Archaeological Topography
YOLO	You Only Look Once

Acknowledgements

First and foremost I am extremely grateful to my supervisor Dr. Jonathon Hare for going out on a limb and believing that my project was worthwhile and that I was able to do it. I benefited a lot from his support, deep insights, knowledge and constant feedback throughout the PhD. I'm also grateful to the Ordnance Survey for funding my PhD and for my OS supervisor Dr. Isabel Sargent for the great discussions on deep learning in the geography domain. I am also grateful to the funding I received from EPSRC through the University of Southampton.

I also want to thank all the members of the VLC group for their support and for bringing a lot of fun to the PhD journey. I'm also very thankful for all the amazing friends that I made during my time in Southampton.

I would like to thank my parents, sister and all of my family for their encouragement and support all through my studies. Finally I would like to thank James Byers, for his unwavering support and belief in me and also for helping me enormously, especially with the final formatting and proof reading.

Chapter 1

Introduction

An essential aspect of archaeology is the protection of sites from looters, extensive agriculture, and erosion. Under the constant threat of destruction, it is of utmost importance that sites are located so that they can be monitored and protected. This is mostly done by archaeologists on the ground or through manual analysis of remote sensing data such as aerial images or Light Detection and Ranging (LiDAR) derived elevation models. This task is time consuming and requires highly specialised and experienced people and would thus immensely benefit from automation.

The recent explosion in the availability of high resolution imagery and in the variety of new remote sensors underscores the need for automated methods. The increased resolution has the improved the detail that can be recorded but has also increased the amount of time that is spend per km^2. Likewise the number of sensors that are available for the detection of archaeology has increased the and brought a realisation that there is more data than is humanly possible to assess Bennett et al. (2014).

Automation of archaeological objects in remote sensing data is highly challenging as some of the most 'overwritten' signatures of the landscape need to be extracted from petabytes of imagery. Despite previous attempts, researchers have not been able to develop a method that is able to generalize well across archaeological objects, geographical locations and remote sensing data sets. In order to generate a satisfactory method, it is argued in this thesis that only a machine learning approach can reach the desired generality. Even though traditional machine learning required extensive feature engineering, recent developments have moved towards automated feature learning. Deep learning using Convolutional Neural Networks (CNNs) have been particularly successful in this space and are therefore the main focus of this research.

In machine learning, the goal is to label image pixels into one or several classes. Image classification can be binary where images are given a single object class or groups of pixels within an image can be classified into different classes with semantic segmentation or object detection.

The aim of this thesis is to discover approaches utilising deep learning that can be applied to the detection of archaeology. In chapter 2 we review the challenges that are be prevalent in the detection of archaeology on different remote sensing resources and research and implement solutions. We have highlighted three main challenges:

- **Small datasets:** One of the essential requirements for deep learning is a sufficiently large training dataset of example objects. Our main concern in the domain of archaeological object detection is that there are often only a few objects known of a specific type, and these objects are sparsely distributed throughout the landscape. Our initial focus will therefore be on optimising different aspects of deep learning for small datasets, and part of this is to include domain knowledge.
- **Non-conventional data format:** Archaeological sites can be detected using different types of remote sensing data including multi-spectral aerial imagery and LiDAR derived elevation models. The information captured in the signal of these individual sensors could be leveraged with an integrated deep learning approach using all of the data modalities.
- **Deep learning architectures:** In our experiments, we have specifically chosen networks and parameters, mainly on regularisation and transfer learning, which are known to work well with small datasets. We also compare the performance of the networks when they are trained on individual remote sensing images, and with those trained on images of stacked multi-sensor data.

In chapter 3 we present the results from our initial experiments to alleviate the highlighted challenges. We focus on the New Forest National Park and the detection of barrows which are well known funerary sites that are found across the world. Our datasets include LiDAR and multi-spectral aerial imagery.

In chapter 4 we further address the challenges that we were not able to overcome in the previous chapter. This case study is focussed on the Isle of Arran in Scotland, uses only LiDAR data and looks at round houses, shieling huts and small cairns. In the case study we use datasets from Historic Environment Scotland (HES) and we incorporate feedback on our results to optimise the approach.

In chapter 5 we discuss best practise that we have gathered from literature and our own experience. We discuss the most important elements of a successful workflow ranging from the creation of a dataset for deep learning to the selection of CNNs and evaluation metrics that suit a specific dataset. We finally discuss the most promising technology innovations that we hope will be used in future research projects for the detection of archaeology on remote sensing data.

The main contribution of this thesis is a systematic workflow that encourages a deep understanding of the dataset and applied methods. It also addresses the challenges that we have identified. The framework starts with optimising an image classification methods and uses the optimised parameters in an object detection approach. In addition, we release a benchmark dataset and share our code to encourage comparison and improvements in new approaches.

The additional goal was to encourage the uptake of automation in archaeology and increase a positive outlook to new approaches. This PhD was designed to follow up the MSc research from Kramer (2015). Since the start of this PhD, the automation discussion has positively changed, highlighting particularly the use of machine learning for automated detection of archaeological sites. In part this shift in mindset has been strengthened by the organisation of events by the community surrounding automation. The core of the discussion has taken place at the largest computer conference in our field, the international conference for computer applications and quantitative methods in archaeology (CAA). For example, at the recurring session run by Arianna Traviglia and Dave Cowley on automation in remote sensing: “Computer vision vs human perception in remote sensing image analysis: time to move on.” (Traviglia and Lambers (2016) at CAA-2016), “Automation is here to stay! The hitch-hiker’s guide to automated object detection and image processing in remote sensing” (Traviglia and Lambers (2017) at CAA-2017) and “Setting the automation agenda for remote sensing: learning to see through a computer?” (Traviglia and Lambers (2018) at CAA-2018). In contribution to this discussion I have presented various papers (Kramer (2016) (for which I won best paper award), Kramer (2017), Kramer (2018b)) and organised a workshop on “The basics of deep learning for archaeological site detection on remote sensor data” at CAA-2018 (Kramer (2018a)) teaching participants the basics of the approach presented in chapter 4. At CAA-2019, in contribution to the wider discussion on applicability of AI to archaeological applications I co-organised a session called “Challenges and opportunities of machine learning in archaeological research” together with Wouter Verschoof-Van Der Vaart and Alex Brandsen (Kramer et al., 2019). Based on its 2019 success we will organise this session “Machine learning in archaeological research; challenges and opportunities” at CAA-2021 with Wouter Verschoof-van der Vaart, Alex Brandsen, Hector Orengo, Arnau Garcia-Molsosa and Francesc Conesa. Aside from the CAA, several other events have taken place including a workshop “Tracing the Past: Combining Citizen Science and Data Science” organised by Karsten Lambers and co-hosted by Dave Cowley held in July 2018 at the Lorentz Center, Leiden University. In November 2019, I also co-organised a two day international conference and workshop for Machine Learning in Archaeology in Rome together with Christopher Stewart (European Space Agency) and Peter B. Campbell (British School at Rome) (Campbell et al., 2019). It is in large part thanks to the discussions held at these various meetings that the automation discussion has shifted away from discussing the potential of automation towards researchers actively working together with their computer science departments to apply state-of-the-art deep learning approaches to archaeological case studies.

Chapter 2

Aerial Archaeology and Automation

In this chapter we discuss the history of: how new archaeological sites are manually detected with the help of airborne techniques (section 2.2); what limits manual detection (section 2.3); how automated methods have been applied in archaeology (section 2.4); and, how deep learning can improve current methods (section 2.5). We critically review the key issues need to be addressed to apply deep learning to archaeological site detection and finish with what we aim to be an inspirational discussion of relevant fields that have similar issues when applying deep learning (section 2.6).

2.1 A Brief History of Aerial Archaeology

Past human activity has left its fingerprint on the landscape. This impression is sometimes observed as standing remains like Stonehenge or Carnac but is most often buried underground. Traditionally excavation is the main approach to studying such remains, but excavations alone do not provide insights into the context of ancient landscapes. Some archaeological features cannot be seen, or fully appreciated, without an aerial perspective (Crawford, 1923). The rise of aerial archaeology brought about the study of landscape archaeology, in which archaeologists disentangle the hierarchies of land use in different periods and find patterns that were previously unknown. Archaeology was one of the first disciplines to use remote sensing in scientific investigations (Barber, 2011). Aerial archaeology has allowed archaeologists to discover "about what lies beyond the site, or the edge of the excavation" (Johnson, 2007). Filling the gaps between the sites providing valuable information about human exploitation of the environment. Country-wide research has especially added to such insights. In the UK, aerial archaeology was pioneered by O. Crawford who worked for the national mapping agency, the Ordnance Survey, where he became its first archaeology officer in 1920. Today Historic England holds the national archive of aerial archaeology and continually adds to it with their National Mapping Programme.

2.2 Aerial Archaeology

Buried archaeological features (hereafter called objects to avoid confusion with the computer science use of feature) in the landscape can be recognised as slight elevation differences (earthwork or shallow buried walls) or through discolourations of the soil or vegetation revealing different moisture content or growth habits to their surrounding, undisturbed, soil. In this section we will discuss two different sensors and how we interpret these signs of archaeology.

2.2.1 Aerial Photography

The visual appearance of preserved archaeological sites can be captured from aerial photographs. The imagery is taken by sensors that measure visible light, this includes sensors that measure other kinds of electromagnetic radiation, such as infrared and hyperspectral sensors. Mainly these sites are apparent through the textures or shadows of earthworks, soil colouring and the difference in stress and enhanced crop growth over buried archaeological remains (Figure 2.1). The appearance of the site is highly dependent on the environmental factors such geology, crop type, soil moisture, time of year and even time of day. These extraordinary and sometimes serendipitous circumstances require aerial archaeology experts to have a deep understanding of the local circumstances. It also means that they need to fly in very specific time frames or, when looking at general purpose aerial/satellite imagery, they need to reflect on the environmental conditions at the time the image was captured. Because of these specific circumstances an image from a single time frame rarely tells the full story and experts try to look for images for the same field from multiple years and at multiple times of the year, mainly to account for crop-rotation. However, this practice is very costly and only really undertaken in commercial archaeology where high accuracy over a single field is necessary and worth the extra investment.

The environmental factors that reveal archaeology, such as soil moisture and crop type and crop stress, have a disproportionate effect on the spectrum beyond visible light. Infrared and hyperspectral sensors can detect subtle vegetation characteristics (e.g. stressed versus healthy plants) and soil properties (e.g. mineral composition) to a much higher extent than any standard photographic method (Traviglia et al., 2006).

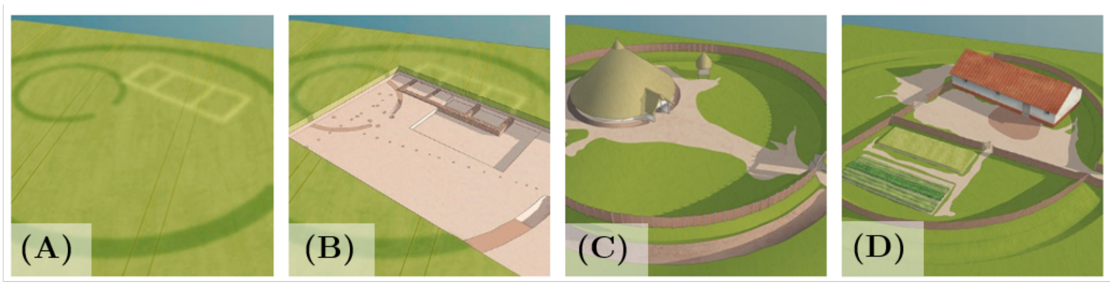


Figure 2.1: A schematic timeline of images demonstrating how archaeology can result in cropmarks. Reproduced from HistoricEngland (2018)

2.2.2 Airborne LiDAR

Different from aerial imagery LiDAR or Airborne Laser Scanning (ALS) sensors do not measure electromagnetic radiation but instead measures the distance to objects from the sensor. Rather than two dimensional data the LiDAR sensor captures 3D (XYZ) coordinates. During an airborne LiDAR survey the land surface is scanned from an aircraft by a high frequency pulsed Infrared laser beam which records the locations of each ground/surface hit and calculates its elevation based on the time it takes for a pulse to return to the transmitter (Hyyppa et al., 2009). Every laser beam may be returned multiple times and could, depending on the track towards the surface, return on several branches of a tree before it returns the terrain elevation (Figure 2.2). When using a full-waveform recording scanner with a high point density this can pierce through dense forest canopy and reveal hidden archaeological landscapes (Doneus et al., 2008; Sittler, 2004).

The resulting data, also called a point cloud, cannot be read by humans without further processing. The raw XYZ point data is generally interpolated to generate a rasterized Digital Elevation Model (DEM) from all the returned points or Digital Terrain Model (DTM) from only the last return. In this process some potential key information could be lost or image artefacts can be created which will perpetuate in further processing and analysis. The resulting greyscale DTM will reveal the general terrain trends. This image stretches over large height difference and displays many shades of grey, often in 16 bit images to retain the terrain detail. However, humans can only distinguish between about 30 shades of grey which means they can interpret the difference between an area at sea level and a hilltop but not the local bomb craters in both areas. In archaeology the immediate neighbouring pixels of a bomb crater are more important than distant pixels. Further image processing where meaningful pixels are grouped together by image transformations including smoothing, sharpening, contrasting, stretching is required to highlight local archaeological earthworks (Figure 2.3)). There are several such visualisations, also called ‘derivatives’, developed for archaeology. For example, Local Relief Model (LRM) emphasises small-scale features by extracting local positive and negative relief variations (Hesse, 2010). In this process a low-pass filter is applied to the DTM to approximate the large-scale landforms. The neighbourhood size of the low pass filter determines the scale of features that will be visible in the LRM. Yet, archaeological objects can be of varying size and height which means that this process can result in the removal of some archaeological earthworks (Doneus, 2013). Therefore, it is advised to use multiple derivatives which requires more interpretation time from the expert. To improve the speed and accuracy of manual analysis, Kokalj and Somrak (2019) propose the combination or fusion of multiple visualisations through different blend modes (e.g. overlay, multiply).

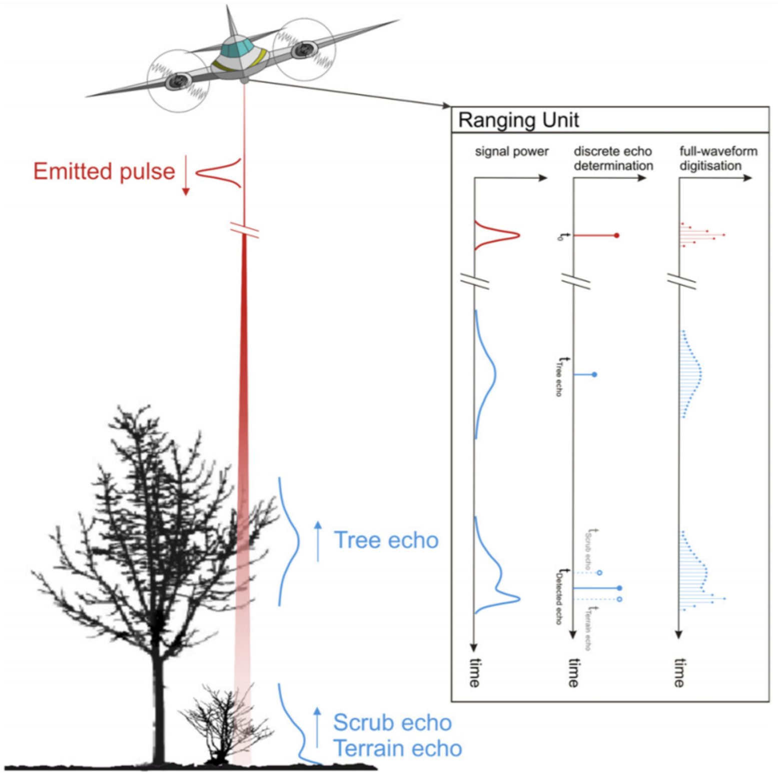


Figure 2.2: Process of a LiDAR survey capturing elevation data (reproduced from Doneus et al. (2008)

2.3 Need for Automation

Whereas in the 1990s aerial archaeologists had just a few aerial images in archives to work with, in 2020 there are petabytes of satellite imagery available on top of yearly national coverage of aerial imagery and frequent updates to LiDAR archives. Somewhere in this data, all of the archaeological sites are captured; we just have not yet found a way to extract it. The increase in data has encouraged many national and county heritage agencies to launch projects for systematic large scale mapping. Historic England undertook the National Mapping Programme (NMP) over 20 years and they were able to cover approximately 1 km^2 per person per day, looking mainly at aerial photography (Bewley, 2003). In Baden-Württemberg (Germany) a single expert was appointed to

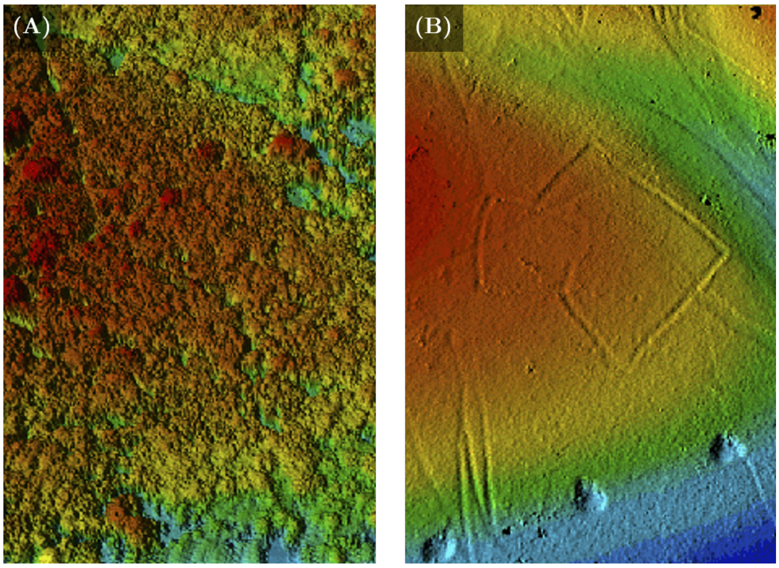


Figure 2.3: Interpolated point data in the area of Savernake, Wiltshire, UK, showing (A) the forest canopy and (B) revealed elevation differences on the forest floor below the canopy. Reproduced from HistoricEngland (2018b))

analyse only LiDAR data and was able to cover 35,000 km in six years (Hesse, 2013). Historic Environment Scotland has also experimented with efficient national mapping through their Rapid Archaeology Mapping Programme (RAMP) in which they used the Isle of Arran as a representative case study to later extrapolate to the rest of the country (Banaszek et al., 2018). They were able to analyse 30 km^2 per person per day on average using only LiDAR data. Whereas (Somrak et al., 2020) noted that it took 8 man-months to annotate 130 km^2 of Mayan archaeology using LiDAR derived visualisation.

Despite significant efforts to speed up and systematise manual analysis there will always be more data then a manual assessment can economically look at. In many scientific disciplines this realisation was made early and has fuelled research into automation approaches. In archaeology, automation has long been a controversial issue as clearly highlighted by Parcak (2009) "Why does there even need to be an automated process for satellite archaeology?". There is a fear that human experts would be replaced with computer vision in archaeological prospection (Casana, 2014). Most importantly automation is meant to become another tool, and not a replacement, for archaeologists. It can be used to quickly create a baseline dataset of the features of interest over large geographical areas, especially for studying high-density off-site features with relatively uniform appearance (Soroush et al., 2020). The baseline of common, easily detectable sites can further be used to infer the existence and preservation of more unique sites that are difficult to detect in specific areas which can further be used in policy making or grant application to research an area. This is especially important for large scale mapping and monitoring of ancient landscapes that are inaccessible for fieldwork, threatened, or permanently destroyed. Globally there are hundreds of thousands, if not millions, of undiscovered ancient sites. If these site locations are unknown they remain at risk from development, warfare, intensive agriculture and climate change. As a profession, we are responsible for saving the archaeological record as best as we possibly can (Institute for Archaeologists' code of conduct). By accelerating the detection rate with automation we will be able to save more of our invaluable human past.

Aside from the speed in site detection, there are several arguments that could favour computational over human interpretation. Humans are biased by their previous experience and their interest. A compelling research on bias within aerial archaeologists has been presented by Cowley (2016). Cowley mainly critiques the traditional observer directed approach where aerial archaeologists observe the landscape from an aircraft around the time crop marking should be visible. For example, in Scotland aerial reconnaissance causes a disparity between known sites in highly fertile arable land and areas where soil types such as heavy clay and poor draining prevents obvious crop marking. Aerial archaeologists expected to find less sites on the poor soils and chose to fly over the attractive areas with beautiful crop marks and re-recording what was already known. He also reflects on experience and notes that an expert will not see what they have not been trained to see; observation doesn't always lead to an interpretation. Another important observation in archaeology is that many of the same sites have different terminology and sometimes even interpretation of their ancient use which causes confusion and wrong interpretation of patterns. By training a machine with data gathered by a mix of experts a more averaged expert will arise. The confusion between the same sites that have different terminology will persist but the critical expert working with the AI should be alerted because of the lower AI performance on a group of similar sites and they can correct such irregularities. Finally there is an argument to be made on the reproducibility of the task when using machine learning. Humans are black boxes and it is difficult to understand what an expert knows and what not. Machines can be black boxes as well, but because they will represent the average of the data that they were trained on, they will be an average of different experts.

Aerial archaeology is also highly subjective to the data choices that the expert makes. Humans can only see the visual spectrum of red, green and blue channels. Multi and hyper spectral channels are not easily representable in a human readable form. Today, these bands are visualised in greyscale, or by swapping it with a visual band, or by transformations of different bands combinations into so called vegetation indexes. The choice of visualisation biases the process and undoubtedly some integral information is lost. Automation approaches could reduce this bias as computers can effectively infinitely stack channels and derive the key information from each channel. When using LiDAR data in an archaeological context the expert has to choose the parameters for the initial point cloud processing, DTM created and the choice of visualisation that highlights the archaeology. This process creates manual bias and also means that inevitably some information is lost. This could be alleviated with automation because computers can

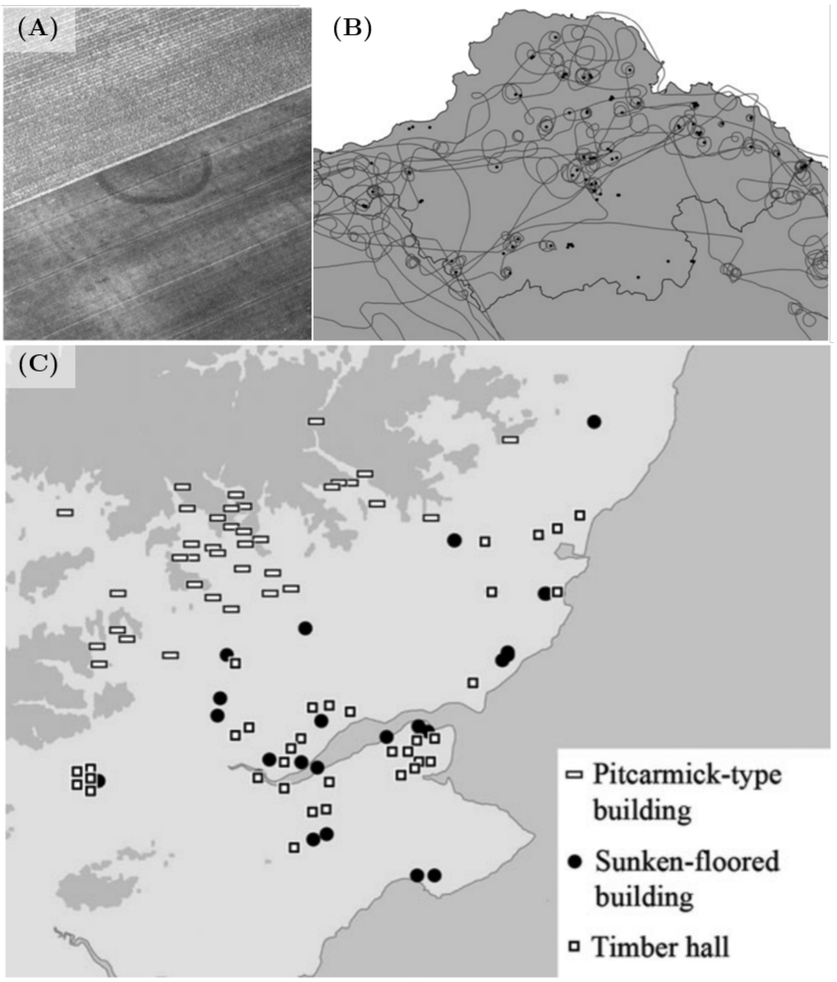


Figure 2.4: Images from different domains that demonstrate the need for a systematic, automated, archaeological detection system on multi-spectral aerial observation data.(A) RGB aerial observation of round archaeological structure at different levels of visibility under different crops. (B) Aerial photography patterns of flying. (C) Map of similar sights using different terminology. Images reproduced from Cowley (2016)

process the single channel numeric grids of height data directly. It is even possible to apply machine learning on the raw point cloud data, as will be discussed in section.

Where LiDAR data is a highly reliable source for detecting earthworks in natural terrain, in agricultural terrain these patterns are ploughed out which means that aerial/satellite imagery is the only usable resource. As previously discussed in subsection 2.2.1, aerial archaeologists often describe their finds on aerial photography as serendipitous. Most sites are found on aerial photography during a summer dry spell when the crop is ripening. Yet there are many other times of year that archaeology can be found in the early/late crops as well as soil marks after ploughing and even with shadow marks in the snow. All of this knowledge reveals that the quantity of time frames matter as well as quality for their selection. It currently takes a very experienced expert to understand the relationship between local geography and archaeological sites. Yet expert knowledge can be used to train machine learning algorithms to find the right time frames and to do it tirelessly for many years.

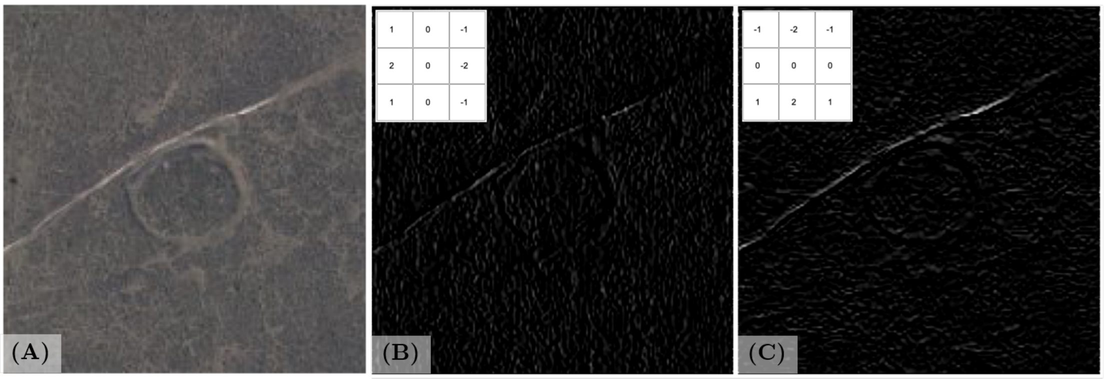
All the information from the different data sources and different time frames can be infinitely stacked by a machine which can then calculate the cumulative accuracy of all the signals and extract the most insightful data points to feedback to the human expert for verification. This isn't humanly possible, but with machine learning it is.

2.4 The Current State of Automation Research

We have already discussed different approaches for image processing that can be applied to visually enhance aerial imagery and LiDAR data. The next level of automation are knowledge based algorithms that can be used for object detection and include explicit feature selection. In archaeology we have seen several specialised algorithms for shape detection (Zingman et al., 2016), template matching (de Boer, 2007; Trier and Pilo, 2012; Trier et al., 2015), and rule based pixel or Object Based Image Analysis (OBIA) (de Laet et al., 2007). These feature engineering techniques rely heavily on the selection of image processing techniques. Kramer (2015) in her Masters thesis reviewed the history of such techniques and their applicability to archaeology. She also created an approach to adaptive template matching and OBIA for round barrow detection using the Slope visualisation of LiDAR data. She concluded that the drawback from knowledge based approaches is that they aren't transferable between sensors and objects, and don't scale geographically. The lack of scalability in combination with high false positive rates might also be the reason why automated methods are not generally (re-) used or picked up by national mapping programmes or commercial archaeology. The recommendation from this thesis was to start applying deep learning techniques which in other complicated fields have reached human level accuracy. Since her thesis several other papers have been published with knowledge based approaches (e.g. Sevara et al., 2016), however, most automation research has followed this recommendation. An up-to-date review of knowledge based approaches can be found in Lambers et al. (2019).

2.5 Machine Learning

Different from feature engineering, with deep learning you ultimately want the algorithm to choose or create the most important features needed for a correct classification. For a visual understanding of features we have created Figure 2.5 with two different Sobel kernels that highlight horizontal and vertical edges. The matrix transformations used in image processing are used in deep learning algorithms like CNNs. Where a filter is convolved over an image (another matrix of pixel values) to detect features such as edges or colour intensities which are important for a correct classification.



1	0	-1
2	0	-2
1	0	-1

-1	-2	-1
0	0	0
1	2	1

FIGURE 2.5: (A) Aerial image of a barrow next to a road in the New Forest (©Crown copyright and database rights 2020 Ordnance Survey). (B) 3x3 pixel vertical Sobel kernel applied to image (A), which highlights vertical lines. (C) 3x3 pixel horizontal Sobel kernel applied to image (A), which highlights horizontal lines.(A) and (B) created using Sobel kernels available at [https://setosa.io/ev/image-kernels]

Deep CNNs were first applied to a large dataset by Krizhevsky et al. (2012). They drastically reduced the state-of-the-art error rate of the ImageNet image classification competition (Deng et al., 2009) from 26.1% to 15.3%. These CNNs consist of several layers of which Convolutional, Fully Connected, Rectified Linear Unit (RELU) and Pooling layers are most important. A convolutional layer consists of a set of learnable filters. This will output a stack of 2-dimensional activation maps which illustrate the responses of the learned filter. This is often followed by a RELU layer which is an activation function that keeps only the positive activations from the convolutions. After this combination of layers, a pooling layer is used to downsample the spatial size of the representation to reduce the computational load in the network. In most common CNNs these three consecutive layers are repeated until the image has been merged spatially to a small size. This is then finally followed by a fully-connected layer which predicts the final score for each of the classes. In their architecture, CNNs appear to parallel mammalian vision by learning filters to perform functions like edge detection at early layers and, at higher levels, specific patterns which we may recognise as objects or their parts. In a way, aerial image classification is a more simple task than classifying the general scenes of ImageNet: it is generally consistent in viewpoint (overhead imagery) and scale (known ground resolution) which reduces the variations of the object's appearance and simplifies the classification task (Mnih, 2013). However, there are many reasons why the detection of archaeology on aerial imagery is more challenging which should be addressed with more complex solutions:

- Small datasets; When a model is trained on only a few examples it is at risk of overfitting to the training data. In this situation the model memorizes the training samples and does not generalize well to new data. Many real-world problems that are being solved with machine learning face this issue and thus several techniques have become available to encourage CNNs to learn more general representations.

- Class imbalance; in archaeology we have unbalanced datasets with only a few examples for each class and globally a lot of background examples against only a few foreground examples. In machine learning a loss-function minimizes the model-error. With class imbalance this is solved most simply by classifying all objects as the majority class (background). A specific loss function that penalises this should be considered to overcome the issue.

- Noise; Archaeological sites are the most overwritten patterns in the landscape, every period following another has added more noise which has an affect on the variability of site appearance and their detectability. Overgrown vegetation, natural erosion, agricultural activities and in some cases looting should be considered.

- Scale; in archaeology we are looking for small objects with detailed variation in a large landscape which presents a harder task than separating a woodland area from agricultural terrain. Undoubtedly this creates a much more complex decision boundary. Due to the need to learn such highly nonlinear decision boundaries, highly advanced machine learning approaches are required.

- Low contrast; In high-resolution LiDAR analysis the task is to separate earthworks from the natural terrain undulation. This is much more challenging than separating modern roads or buildings from their surroundings.

- Non-conventional data format; Data from LiDAR derived DTMs and (multi-) spectral satellite imagery are often supplied in 16 and 32 bit or float images. For a large DTM this is important because large continuous areas can span over 256 meter height difference and is often captured at < 1 metre resolution. When fitting this data into an 8 bit image you will loose vital detail. An easy solution is to use one or more of the previously mentioned DTM data visualisations that highlight the local terrain differences. However, inevitably detail is lost with such visualisations. Raw DTM processing can be achieved with proper rescaling.

- Changing appearance; the archaeological remains have subtle changes in appearance depending on the geology in different geographical areas.

- Fuzzy site definitions; Finally, archaeological sites are often classified according to rough rules but show a lot of variance between them (e.g. banjo enclosures or hillforts). The opposite is true for Roman sites which are often built according to strict patterns which are similar to modern building practices.

In the following sections we present the known deep learning approaches to image classification (subsection 2.5.1), object detection (subsection 2.5.2) and object segmentation (subsection 2.5.3). In image classification a class is predicted for the whole image. In object detection all objects in an image are given a bounding box and class. In semantic segmentation all pixels in an image are given a class. The results and key considerations from the reviewed papers are summarised in Table 2.2, Table 2.3 and Table 2.4. We have mainly focused our review on their fundamental approach (Table 2.2) and their additional efforts to prevent overfitting (Table 2.3). The results Table 2.4 reports on the true positives, false positives, false negatives, precision, recall and F1 score (precision, recall and F1 defined in Equation 2.1, Equation 2.2 and Equation 2.3 respectively). Researchers in deep learning will generally try to optimise the trade-off between recall and precision to get the highest F1-score. The same table also reports on the number of foreground and background examples in the validation dataset of image classification. For object detection and segmentation we report on the area size of the validation area. This addition is important because a larger area is more prone to have false positives and it is more impressive when their results are good.

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F1 = 2\left(\frac{Recall * Precision}{Recall + Precision}\right) \tag{2.3}$$

Where TP = true positive, FP = false positive and FN = false negative.

2.5.1 Image Classification

The first deep learning approach in aerial archaeology was presented by Zingman et al. (2016) who compared their research of a knowledge based rectangular-shape feature detection algorithm with a CNN approach. Their pre-trained CNN was trained with only 9 real examples (plus 135 augmented examples) and 49584 negative samples. They concluded that their handcrafted features yielded higher performance but recognised that the actual detection rate of the CNN cannot be reliably estimated due to a very small number of positive examples. Visually the results were interesting for a first time approach, the CNN classified images that are conceptually close to rectangles (Figure 2.5).

Another case study with image classification was presented by Trier et al. (2018) (research was presented at a conference in 2016 (Trier et al., 2016)) who used greyscale DTM (scaled between 0-255) to detect charcoal kilns. They had 375 (0.15 km ^{2}) positive examples and 10,027 (4 km^{2}) negative locations of lookalike heap objects each cropped to 101x101 pixels (20.2 m x 20.2 m). It's interesting that they have chosen a challenging background class. The approach will show how well it can discriminate between similar classes, our expectation in such a case would be that the trained CNN would not generalize well to the overall background terrain. For training the image classifier they used

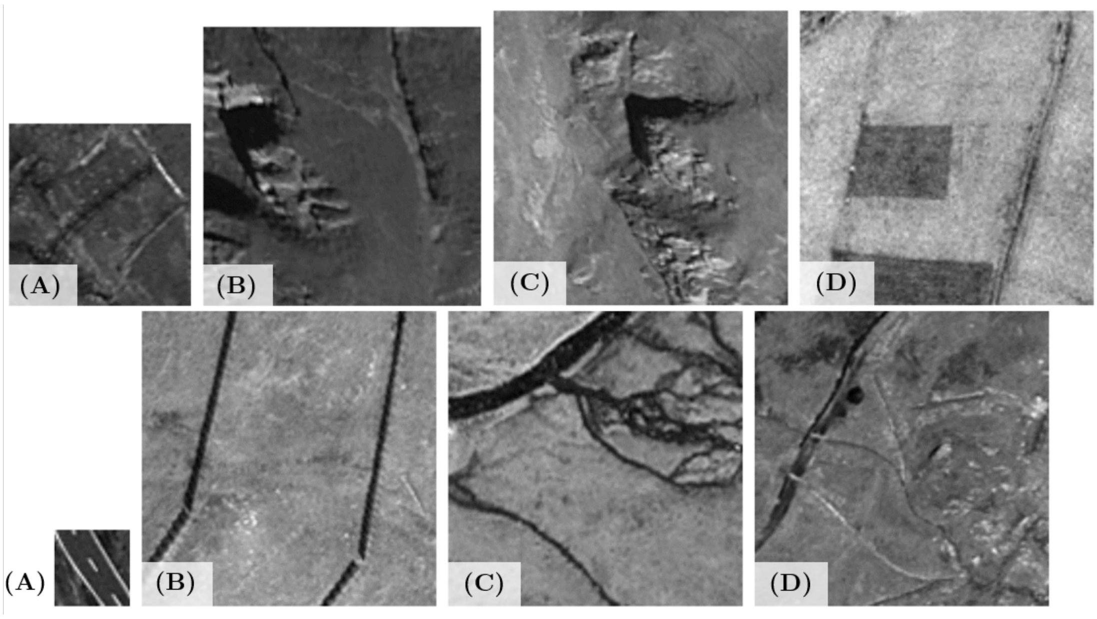


Figure 2.6: Four patches that generated the highest responses by AlexNet (top row) and Vgg-f (bottom row) architectures of pre-trained CNNs. Images from Zingman et al. (2016).

the AlexNet CNN that was pre-trained on ImageNet and used the second to last layer as input to train a linear Support Vector Machine (SVM). To infer the success of their approach they ran the image classifier over a large area with a sliding window of 1 meter (the threshold for their final classification is unknown). Unfortunately the training data for the image classification came from the same area as used in this inference step so those reported results are biased. The inference area size is 9 km^2 so almost half of the negative examples (the heap class) and all positive examples were already seen previously by the image classifier. Most interesting result between the two approaches is that in the large area assessment the false positives only increased by 184 so the approach was relatively successful at disregarding the background of unseen data. They also found 9 previously overlooked potential sites. Even though the approach isn't solid this was another early stage example and provided an interesting case study.

The Norwegian Computing Center who created the previous case study was also commissioned to apply their approach to a case study from Historic Environment Scotland on the Isle of Arran. Trier et al. (2019) used a Simplified Local Relief Model (SLRM) visualisation of LiDAR data and trained on roundhouses (121), shieling huts (267) and small cairns (384). They trained a separate ResNet18 (pre-trained on ImageNet) for each class and trained against background/negative images. They excluded images with common confusion objects such as burial cairns, enclosures and modern cattle feeders to artificially reduce the false positive rate. We would argue against that practise because the CNN should learn to classify those objects as background, such discrimination quality is especially important when it is applied over a large area. For inference they have applied the same approach as Trier et al. (2019), running their image classification

Table 2.1: Data reproduced from Trier et al. (2019), showing training accuracy and validation accuracy change for each epoch on the roundhouse dataset. The training accuracy is only improving from classifying all objects as background at epoch 7, 9 and 10.

Epoch	**Training Accuracy**	**Validation Accuracy**
1	0.9847	0.9936
2	0.9876	0.9925
3	0.9890	0.9946
4	0.9856	0.9946
5	0.9895	0.9834
6	0.9872	0.9845
7	0.9894	0.9791
8	0.9883	0.9914
9	0.9903	0.9925
10	0.9907	0.9888

model that was trained on the 101x101 pixel images on large 2048x2048 tiles. In this case they have visualised the accuracy for each image with a probability map on top of the SLRM (Figure 2.7). They didn't threshold the results and decided true/false positives on visual inspection. This defies the purpose of automation because the archaeologist still has to look at each pixel to verify the results. Similar to Trier et al. (2018) it is not clear whether the results shown were previously used for training so that should be kept in mind when reviewing those results in Table 2.4. The results overall are poor despite a seemingly good training and validation accuracy at the image classification stage. At closer inspection the training accuracy is only improving from classifying all objects as background at epoch 7, 9 and 10. This is evident because only 80 foreground and 7355 background examples were used. Class imbalance is a known issue in machine learning and sometimes overlooked because of a high accuracy as it was done here. The authors should have done more experiments to better understand this issue and find ways to improve it using hyperparameter tuning. A simple confusion matrix would have helped them and the reader to understand the issue at hand. Instead of noticing the issue they used the model weights at epoch 3 or 4 for their inference because the validation accuracy seemed highest (Table 2.1). Probably, they used both training and validation images in the inference because it would explain the poor overall results. The discussion and conclusion section reflects that sometimes "artificial intelligence is being applied without proper understanding" and "as the study presented here demonstrates, the reasons for differing performance of deep neural networks are complex, and there is a pressing need to explore the reasons for this variability in output.". They argue the main reasons for poor performance are the neural network structure, the "black box" problem and the number of training examples. However important those points are, we argue that the authors made some fundamental faults that can easily be addressed.

Caspari and Crespo (2019) presented an image classification approach to detect burial mounds on satellite imagery. Their data was split with 75% for training and 25% for

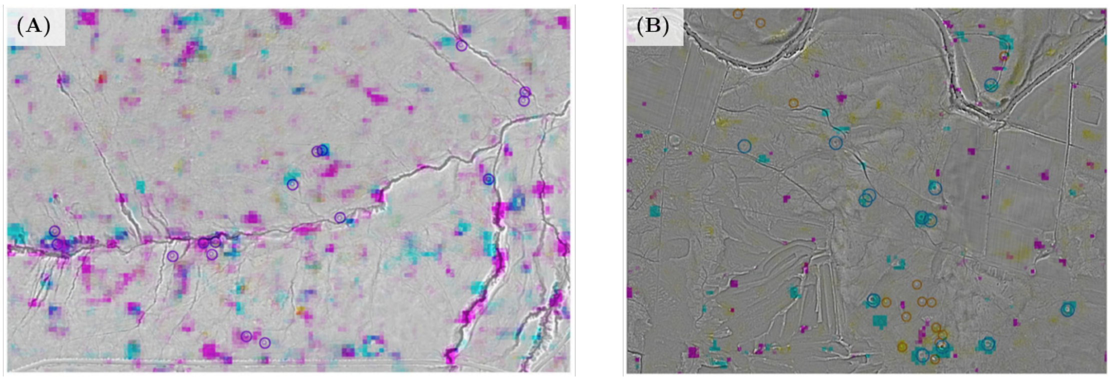


FIGURE 2.7: Heatmap detection results (coloured overlay) from Trier et al. (2019). (A) are the results for Glen Shurig, showing probability of roundhouses (cyan), shielings (magenta) and small cairns (yellow) and verified sites are depicted as circles. The results in this area were chaotic with a large number of false positives for shieling huts and for roundhouses. (B) are the results for Machrie Moor with less chaos and some correct detections.

testing and validation. They created their own CNN using 3 convolution and pooling layers with ReLU activations and two fully connected layers before the final activation with a sigmoid. We would recommend using a State-Of-The-Art (SOTA) model rather than making custom networks. These SOTA models have been extensively benchmarked on various datasets and are well understood. These also have pre-trained versions available which significantly reduces overfitting. That being said, we do appreciate the CNN is benchmarked against another machine learning approach, here SVM. Overall the approach is simplistic but we appreciate that the authors were cautious in their approach and understood key concepts.

Somrak et al. (2020) used images classification to map Aguada's, Buildings & Platforms in a 230 km^2 area around Chactún, Mexico using LiDAR data. In their approach they used several tests with different hyperparameters to find the best performing model. They tried 2 and 15 pixel buffers to understand the importance of context around objects. They also experimented with data augmentation and varied the trainability of their VGG-19 architecture with 3 or 5 frozen layers at the top. Potentially the most interest hyperparameter they experimented with are 6 different visualisation techniques all of which they found worked well in the local environment. For example, Visualization for Archaeological Topography (VAT) is a blend of analytical hill shading, slope, positive openness and sky-view factor into a single greyscale image. They also made several adaptions, for one they placed slope, positive openness and sky-view factor into several channels to create RGB images. Their best performing model used this "VAT-HS channels" visualization, image samples with 2-pixels edge buffer, data augmentation and five frozen layers. They also extensively reviewed the confusion classes to better understand how the deep learning model is performing which is key benefit of the image classification approach. The results noted in Table 2.4 is the micro average of all the classes including the background terrain.

2.5.2 Object Detection

The first approach for object detection was published by Verschoof-van der Vaart and Lambers (2019). The first author has developed and improved this approach for his PhD in Archaeology (collaboratively with the Leiden Centre of Data Science). He has developed a workflow called WODAN (Workflow for Object Detection of Archaeology) to detect barrows and Celtic fields on a 440 km^2 area using LRM visualisation of LiDAR. They applied the state-of-the-art (at the time of publishing) Faster R-CNN which generates object proposals within an image, extract features from the proposals using the CNN, and then classify those. The authors are very upfront about potential shortcomings. For example they have cut the large case study area without overlapping tiles. This has dissected 3% of their target objects which will have an adverse effect on their detectability. They also note that they have manually found common false positives in potential barrows and small dunes (caused by drift-sand) in image patches which they have excluded from the analysis to avoid unbalanced increase in false positives. Only 12-18 epochs were used to train the model, to avoid overfitting. They have also experimented with different backbone network architectures and found that they weren't able to train Resnet50 for multi-class detection so they decided to favour VGG16. That must be a bug in their implementation because any network can be reformed to detect multiple classes. Nevertheless it is a really good first paper to apply deep learning to aerial archaeology object detection.

2.5.3 Object Segmentation

The first case study using segmentation was presented by Gallwey et al. (2019) who looked at mining pits on LiDAR derived DTM. They used a U-net model which has proven to be successful in many domains. It was created by Ronneberger et al. (2015) for the detection of cell tracking in biomedical image analysis. This domain and data source shares similarities with aerial archaeology such as small datasets, fixed scale, high resolution, small objects, indistinct boundaries and greyscale images. Because the authors used raw DTM they had to rescale their 16-bit float images. They applied min-max normalisation to rescale the individual patches between 0–1 which maintains the original distribution before converting them to an 8-bit integer format. To enhance contrast they further rescaled the image tiles linearly prior to model input. By quantising from 16-bit (65,536 distinct values) to 8-bit (256 distinct values) they lose a lot of information. This is not necessary because the model transforms the input image to 0-1 floating point. Using 8-bit images will especially have a large effect on mountainous regions where there is a high variance in height - sites would become visually indistinctive. We further address this issue in subsection 5.2.2. Because they work with greyscale images they realised that transfer learning using ImageNet weights would probably only slightly improve their performance. Instead they used a model that was pre-trained on a large planet scale DEM dataset that was used to detect craters on the Moon. They kept the hyperparameters the same and only retrained for 4 epochs on the 520 images (1568 mines) of their own case study and they applied several data augmentation techniques. The results can be found in Figure 2.8. They compared the approach results from the raw DTM with several visualisations of the DTM and found that the raw DTM worked best. This is expected because the model they used was pre-trained on a similar DEM dataset and not on a dataset that is optimised for the human visual spectrum. Nonetheless this is a great example of clever domain adaptation.

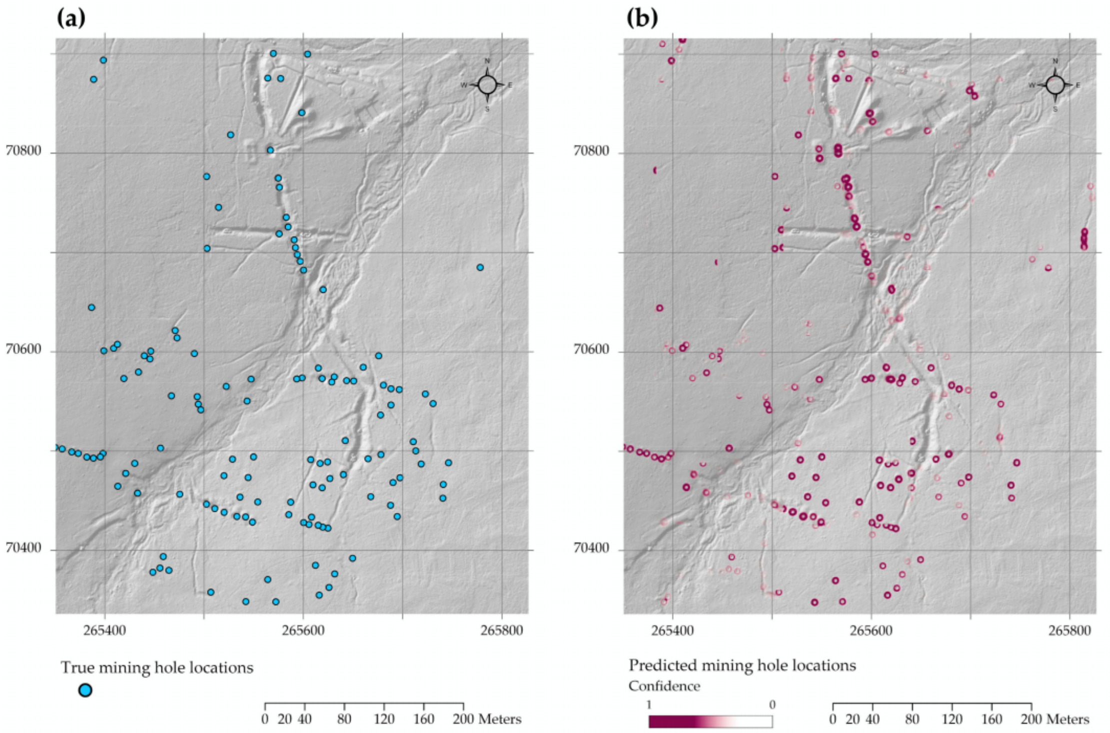

(a)
(b)
70800
70600
70400
265400
265600
265800
True mining hole locations
Predicted mining hole locations
Confidence
1
0
0 20 40 80 120 160 200 Meters
N
W
E
S

FIGURE 2.8: Detection results from Gallwey et al. (2019) on the Dartmoor Hexworthy mine test area. (A) shows the true mining hole locations in blue and (B) shows the model's predicted mining hole results depicted with a graduated transparency colour scale representing model confidence in magenta.

Kazimi et al. (2019) used a variation of DeepLabv3+ to detect bomb craters and charcoal kilns on LiDAR derived raw DTM. The researchers of this paper also experimented with min-max normalization on the whole dataset vs single images and they found that it was essential to apply this on a per-image basis. To extract training data they cropped 256x256 pixel images from each object out of their large DTM into which they then randomly cropped to smaller 128x128 images to ensure that not all objects had a centre object. The input and output data to the original DeepLabv3+ model is 128x128 pixels and the authors changed the output size to 64x64. This improved their result from the baseline model.

A modified 3D version of a U-net was used by Soroush et al. (2020) to detect qanat shafts on Cold War-era CORONA Satellite Imagery. This type of U-net was created for 3D

Table 2.2: Summary of methods applied by key papers in the literature.

Reference	Sensor	Objects	Method	Deep Learning Platform
Zingman et al. (2016)	Satellite Imagery	Enclosures	Image Classification, Alexnet	Matlab toolbox, MatConvNet
Trier et al. (2018)	LiDAR	Kilns	Image Classification, Alexnet + SVM	CNN in Caffe, SVM in Scikit Learn
?	LiDAR	Cairns, Shieling huts, Roundhouses	Image Classification, ResNet18	PyTorch
Caspari and Crespo (2019)	Satellite Imagery	Burial mounds	Image Classification, Custom CNN	Keras & TensorFlow
Somrak et al. (2020)	LiDAR	Aguada, Building, Platform	Image Classification, VGG-19	Keras & TensorFlow
Verschoof-van der Vaart and Lambers (2019)	LiDAR	Burial mounds, Celtic fields, Charcoal kilns	Object detection, Faster R-CNN with VGG-16	Keras
Gallwey et al. (2019)	LiDAR	Mining Pits	Image Segmentation, Unet	Keras
Kazimi et al. (2019)	LiDAR	Bomb craters, charcoal kilns	Image Segmentation, DeepLabv3+	Keras
Soroush et al. (2020)	Satellite Imagery	Qanat shafts	Image Segmentation, 3D Unet	Keras

medical data such as scans of the brain. It is unclear why the authors preferred this model over the traditional U-net. They did not use a pre-trained model but instead focused on several hyperparameters including a specific loss function, batch-normalisation, drop-out and they added data augmentation to reduce overfitting. They consider adding more approaches to artificially remove false positives with post-processing steps. Rather than simply masking certain areas they propose to use their domain understanding of the linear pattern in which these qanats were placed. It would be interesting to see if that is a pattern that can be found with machine learning or whether that has to be hard coded. This is especially so because context is important for almost all archaeological sites. Unfortunately the authors didn't report on training and validation results separately and provide the result on all 11 patches. This means that readers cannot interpret the transferability of the approach from one area to the next. Moreover, the trained model will have memorized the examples in the training data so the recorded results are worse than the results reported in the paper.

2.6 Discussion

At this early stage of deep learning most case studies in aerial archaeology are still in the feasibility phase. There are several trends, critical observations and issues we will discuss in this section and further address in the thesis.

Most studies face overfitting as their key issue. This can be seen because most studies only train their approach between 3 and 20 epochs. They stop training because their validation accuracy drops which means their model is overfitting. At that point the model is learning the exact examples from the training data which doesn't generalize

Table 2.3: Summary of data processing types applied by key papers in the literature.

Reference	Image Pre-processing	Pre-training	Data Augmentation	Notes
Zingman et al. (2016)	Raw greyscale	ImageNet	16 rotation angles were taken uniformly in the interval [0, 360] degrees	
Trier et al. (2018)	Greyscale, normalized contrast and mean values from dataset (scaled 0 - 255 with scaling factor limited to 25 or less)	ImageNet	8 variation of rotating and flipping	
Trier et al. (2019)	SLRM (repeated for R-G-B channels)	ImageNet	Horizontal flip, rotation, random scaling and random translation	
Caspari and Crespo (2019)	Raw colour (RGB)	No pre-training	Horizontal flip, random zoom and shearing	
Somrak et al. (2020)	VAT, Flat VAT, VAT-HS, VAT-HS channels, PRIM, LD	ImageNet	Pre-processed 3 rotations to maintain consistent hill shading. In Keras; random zoom, width shift and height shift	1. Different numbers of untrainable, frozen layers at the beginning of the network. 2. Oversampling the aguada minority class by rotation (creating multiple hill shading directions)
Verschoof-van der Vaart and Lambers (2019)	LRM	ImageNet	Horizontal and vertical flip and 90° rotations	
Gallwey et al. (2019)	Greyscale with min-max normalisation and linear rescaling	Lunar DSM	Randomly flip, rotation and shift	Pre-training on a dataset similar to the target.
Kazimi et al. (2019)	Greyscale with min-max normalisation	No pre-training	Random cropping, random rotation	
Soroush et al. (2020)	Raw greyscale	No pre-training	Horizontal and vertical flip	Loss function was designed for class-imbalanced datasets.

Table 2.4: Summary of results by key papers in the literature.

Reference	Foreground/background for total km^2	True positives	False positives	False negatives	Recall	Precision	F1
Zingman et al. (2016)	Test: 1/57,504						
Trier et al. (2018)	Test: 375/10,027	317	35	58	0.85	0.90	0.87
	Train/Test: 9 km^2	372	219	60	0.86	0.63	0.73
Trier et al. (2019)	Test 1: 1 km^2	15	54	20	0.43	0.22	0.29
	Test 2: 1 km^2	5	65	14	0.26	0.07	0.11
?	Test: 42/261				1	0.84	0.91
Somrak et al (2020)	Test: 2492/2168	4453	184	23	0.99	0.96	0.98
Verschoof-van der Vaart and Lambers (2019)	Test: 10.9 km^2	55	6	23	0.71	0.79	
Gallwey et al. (2019)	Test 1: 1 km^2	155	37	38	0.81	0.81	0.81
	Test 2: 0.2 km^2	142	13	30	0.83	0.91	0.87
Kazimi et al. (2019)	Bomb crater area: ?	49	?	2	0.96		
	Charcoal kiln area: ?	228		5	0.97		
Soroush et al. (2020)	Train/Test: 60.5 km^2	2863	1785	989	0.74	0.62	0.705

well to the unseen data in the validation dataset. Generally, deep learning requires large training datasets but in aerial archaeology there aren't many examples to train on. In table x we have noted the various methods that the researchers have taken to prevent overfitting. Most popular approaches are pre-training and data augmentation. However, there might be a more fundamental problem such as a noisy dataset or class imbalance. We fear that uncritical approaches fuel the disbelievers of automation approaches such as the recent publication by Casana (2020). Quoting Verhoeven (2017): "Too often, incorporating (new) digital technologies in archaeology while lacking any theoretical framework is said to be meaningless and even erroneous conclusions are drawn". Since deep learning is completely reliant on digital data we need to be especially aware of bias.

Approaches have to be appropriately backed by theory and the results should be interpreted alongside theoretical frameworks. Contrary to knowledge based approaches, in deep learning researchers do not explicitly model their domain understanding. This abstraction can seem like the approach is a black box and researchers like Trier et al. (2018) and Trier et al. (2019) have also reiterating that statement in their work. However, with the proper understanding researchers can get a lot of feedback from their models which they should use to tune several hyperparameters. We have seen extensive hyperparameter tuning in Somrak et al. (2020). We have also seen that Soroush et al. (2020) chose a loss function that was designed for class-imbalanced dataset and Gallwey et al. (2019) used a network that was pre-trained on a LiDAR dataset and objects that were similar to their target. Overall we think the right approach is to start with image classification to test different parameters that are transferable to object detection and segmentation. At this point researchers can quickly review confusion matrices to understand the resulting accuracy and they can also visualise images that are most confused to quickly understand whether the dataset is noisy (Somrak et al. (2020)). Understanding the potential of a dataset at an early stage can speed up the research and helps researcher to find out where to invest time to gain accuracy increase. Somrak et al. (2020) tried several visualisations and found that Local Dominance didn't work well on the dataset despite being one of the most important manual tools for classifying the objects. Object detection and segmentation take much longer to train than image classification and it is thus more expensive to tune. If they went straight for large scale mapping than they may have not had the means to tune the visualisation and would have had to conclude that automation was useless on their case study.

Most researchers evaluate their approach with the false positive rate and the final F1 score. To improve their false positive rate Verschoof-van der Vaart et al. (2020) published an updated version of their approach using Location Based Ranking to mask built-up areas, and areas with drift-sand that were known to have low likelihood of archaeology but a high number of false positives. Ultimately the success of an approach is not dependent on one metric, it depends on what is most suited for a specific task (Soroush et al., 2020). In the medical profession classifying a sick person as healthy has a different cost than the opposite case and so doctors prefer to review more false positives and accept a higher recall with lower precision. In the case of Verschoof-van der Vaart et al. (2020), their focus was on large scale mapping where it was accessible to miss a few objects for a higher precision to increase the overall success measured in the F1 score. Automation in archaeology is still at an early stage where researchers are trying to locally optimise their approach. In the future we foresee that a heritage managers may accept high recall with lower precision when it only takes them a short while to sift through the detections. The same is apparent in commercial archaeology where high recall is the most important metric.

Our final observation is that some researchers are not concerned with geographically separating their results. However, it is a really important test for the transferability of the approach and only then can the validation and testing be really attested for. The testing areas of Gallwey et al. (2019) were 20 km^2 and 500 km^2 away from the training data which is most impressive but also the approach of Soroush et al. (2020) works where the whole study area is split in a training and validation areas.

In the remainder of this thesis we will further address our observations and propose several of our own solutions.

We expect the next phase will include more research using object segmentation. This approach will address the major flaw in object detection which is that most archaeological sites do not fit within bounding boxes. Rather than focusing on specific sites we suspect that the most value will initially be in the detection of concave/convex earthworks on LiDAR data or positive/negative crop marks on aerial imagery.

After this stage it starts to become more important to include more geographical sources such as soil type, hydrology, land use and vegetation cover. It is a highly specialist job for an archaeologist to distinguish natural and modern features from archaeological features. Often just one source of data is not enough and specialists use other earth observation sources or geographical maps.

Chapter 3

The New Forest Case Study

This chapter presents our initial approach to image classification and a first attempt at object detection. Most of this research was undertaken in the first year of the PhD (2017) when the literature on deep learning for remote sensing datasets was scarce, especially in relation to archaeology. The objective of the experiments was to find out whether CNNs could be trained to detect barrows in multi-spectral imagery and LiDAR derived DTMs. We have particularly focused on our identified data challenges in chapter 2. We first discuss the case study area and how we created the dataset (section 3.1). Our initial experiments are divided into basic image classification (section 3.2) and object detection (section 3.3). We will conclude the chapter in section 3.4 with a discussion on the challenges that we addressed and describe which challenges need more work in the following chapter and which we cannot overcome within the PhD and are classed as future work.

3.1 Dataset

Despite the availability of countrywide remote sensing data for the UK, we are still unable to process the petabytes of data. We thus limit our research to a 600 km^2 area of the New Forest in the south of England. The area is known for its diverse land cover and rich archaeology, and will thus be a good testing ground for including data from multiple remote sensing. During the initial stages of this research we established a collaboration with the New Forest Archaeological Mapping Project. Their extensive research of discovering new archaeological sites, especially using remote sensing data, has provided us with a very good dataset of known sites (subsection 3.1.1). In addition to the site locations they have kindly provided us with LiDAR data which complements the aerial imagery available from Ordnance Survey, who sponsored this research (subsection 3.1.2).

To view our datasets in their geographic context we used QGIS which is an open source Geographical Information System (GIS) software used to process remote sensor (raster) data and location (vector) data (QGIS Development Team, 2020). In GIS software vectors can consist of three types: polygons, lines, and points. In this software there are several pre-existing tools for both vector and raster processing which can be pulled together to process the large datasets into a format that be used for machine learning (Verschoof-van der Vaart and Lambers, 2019). However, we found it easier to process the data directly in Python with specific geography packages. For vector data we used the OGR and Fiona libraries and for raster data we mainly used the GDAL library.

3.1.1 Site Locations

The objects chosen for this initial case study are barrows (Figure 3.1), also known as grave mounds, which typically date back to the early-middle Bronze Age around 3,500 years ago (Field, 2011). These objects are amongst the most common monuments of prehistory all over the world and have been the target of many other automation projects (de Boer, 2007; Riley, 2009; Trier et al., 2015). Barrows appear as circular mound structures and have similar appearance to naturally occurring elements (e.g. fairy ring [1]) and modern human-made structures (e.g. roundabout), which are likely to cause false positives and may challenge the accuracy of a CNN as noted by Trier et al. (2016).

The known archaeological site locations are a combination of the record from the local archives Historic Environment Record (HER) and the more recent discoveries made during the New Forest Archaeological Mapping Project. The locations of barrows are shown in Figure 3.2. The data is provided as shapefiles[2] with central points of every site.

The archaeological sites used for this research have been discovered over the last 100 years using different methods including remote sensing, but also ground survey methods such as geophysical techniques. This means that some objects are not visible on both or even on either of the data modalities. Additionally, some objects are historically classified and have since been destroyed. Experiments will determine whether the given data can be used as a raw resource or if further manual tuning is required.

3.1.2 Remote Sensor Data

Remote sensing data is often very high resolution which over large area creates big data. This big data is difficult to load into memory or transfer and therefore datasets are generally provided in multiple different files. These files can be processed individually or as Virtual (VRT) Files. VRT files contain links to the individual images that are available for a RS dataset. This file type significantly speeds up image cropping because

[1] A fairy ring is a naturally occurring ring or arc of mushrooms.

[2] The shapefile format is a geospatial vector data format.

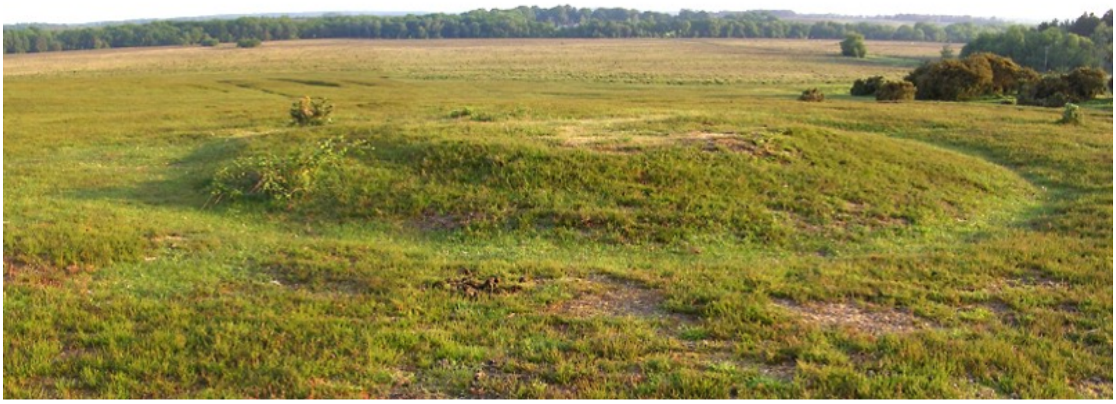


Figure 3.1: Photograph of a barrow captured by Champion (2006), at Longdown (New Forest, grid ref SU36280830). This barrow is 8 metres in diameter and up to 0.5 metres high.

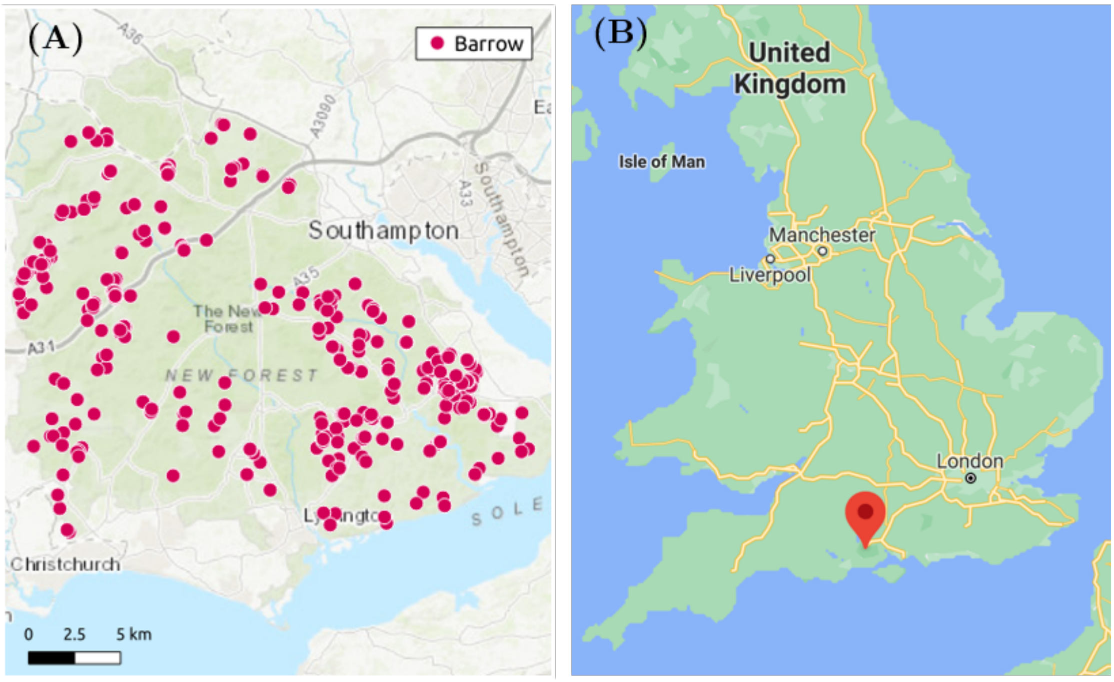


Figure 3.2: (A) New Forest National Park overlaid with the locations of known barrows. (B) Location of the New Forest marked by a red indicator on a map of the United Kingdom (GoogleMaps, 2020b).

it loads only the area of interest into memory and more importantly it includes all the images during this process which allows objects at bordering images to be merged in the process. Additionally, VRT files can be provided with additional instruction for processing the data when images get extracted from it. In our case this includes the re-projection of the coordinate system to EPSG:27700, to set the resolution to 0.5m pixels and to scale all images at 8 bits per sampled pixel.

3.1.3 Aerial Photography

The Ordnance Survey aerial imagery is captured with Red-Green-Blue-Near infrared (RGBN) bands and is provided in 16-bit unsigned integer format. The images have undergone some basic pre-processing for merging the individual photos captured during the flight but are not colour corrected and thus show colour imbalance and artefacts at the seams (Figure 3.1). In total, 6 grid tiles of 10 kilometres along each side (tiles SU20, SU21, SU30, SU31, SZ29, SZ39 in the British National Grid) have been provided, with 0.5m ground resolution. Each of these is about 3.5-4 GB in size (Figure 3.3 (A)). This data was captured in August 2016 which makes it possible for cropmarks to be seen, especially in the near infrared band.

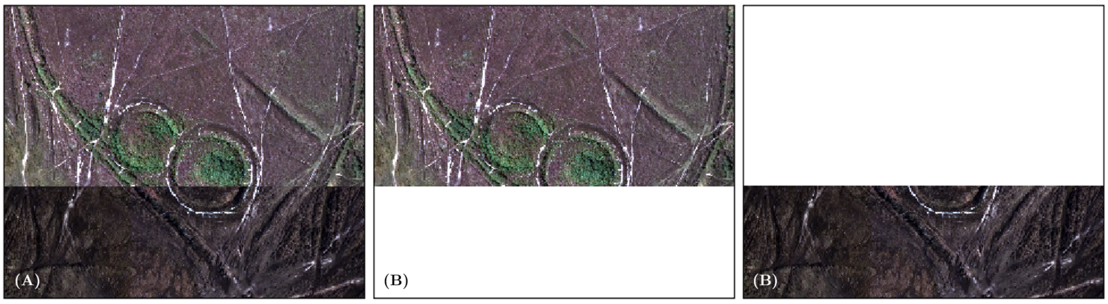


Figure 3.3: Colour imbalance between provided tiles SU20 (middle) and SZ29 (right). ©Crown copyright and database rights 2020 Ordnance Survey

3.1.4 Airborne LiDAR

The LiDAR data was captured in two different surveys in December 2011 and January 2015 (University of Cambridge, 2011, Natural England, 2015). Both surveys were performed during the winter when the broadleaf trees are devoid of leaf cover and the understory vegetation is at a minimum. The 2011 survey covers about 400 km^2 and has a minimum of 2 laser points per m^2 (ppm) and reached up to 6 ppm. The point-cloud data was processed to produce both Digital Surface Models (DSMs) and DTMs as IMGs formatted rasters with a 0.5m cell size. These files cover regions of 30-40 km^2 with maximum IMG file sizes of 1 GB (Figure A.1 (B)). Unfortunately, the DTMs include many ‘no data’ patches where no ground points were returned (e.g. houses or dense tree coverage). These areas are not interpolated and might cause a problem for the training of networks. The 2014 data covers about 650 km^2 and is captured with 2 ppm and delivered as a processed 1 m DSM and DTM. The coverage of this survey is significantly larger than the 2011 survey and the images are interpolated without ‘no data’ patches (Figure A.2 (B)).

3.2 Image Classification

To process the geographical data in Python we used the GDAL (raster-data) and OGR (vector-data) libraries. We have implemented the workflow in a Jupyter Notebook to include intermediate feedback steps for printing details about the loaded files, their geographical information and to show plots of the image crops (using Matplotlib) and their geographical point locations (using Basemap from the Matplotlib toolkits).

There are various reasons why image classification is a good approach for the task at hand. Image classification generally gives a bigger window around the object then approaches which perfectly localise objects (e.g. object detection and segmentation). This bigger window is useful as barrows are often found in clusters (Field, 2011), so context might be important for classification. Additionally, our barrows in the dataset are not always accurately located / digitised at the given centre point, and so, a bigger window may slightly alleviate this noise. As we are looking to classify a single object type we mainly base our approach for this section on research from single class image classification techniques which have previously been applied to large datasets like ImageNet.

The success of deep learning has mainly been shown by training large datasets such as ImageNet and it has been argued before that better results come with deeper and more advanced network architectures (He et al., 2016). In our case, however, we have only very small datasets of 260 - 431 barrows and therefore need a different approach where we carefully consider overfitting. Overfitting happens when a network trains on too few examples and learns patterns that do not generalize well to new data. This effect can be witnessed when the validation accuracy is much lower than the training accuracy. In this section we will discuss a range of approaches to alleviate the chance of overfitting. In section subsection 3.3.2 we discuss data pre-processing and especially the approach to overfitting by increasing the dataset by making minor alterations to our existing dataset using techniques like flipping, rotating, scaling, cropping, translating, or adding random noise. For a CNN which is invariant to these changes such augmentation will be interpreted as distinct data to learn from. Besides increasing the amount of data, data augmentation is also good for other reasons. Our images (supposedly) have the object of interest in the centre. A network without augmentation might therefore fit to images with objects in the centre and would not recognise an ‘unseen’ image of a barrow that is not in the centre. So, to an extent, data augmentation can be used to prevent overfitting, however, we need additional approaches. In subsection 3.2.2 we research the best approaches to work with small datasets and experiment with different CNN architectures. In subsection 3.2.3 we analyse the usefulness of transfer learning and compare results from a network pre-trained on general image scenes to one that was trained on aerial images. In subsection 3.2.4 we experiment with different image pre-processing techniques that reduce the complexity of the LiDAR derived DTM and highlight the local archaeology.

3.2.1 Data Pre-processing

The central site locations (XY coordinates) are used to crop images from the VRT-files. Starting with a 100x100 meter around centre location, shift this location by 20 meters (up, down, left, right), zoom to create an area of 80 meters at 0.4 m pixel size (maintaining the same image size as other cropped images) and finally perform the same shift on the zoomed locations (Figure 3.4). The augmentations were carefully chosen to always include the full barrow and have a zoom within the range of expected barrow sizes. Other augmentation options such as rotating, shearing and flipping were considered unsuitable as they might confuse the network (due to human choices to shape the barrows and natural effects to the structure over centuries such as erosion due to prevailing wind directions). Before saving the files, we exclude any images with exclusively “no data” pixel values. We will train our network to distinguish “positive” images with barrows from negative examples and thus create an equal number of negative examples that are extracted at a buffer of 100 meters from all the known barrow.

Finally, all cropped images are saved as GeoTIFF files and separated into folders of 75% training and 25% validation data. In order to later be able to assess the robustness of a trained network to a new area, we have split the data into east and west sets rather than a random division (Figure A.1 (B) and Figure A.2 (B)). Even though it is not expected to affect the RGBN dataset, this might influence the DTMs which have a higher general elevation to the west than to the east (Figure A.1 (A) and Figure A.2 (A)). In the next section this data will be used to train a CNN.

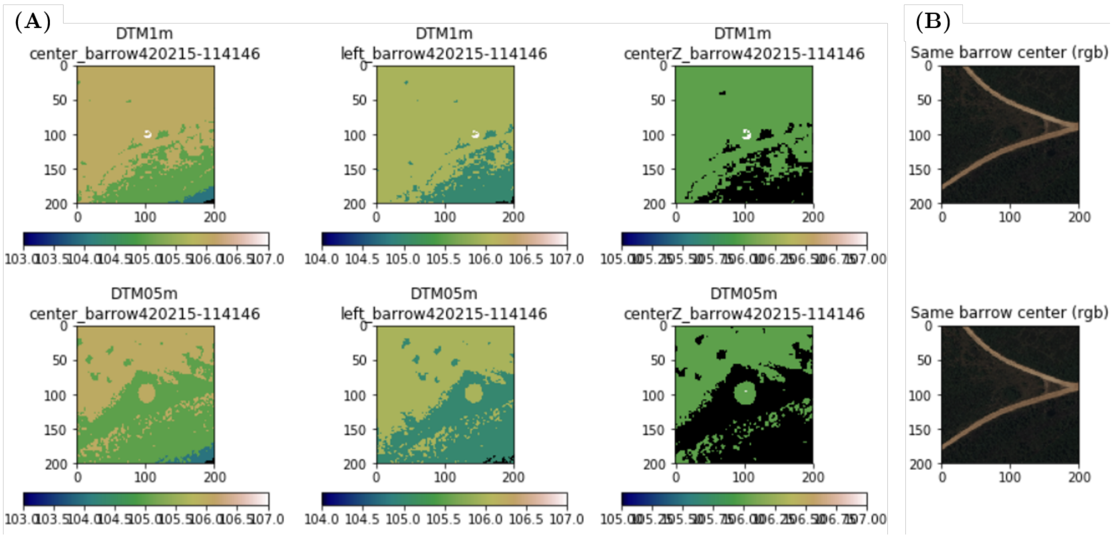


Figure 3.4: Different augmentations from the datasets showing the same barrow.

3.2.2 Experiment 1: Simple Network with Added Regularisation

According to Chollet (2016), the choice of a CNN should depend on the size of the dataset. Complex networks with many layers have more space for information to be stored which has the potential to generate high accuracy. However, when having very little data going through a complex network, this may lead to the creation of irrelevant features and thus lead to overfitting. Whereas a network that can learn less features will have to focus on the most significant features found in the data, and these are more likely to be truly relevant and to generalize better. This argumentation is supported by a demo created by Karpathy (2018) where they show that larger Neural Networks can represent more complicated functions but at the same time it's likely fit to noise/outliers (Figure 3.5). Conversely, they also argue that the complexity of larger network can still be leveraged when the network has sufficient regularisation (Figure 3.6). Below we employ a simple network as proposed by Chollet (2016) and compare different regularisation techniques.

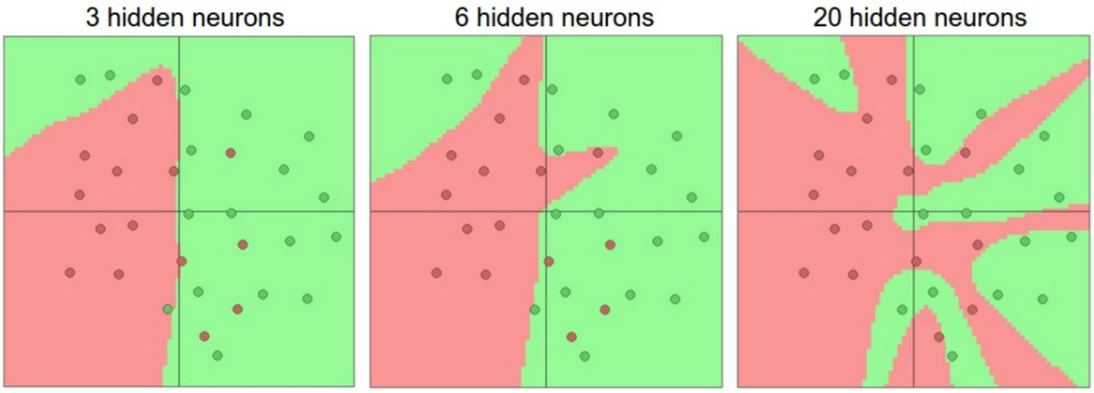


Figure 3.5: These images depict the effect of network depth when classifying two classes. The changing decision regions show that larger CNNs can represent more complicated functions.

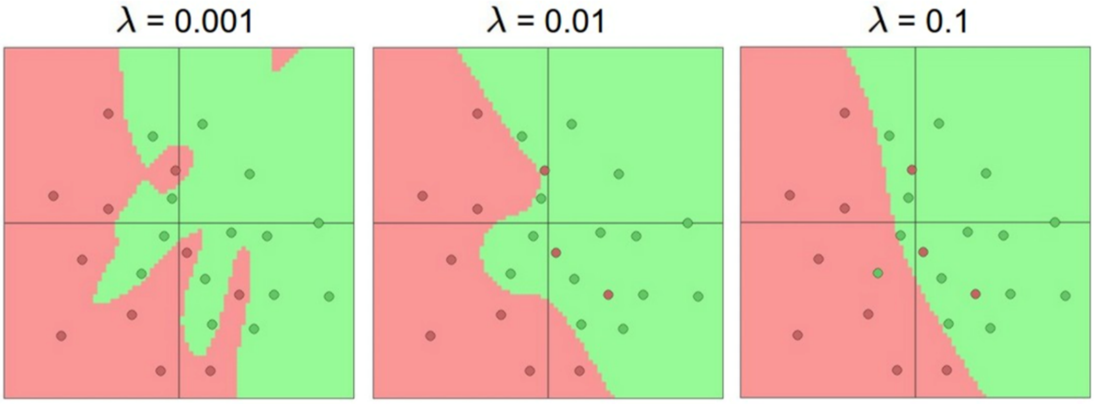


Figure 3.6: These images depict the effects of regularization strength on a large network (20 hidden neurons). With lower λ, the model can increase its complexity by assigning big values to the weights. On the other hand, when increasing λ, the network becomes simpler and smooths its final decision regions.

For this experiment we use the Keras (Chollet, 2015) deep learning library with a TensorFlow (Abadi et al., 2016) back-end. Unfortunately, Keras does not natively accept GeoTIFF files nor can it load images with more than 3 bands, so we have made custom adaptions to the `preprocessing/image.py` file to load our data.

We initially used a simple network consisting of a stack of three convolution layers with a RELU activation and max-pooling, and ending with two fully-connected layers, dropout, a single unit and a sigmoid activation (following Chollet, 2016). Dropout is one of the regularisation techniques mentioned by Karpathy (2018) where a layer randomly switches off part of the neurons to decorrelate the learning of different neurons. Additionally, we applied the more common L2 weight regularization on the convolution and dense layers. This technique forces a network to learn information from all the given data instead of focusing on a specific pattern and does so by penalise spiky weights and favouring diffuse weights. To compare different rates of weight regularization we trained the network 6 times, varying the rate from 0 to 10-6. To train our model we used binary cross entropy loss as we have a two-class problem and ended our network with a sigmoid activation. Additionally, after several attempts with different optimisers we found that RMSProp (Tieleman and Hinton, 2012) provided the most stable results. From the first results we conclude that the network was able to train on RGB but did not generalize for the DTMs. The poor results on DTMs were likely caused by the minimal pre-processing of the DTM. The elevation in the New Forest ranges between 0 and 123 m, feeding the raw DTM with absolute height data has likely confused the model. To overcome this, we normalized the inputs with the means of the training data which improved the results and developed more stable training and validation curves across the datasets.

The results on this network trained for 150 epochs on every dataset are presented in Table 1 & 2 (best accuracy). From these results, we observe that:

- In all cases, we were able to obtain >50% accuracy, demonstrating there is an underlying pattern to the images.
- The RGB and RGBN training show very similar patterns and have best validation accuracies of 78.20% and 77.58% respectively. To further compare this, we also trained the infrared as a greyscale image which on its own got up to 75.78% accuracy.
- The DTM-1 m has the best validation accuracy of all datasets with 83.57% on the maximum training data but during the equal comparison this dropped to 69.32%.
- The networks trained on the DTM-0.5 m show a very unstable validation accuracy which may be a result of the noisy 'no data' patches in the training images.

- The networks trained on the DTM-0.5 m show a very unstable validation accuracy which may be a result of the noisy 'no data' patches in the training images.

- The combinational images do not perform better than RGBN trained networks. Overall, combinational images perform better with the Near InfraRed (NIR) band.
- he networks trained on the RGBN and combinational images do not learn without weight regularization.
- The weight regularization rate of 0.01 seems to be the best on the RGBN data and combinational images. For just DTMs, the 0.0001 works better.
- The results vary a lot across rates and data combinations. This seems to confirm the statement of Karpathy (2018) noting that most local minima in small networks have a high loss and that you have to rely on luck not to get trapped in a bad local minimum.

After various attempt we can confirm that a much deeper, state of the art, network (VGG16) in its most original form without controlled regularisation did not learn no matter the RS data, optimizer or with the addition of weight regularization. We posit that is because of the relatively limited size of our dataset compared to the number of parameters of the network. For better implementation of Karpathy (2018)'s argument and comparison to Chollet (2016)'s argument we will look at a deep network with controlled weight regularisation in the next section.

To further test the success or increase the accuracy of this experiment, the next steps involve:

- Using a deeper network.
- Increasing the dataset size.
- Add other augmentation techniques.
- Add image pre-processing techniques.
- Trying the networks with other object types.
- Applying the trained network on other areas.

3.2.3 Experiment 2: Transfer Learning

Transfer learning is a very commonly used technique for training on small datasets (Razavian et al., 2014). In this process, a network is pre-trained on a very large dataset (e.g. ImageNet contains 1.2 million RGB images with 1000 categories of objects) and is used for its fixed features or to initialize feature extraction. During training, they start by learning more abstract features and further on in the network start to generate

Table 3.1: Accuracy results on networks trained for 150 epochs on the maximum available data. Best result for each data type highlighted in red.

LR	RGB	N	RGBN	DTM	DTM	DTM	DTM	DTM	DTM
				1 m	0.5 m	1 m	1 m	0.5 m	0.5m
						+ RGB	+ RGBN	+ RGB	+ RGBN
0	75.07	75.78	50.43	65.87	59.94	50.43	50.43	50.47	50.47
0.1	76.28	64.7	56.53	51.74	51.58	68.11	66.76	56.02	66.17
0.01	78.20	69.58	77.34	51.98	63.03	75.14	75.21	73.83	77.42
0.001	75.85	75.14	74.50	64.50	63.55	71.59	74.22	69.61	73.28
0.0001	76.49	74.57	76.70	83.57	63.86	71.66	73.72	69.30	75.63

Table 3.2: Accuracy results on networks trained for 150 epochs on 360 barrows (max available barrows that have data in all modalities).

LR	RGB	N	RGBN	DTM	DTM	DTM	DTM	DTM	DTM
				1 m	0.5 m	1 m	1 m	0.5 m	0.5m
						+ RGB	+ RGBN	+ RGB	+ RGBN
0	50.47	51.42	50.47	58.44	57.49	50.47	50.39	50.47	50.47
0.1	65.69	60.33	74.45	52.29	52.52	75.31	64.95	56.02	66.17
0.01	75.71	75.55	77.58	51.66	62.38	75.63	77.65	73.83	77.42
0.001	73.50	74.45	73.59	52.68	61.20	71.02	76.48	69.61	73.28
0.0001	75.53	74.68	72.81	69.32	58.91	68.52	69.63	69.30	75.63

features for the specific classes. Transfer learning has already been successfully applied to satellite imagery (Penatti et al., 2015) and seems to have become essential as most remote sensing-projects lack a large labelled dataset or time to train from scratch, e.g. in disaster response (Zhu et al., 2017).

In this section we will compare a 50-layer ResNet (He et al., 2016) adapted by Chollet (2016), pre-trained on everyday objects (ImageNet) and on aerial photography (TopoNet). TopoNet was recently created by the Ordnance Survey and is trained on 1.4 million RGB images of RS-data from all over Britain captured with the same sensor as our dataset. Even though the ImageNet dataset does not include our object types or any aerial images for that matter, its trained network can still be useful for the features learned by low-level convolutional blocks. For the comparison, we derive fixed features from our data and use those to train a linear SVM classifier (like Penatti et al. (2015)). We will run the classifier at each of the 50 activation layers in the network to analyse how well lower (abstract features) and higher (specialised features) layers in the trained networks perform on our dataset. Both networks were trained using Keras so for this experiment we use the same library to extract the features from the networks. To implement the SVM we use the scikit-learn library (Pedregosa et al., 2011) which is a specialised library of various machine learning approaches.

The outcome of this experiment will not perfectly compare to the previous section. The experiments are basic and mainly implemented to show how weights from different datasets translate to a new target dataset. To this extent we have not applied data augmentation and have also randomly selected train/valid/test data for the different experiment. We also only use the RGB images because we can only use data in the same format as the network it was trained on, which in these cases is 3-band RGB. For baseline comparison we have trained the linear SVM from scratch (without pre-training) on our image data (colour histogram) which obtained a maximum test accuracy of 70%.

The best test accuracy result of ImageNet reached 55% and is worse than the SVM trained from scratch (Figure 3.7). The graph is unstable and varies a lot between layers so there seemingly is not a distinct favouring for earlier/later layers. We see more stable results from the TopoNet data shown in Figure 3.8. The validation accuracy on TopoNet gets to a maximum of about 71% and show an average increasing of accuracy towards the later layers. This reaction of our data to the pre-trained TopoNet shows promising results so we should continue to do more experiments with this is the future.

The next step from pre-training is fine-tuning (Yosinski et al., 2014). In this process a pre-trained network is retrained to adapt to the target dataset. With a trained network on images very similar to your target dataset one would only retrain the last layers of the network and freeze the weights of earlier layers. Conversely, for a dataset that is very different you would keep only the first layers frozen. To fine-tune ResNet50 we chop of the fully convolutional layers at the end, which were tuned to classify ImageNet into 1000 categories and TopoNet into 12 categories. We flatten the outcome at the chopped layer and add a dense prediction layer for our 2 classes. Initial results of fine-tuning the ResNet using both ImageNet and TopoNet weights on barrow data are poor. We have applied fine-tuning on early, middle and late layers and with each of these approaches we find that the networks easily over-fit and that it's difficult to find stable hyper-parameters. So far it seems that retraining the networks without frozen layers gives the best results with around 83% on TopoNet and 80% on ImageNet.

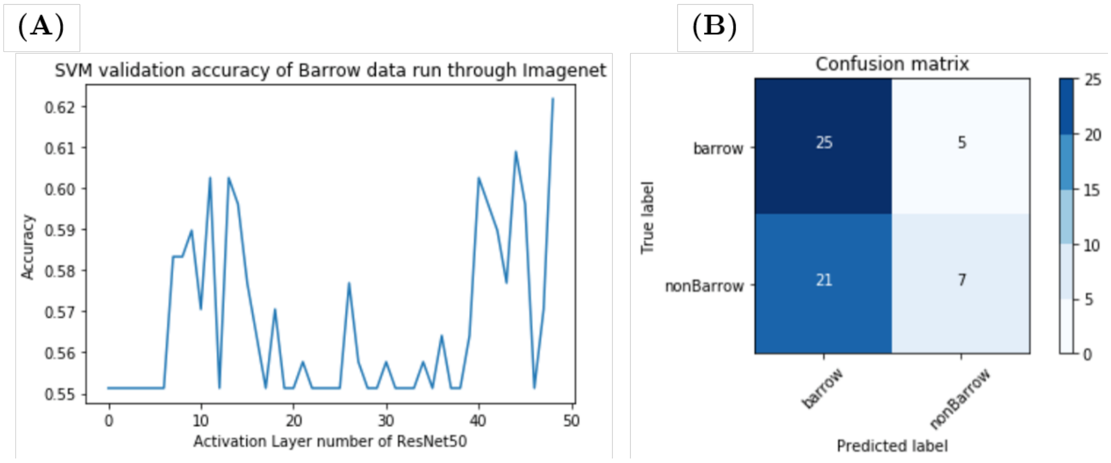


FIGURE 3.7: Results of SVM trained on different layers of ResNet50 trained on ImageNet. (A) the best performing layer was 48 with a validation accuracy of 62% and test accuracy of 55%. (B) the confusion matrix for this layer shows mostly barrow predictions.

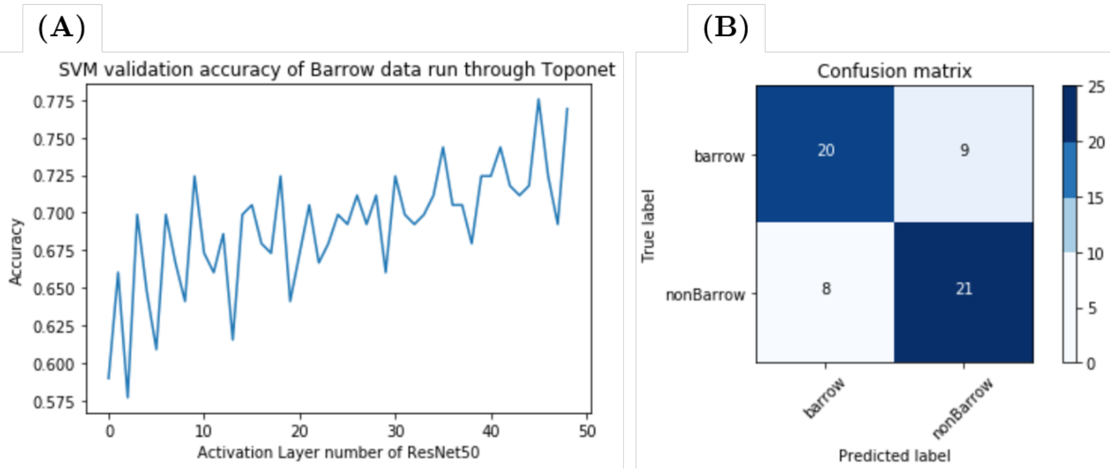


FIGURE 3.8: Results of SVM trained on different layers of ResNet50 trained on TopoNet. (A) the best performing layer was 45 with a validation accuracy of 78% and test accuracy of 71%. (B) the confusion matrix for this layer shows a balanced classification pattern.

3.2.4 Experiment 3: Improving the Pipeline

In our following experiments we researched the most effective improvements. We found it especially useful to visualise the model predictions and analyse what barrows the model found easily and which were most challenging. In the top row of 3.9 the true positives are shown where the model was most certain that the image showed a barrow. In most of these images it's rather difficult to see a barrow and it's mainly through shadow and lack of vegetation that ditches around the barrow are seen for the first images and the final two images are also visible through a different type of vegetation on and around the barrow. In the second row we have depicted the false negatives where the model was most certain that the image belonged to the background class. Visually, archaeology experts are also unable to see a barrow because in the last two images barrows are hidden under a forest and in the first two images agriculture has flattened the barrow and because the field is recently ploughed the archaeology is also not visible through the proxy of vegetation stress. Both results tell us that the dataset is noisy and that we should improve the data with a one to one match between the site location and the RS data to ensure the training data is useful and the results are not biased by the training data. In the third row we show the most certain true negative predictions which show modern buildings, straight lines and corners. The final row shows the false positives where the model was most wrong and sure they were barrows. Again, these images are very interesting and show locations of natural terrain with some curved paths.

Based on the results from the RGB and our intuition from manual analysis we suspected that the DTM data should yield better results than we have previously seen because the LiDAR sensor would have pierced through the forest which should reveal many more sites than the aerial image. We previously found that we had to normalised the DTM to the mean of the training data for the CNN to be trainable. Yet we are not satisfied with

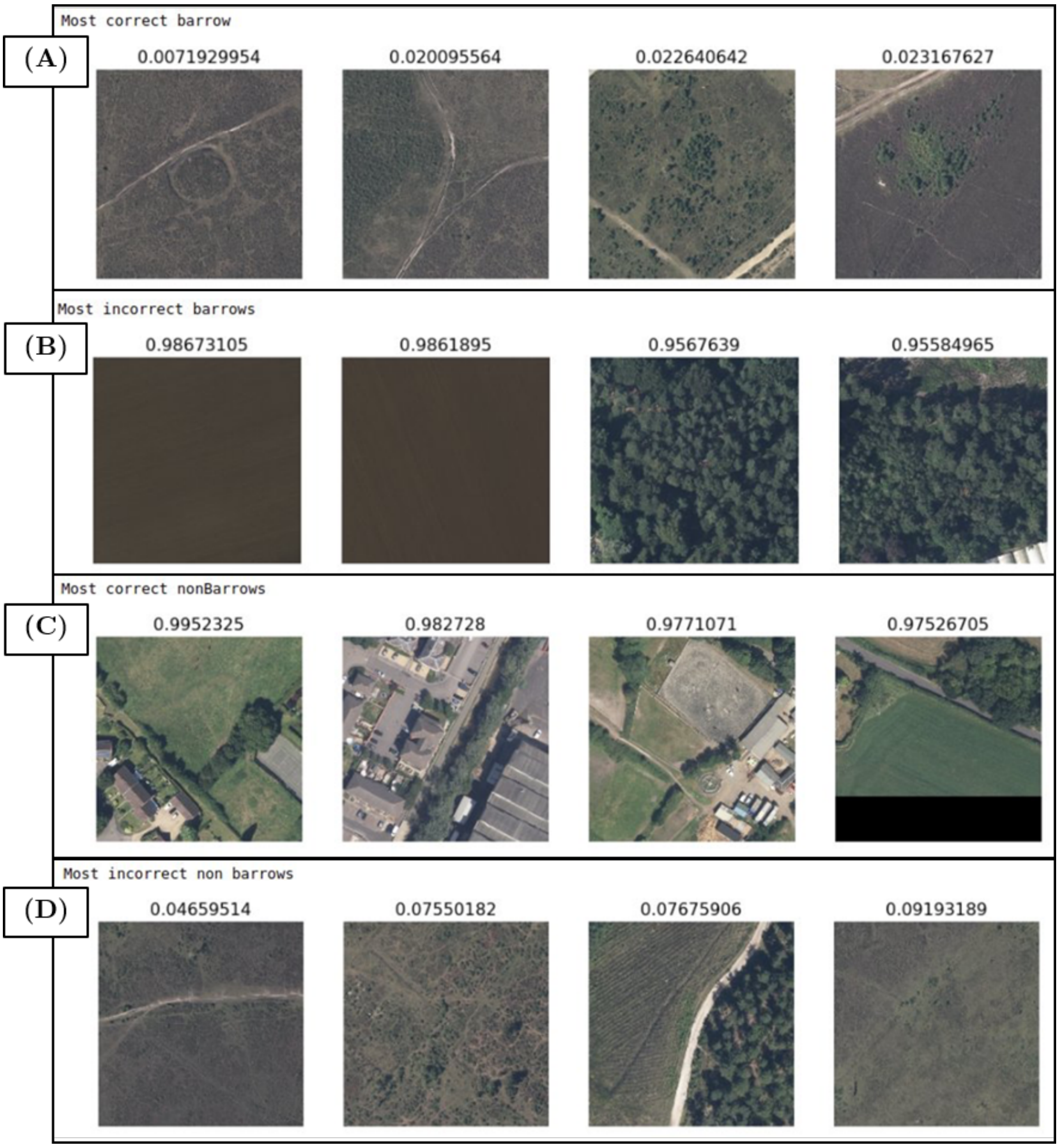


Figure 3.9: Results from experiment with RGB aerial images showing (A) the most correct barrow, (B) most incorrect barrow, (C) most correct background, (D) most incorrect background. ©Crown copyright and database rights 2020 Ordnance Survey

the result and try to further reduce the data complexity with visualisation techniques that highlight local archaeology. Just like the previous experiments, this reduces the complexity and should help the model to converge faster and prevent overfitting. For this experiment we used a ResNet50 that was pre-trained on ImageNet weights. This network required an input of 3-band imagery and thus we chose a multi-directional hill shade which combines 3 hill-shade images that illuminate the image from 3 different directions at a highly oblique angle (225°, 270°, 315° azimuth). The model was trained for 50 epochs and reached an accuracy of 0.8133 which is still not as good as we hoped so we again reviewed the results visually. In the top row of Figure 3.10 again the true positives are shown where the model was most certain that the image showed a barrow.

In these images the barrows are really prominent, the majority show multiple barrows and also a strong ditch surrounding the barrow. In the second row we have depicted the false negatives where the model was most certain that the image belonged to the background class. The first barrow is probably plough levelled or removed and a second other barrow that damaged by a ditch seems visible at the top. The second is also destroyed and the third has been heavily ploughed out. The final barrow is small but should have been found, the model is probably confused by the other disturbances. To improve the detection rate, we could crop the images to the actual object size. In the third row we show the most certain true negative predictions which show very irregular shapes and texture that are mainly modern. The final row shows the false positives where the model was most wrong and sure they were barrows. The first image are modern circular silos, the second is a small hill of some sort but not a barrow, the third might be archaeological nature but is not a barrow and the final image might actually show three barrows that were not known before or not in our labelled dataset. Overall each of these results have sensible explanations and we have mainly learned that the dataset is noisy which we can improve. To further understand if the model is genuinely "seeing what we see" we have also experimented with a Class Activation Map on one of our barrow images (Figure 3.11). To make this image first a gradient image is created using the weights of the second to last layer in the network. The weights for barrow have been overlaid on our image which provides an insight of the region where a CNN is looking to classify the barrow. Again, we can confirm that the model is finding the right pattern.

As we found that the main issue was our noisy dataset, we manually improved it over several days. We used our improved DTM dataset to further research the effect of DTM visualisations. Most visualisations are greyscale and in order for the pre-trained ResNet50 to work we concatenated the same image three times along the channel axis before feeding it to the network. We trained the models for 50 epochs and report on the best validation accuracy. The results are shown in 3.3. Open Positive visualisation works best and Hillshade performed the worst. With Hillshade the direction of the light source changes the appearance of a barrow depending on the size which might have an adverse effect on the learned pattern. To further research the effect of visualisations we experimented with different visualisation combinations to fill the 3-band image that feeds into the ResNet50. The results in 3.4 show that this improves the overall accuracy and mostly improves the result for a combination of Open Negative, Slope and Open Positive. The Hillshade visualisation combinations produces the lowest accuracy.

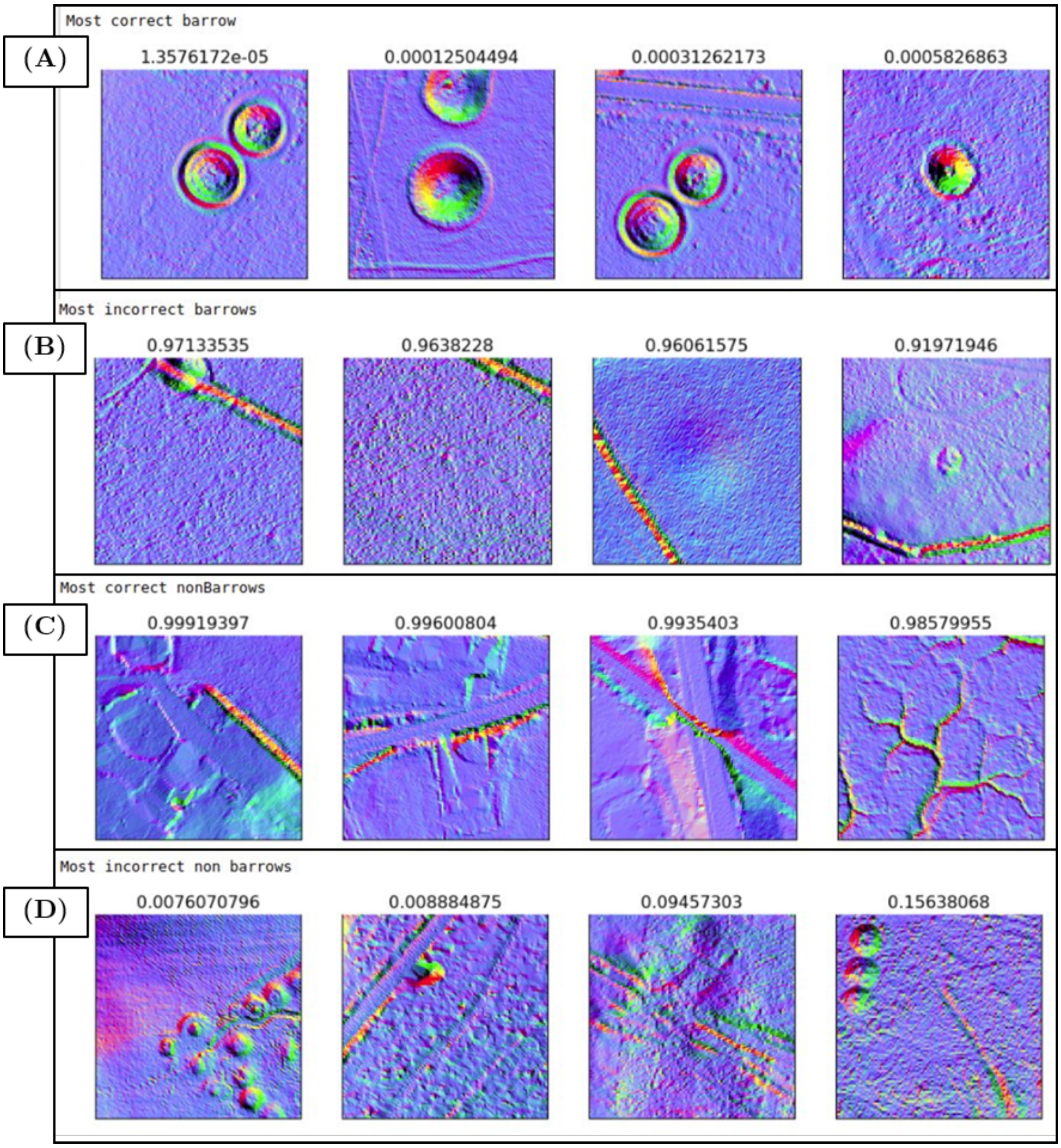


FIGURE 3.10: Results from experiment with DTM 1 m resolution images showing (A) the most correct barrow, (B) most incorrect barrow, (C) most correct background, (D) most incorrect background

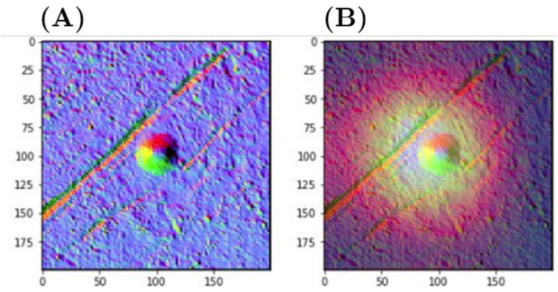


FIGURE 3.11: Class Activation Map (brighter yellow indicates the region is more "barrow-like").

Table 3.3: Validation accuracy for each DTM visualisation.

Visualisation	**Validation Accuracy**
Hillshade	0.85
Open Positive	0.96
Open Negative	0.95
Slope	0.89
Sky View Factor	0.91

Table 3.4: Validation accuracy for each combination (early fusion) of DTM visualisations

Visualisation	**Validation Accuracy**
Open Negative, Sky View Factor, Open Positive	0.9739
Open Negative, Sky View Factor, Slope	0.9674
Open Negative, Sky View Factor, Hillshade	0.9587
Open Negative, Slope, Open Positive	0.9761
Open Negative, Slope, Hillshade	0.9609
Open Negative, Hillshade, Open Positive	0.9674
Sky View Factor, Slope, Open Positive	0.9609
Sky View Factor, Slope, Hillshade	0.9609
Sky View Factor, Hillshade, Open Positive	0.9435
Slope, Hillshade, Open Positive	0.9609

3.3 Object Detection

After having trained a good image classifier this can be used to localise objects across all the available images. Traditionally this has been done using a sliding window approach where one gets a probability for each path that the sliding window comes across. Although accurate, this is done at a very high computational cost. Recently, there have been many improvements with methods such as region-CNN (R-CNN), You Only Look Once (YOLO) and Single Shot Detection (SSD). Of these methods, R-CNN seems most accurate but YOLO and SSD are most efficient (Liu et al., 2016). As we have a very large area to cover and are merely testing object detection we have chosen an SSD for now.

3.3.1 Approach

Ideally, we would use the specialised trained networks of the previous chapter to apply to the object detection problem. However, for now, we have sought a more basic approach by implementing the open source Raster Vision API (Azavea, 2018). This API has a variety of functionalities to make training data, train models, make predictions, and evaluate created models for object detection. It's especially useful to have their functionalities as they have designed their API to work efficiently on very large GeoTIFF files and on objects that are sparsely located within the images. Additionally, they can process the input of objects located with geospatial coordinates using GeoJSON files and output predictions in the same format which makes it easy to implement in a workflow. Their object detection itself uses the TensorFlow Object Detection API with an SSD approach using the MobileNet CNN pre-trained on COCO dataset (Huang et al., 2017). If we choose to continue using Raster Vision in the future we can change the approach, CNN and pre-trained weights accordingly.

3.3.2 Data Pre-processing

For input data the API requires training, validation regions and their respective GeoJSON files with coordinates for the object squares in the images. The API will further create training data by cropping the regions into smaller patches of 300 × 300 and translating the coordinate system of the patch and labels to a local system. For data augmentation we use a horizontal flip and random crop. For this experiment we use a DTM multi-directional hillshade visualisation (Figure 3.12). This experiment was undertaken before the noise was removed from the dataset which reflects in the results.

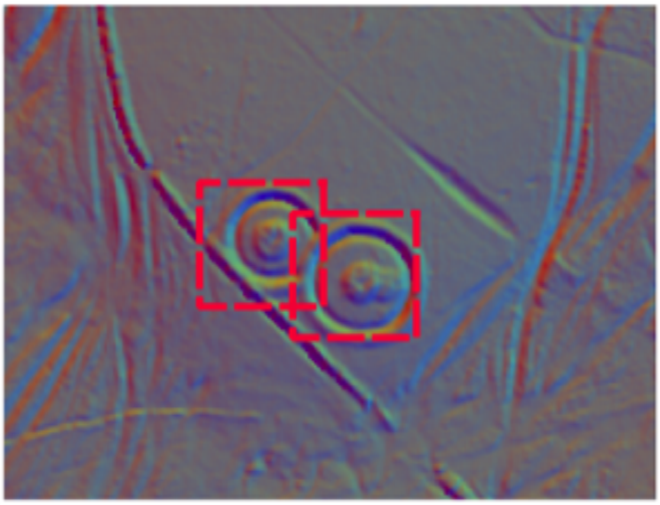


Figure 3.12: Multi-directional hillshade overlaid with the bounding box locations around known barrow site locations.

3.3.3 Results and Evaluation

There are various hyper-parameters that we could tune in the API but for the sake of this initial experiment we have kept most of the default settings. We have experimented with various 3-band combinations for RGBN and different DTM derivatives. We found that the RGBN results did not fluctuate much with different combination so, for now, we'll report only on the RGB results. We have experimented with different score thresholds and eventually chose a rather low threshold of 0.4 which means we should expect low

Table 3.5: Recall, precision and F1 result for DTM and RGB data model.

	Recall	Precision	F1-Score
DTM	0.180	0.333	0.234
RGB	0.034	0.013	0.019

recall but also allows to detect more barrows. Going further we should find how we wish to trade off true positives/false positives ratio's and set an appropriate threshold accordingly. In total we had 299 detections for RGB and 98 for DTM. In Table 3.5 we show the results of precision, recall and f1 scores on both data modalities.

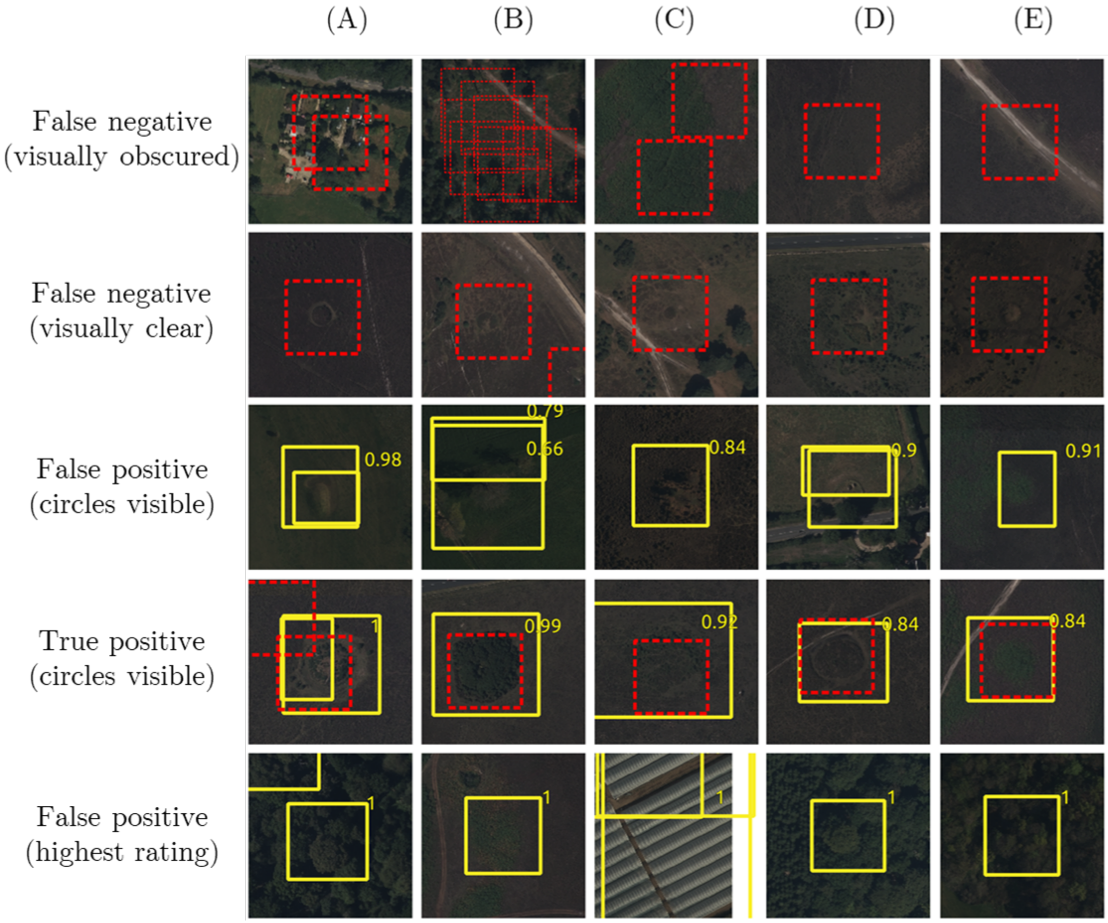

(A)
(B)
(C)
(D)
(E)
False negative (visually obscured)
False negative (visually clear)
False positive (circles visible)
True positive (circles visible)
False positive (highest rating)
0.98
0.79
0.56
0.84
0.9
0.91
1
0.99
0.92
0.84
0.84
1
1
1
1
1

Figure 3.13: Object detection results for RGB (yellow) compared to known barrows (red). From top layer at each data modality: missed barrows that are obscure to the human eye, missed barrows that are visually recognisable, interesting detections, sample of rightfully detected barrows, most likely detections. ©Crown copyright and database rights 2020 Ordnance Survey

For RGB images we gained rather poor results; out of 84 barrows in the validation area it detected 8. This may seem surprising as we previously gained accuracies to over 80%. However, those experiments had an equal split of background vs object examples whereas now we have the challenge of detecting barrows against a large amount of background examples. Seemingly the objects are not distinct enough or include too much noise for the classifier to extract a robust pattern. To analyse the reason for the poor results we

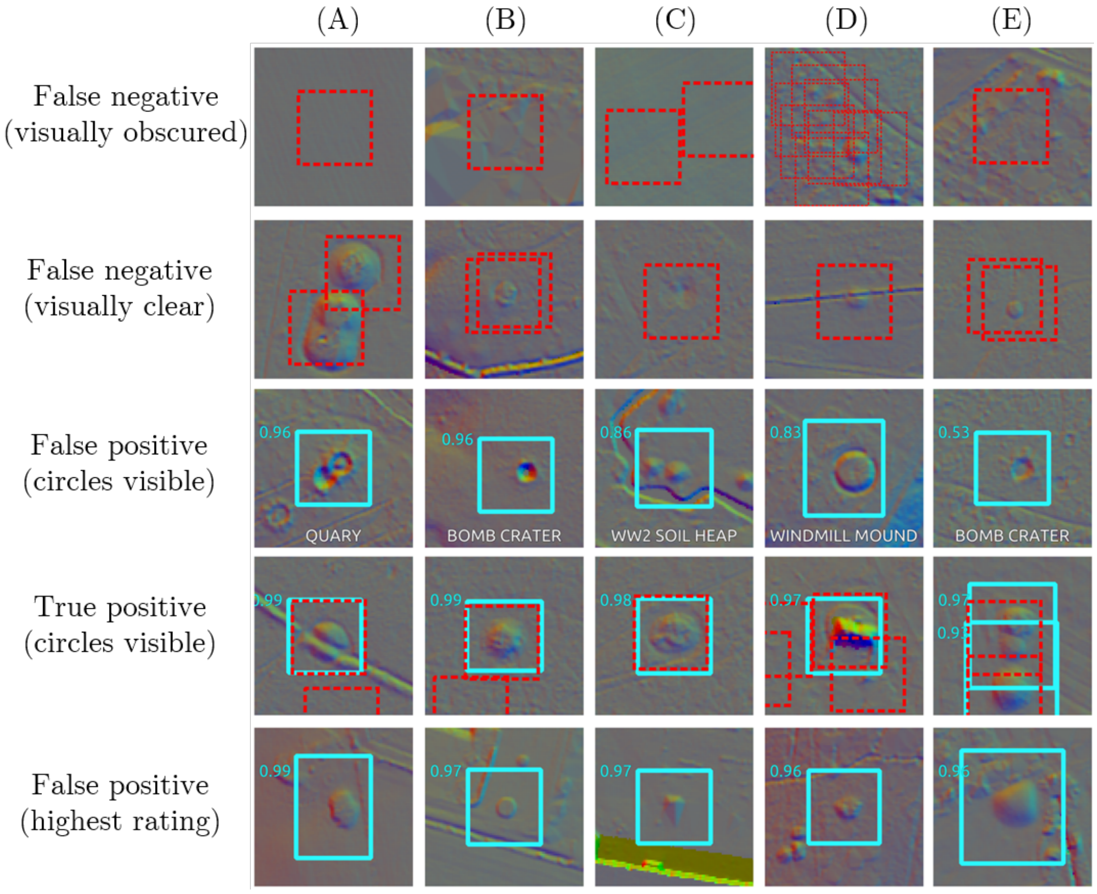

(A)
(B)
(C)
(D)
(E)
False negative (visually obscured)
False negative (visually clear)
False positive (circles visible)
0.96
QUARY
0.96
BOMB CRATER
0.86
WW2 SOIL HEAP
0.83
WINDMILL MOUND
0.53
BOMB CRATER
True positive (circles visible)
0.99
0.99
0.98
0.97
0.97
0.9
False positive (highest rating)
0.99
0.97
0.97
0.96
0.95

FIGURE 3.14: Object detection results for DTM (blue) compared to known barrows (red). From top layer at each data modality: missed barrows that are obscure to the human eye, missed barrows that are visually recognisable, interesting detections, sample of rightfully detected barrows, most likely detections.

have taken a closer look at the kinds of objects that were detected. In Figure 3.13 and Figure 3.14 we show an extensive comparison between detections and known barrows. From this we note that the quality of the known barrows is poor; some barrows do not show a distinctive pattern and others have been destroyed for construction of houses or roads. There are also barrows are surprisingly not picked up on, possibly because they are too small. To avoid this, we could create bounding boxes that perfectly fit around the barrow. Another insight from the box detection is that we see a lot of variation in length and width which does not appear in barrows. We might have to constraint the proposal of boxes to be a perfect square with a range of sizes. The false positives prove most insightful as to what the classifier has learned. Some show imaginable parts of barrows such as a circular object in the centre of the box, mostly trees or low vegetation. In other cases, mainly over forest or houses, the false positives do not have any resemblance to a barrow. Another interesting pattern is that we see high concentrations of detections over certain distinct types of agricultural fields (Figure 3.15). This is most likely caused by similar examples in the training dataset that do not visually show a barrow. Our main conclusion from the experiment on RGB data is that we need to re-evaluate the barrow dataset to remove the obvious destructed barrows and experiment with some of the more ambiguous types. Despite the misleading detections we noted some possible new barrows that require further inspection from an expert and possibly a field visit.

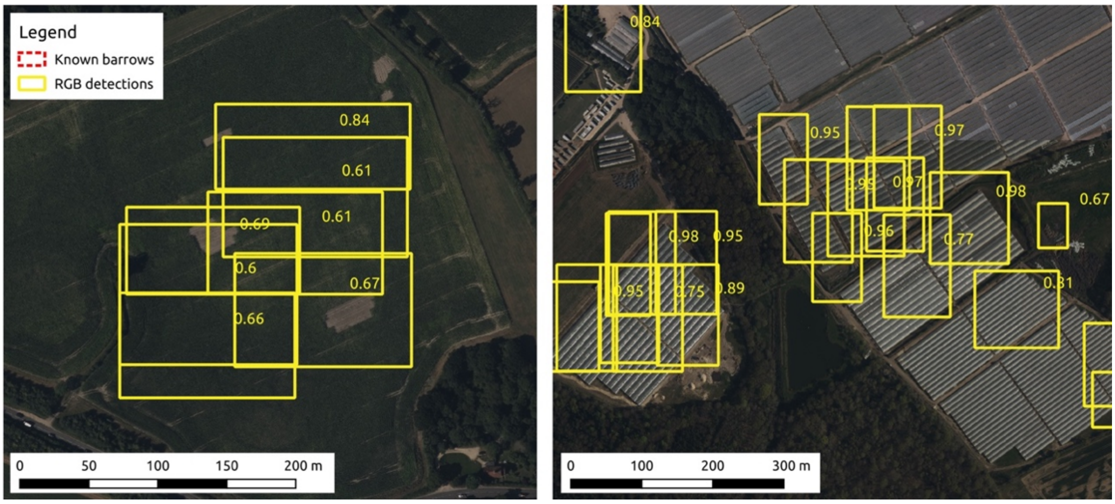


FIGURE 3.15: Clusters of RGB detections over distinct agricultural fields. ©Crown copyright and database rights 2020 Ordnance Survey

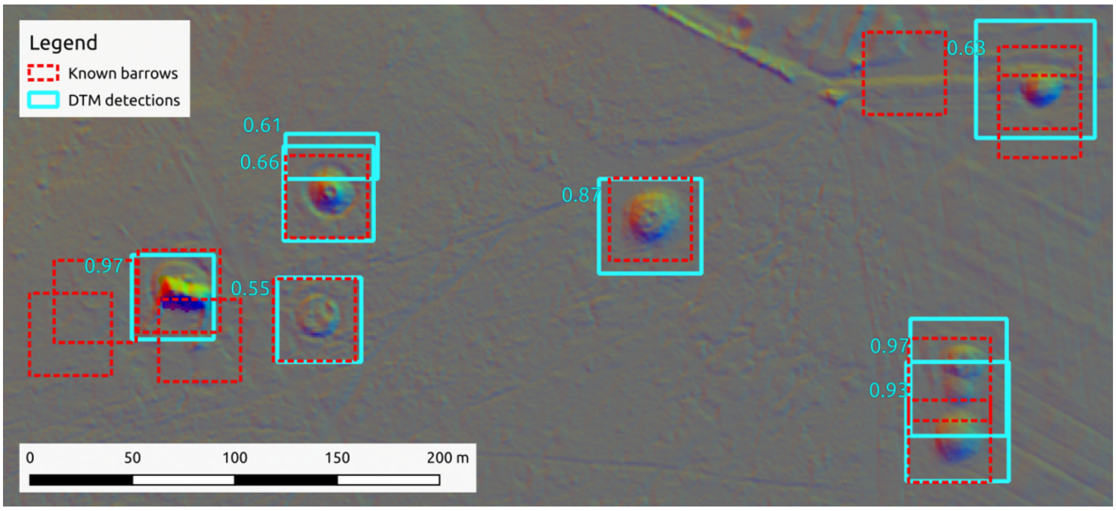


FIGURE 3.16: DTM detections including damaged barrow on the left. The more prominent barrows are detected whereas the levelled or otherwise severely destructed barrows are not.

DTM results are more interesting with a low recall against a somewhat high precision. We may conclude that out of all known barrows we detect relatively few (low recall) but out of all detections a high proportion are known barrows (high precision) (Figure 3.16). In this case we detected 55 barrows out of 143. From the known barrows we note that some have been levelled due to agriculture, some are destructed for other reasons and some show poor interpolation of the DTM possibly due to buildings or impenetrable forest. Same as with RGB we did not manage to detect very small barrows (see example in Figure 3.17) and, in this case, we also failed to detect an unusually large barrow. We again see a few odd size rectangles of false positives, so, further tuning of relevant

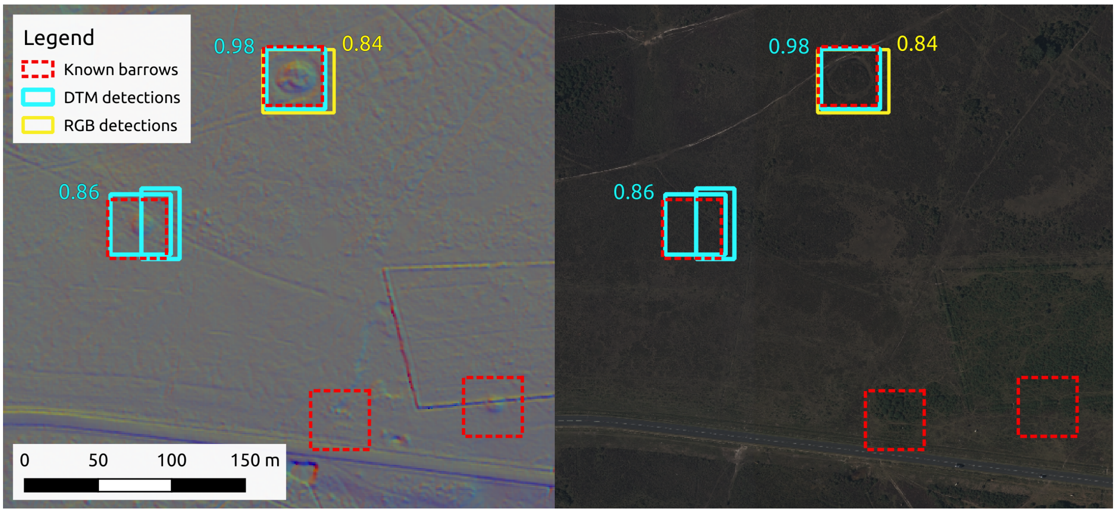


FIGURE 3.17: Overlapping DTM and RGB detections where both detect most prominent barrow.

hyper-parameters is needed. Amongst the false positives we found some notable objects such as bomb craters, post medieval quarries, and a windmill mound. On the DTM derivative all false positives show similar patterns to barrows. However, bomb craters have the inverse pattern of a mound, so we argue that this is caused by our choice of DTM visualisation which could be avoided when using the raw DTM or a more suitable visualisation. Notably, of the many true positives we see a robustness to quite significant destruction. Amongst the DTM false positives there seem to be promising detections that need further analysis.

Most observations for improvements to the object detection overlap between the RGB and DTM datasets and focus on tuning hyper-parameters in the API and improving the dataset of known barrows and their visibility on the data and within the boxes.

3.4 Conclusions and Discussion

We have shown that we were able to successfully train CNNs for both image classification and object detection on both LiDAR derived DTMs and multi-spectral aerial imagery. Somewhat surprisingly, DTM derived visualisations preformed better than the raw and more detailed greyscale DTM. Based on those results we have learned that transfer learning is one of the most important approaches to include when training on a small dataset. With a larger dataset or when using transfer learning with a CNN that is pretrained on a similar dataset we expect that the raw and detailed DTM will outperform the visualisation.

Our main challenge in this case study has been the noise in the dataset. The known sites were derived from local HERs which include legacy data which has meant that many sites were no longer present. Similarly, many sites were not visible on one or both of the remote sensing data used because of the time of year, forest canopy or because of ploughing. Because we are not aerial archaeology experts we have chosen not to improve the dataset but to look for another clean dataset which will be presented in chapter 4.

Chapter 4

The Arran Case Study

This chapter presents a case study on the Isle of Arran where we further address the optimisation of approaches on small datasets. The case study was offered by HES who have used the Isle of Arran in multiple case studies to showcase different approaches for improved mapping of archaeological sites. The known site locations have been mapped over a short period and on the same LiDAR dataset that we will use in our case study. This overcomes the major blocker for our previous chapter. HES also offered to provide continuous feedback and discussions on the approach and results which have made this a collaborative case study between local experts and computer scientists. This collaboration has contributed a lot to the optimisation of the approach to the local conditions.

We will first introduce the dataset and case study area in section 4.1. We then move into our image classification approach in which we experiment with different hyperparameters. In section 4.2 we use these optimised settings for the dataset to train our object detection approach. Our image classification approach follows the experiments that we optimised in section 4.3. The object detection approach introduces a CNN called RetinaNet which is optimised for sparse, small datasets. This network has further improved the outcome and contributed to the detection of many previously unknown archaeological sites.

4.1 Dataset

Arran lies in the west of Scotland and is known at HES as 'Scotland in miniature' because it has a range of landscapes from highlands to lowlands that are generally representative of the rest of Scotland. Arran is being used by HES to develop approaches to rapid large area mapping using remote sensing datasets (Banaszek et al., 2018; Cowley et al., 2020; Cowley and López-López, 2017). Manually they have optimised their approach on the 432 km^2 island systematically using specific DTM visualisations, orthophotographs, and supporting information such as 19th century maps. They were able to achieve an average coverage of 90 km^2 per day in their desk-based approach. The results from the different experts were gathered and analysed to select the least confident sites for field verification. They only looked at the less confident classifications because they are very confident that LiDAR is a reliable source and a field visit doesn't always add more information about commonly known sites (e.g. Figure 4.1). This consideration is an important part of their strategy for rapid mapping.

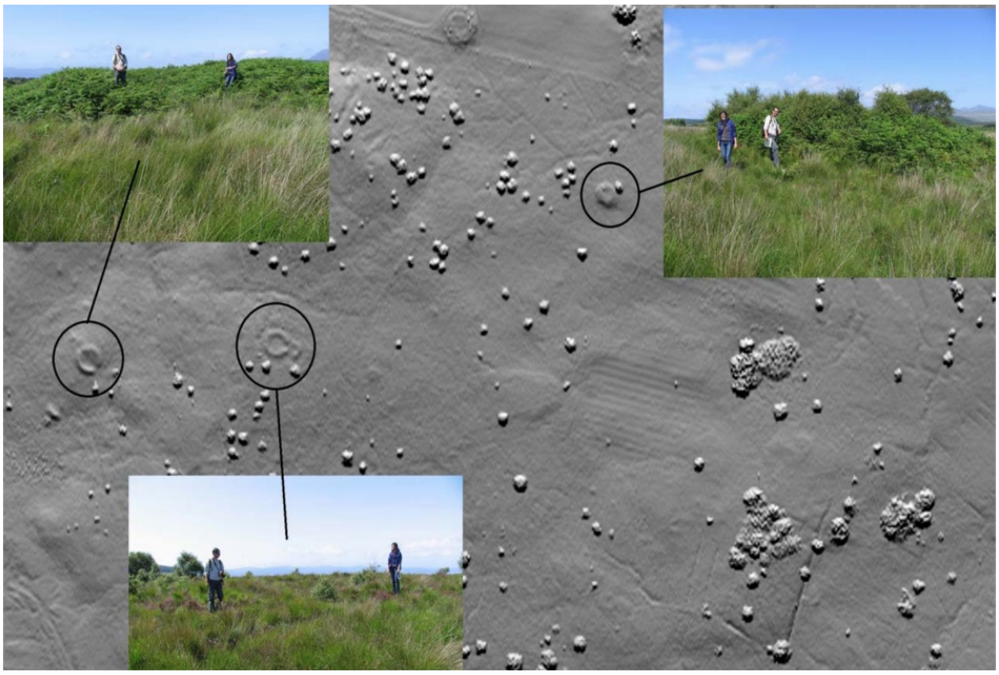


Figure 4.1: LiDAR derived hillshade visualisation from a DEM with insets showing the round houses on the ground during a field survey. Despite uneven vegetation, the LiDAR effectively captured most archaeological remains hidden by the vegetation. Images reproduced from Cowley and López-López (2017).

4.1.1 Site locations

The HES survey has more than doubled the number of known archaeological sites from on the Island that are available from the National Record of the Historic Environment (NRHE). The new discoveries include sites in what today are remote locations, such as the tops of valleys, but also in areas of dense known site distributions. The site locations that we are looking at in this study are prehistoric roundhouses (203), shieling huts (transhumant grazing) of medieval or post-medieval date (344), and small field clearance cairns (403) which are remains from agriculture (Figure 4.2, Figure 4.3). Finding such sites helps HES to understand the pattern of prehistoric settlement on the island, and the use of upland grazing in medieval and more recent times. Round houses appear as circular wide doughnut rings with opening(s) and range from 8 m to 15 m in diameter. Shieling huts and small cairns are smaller at 2–6 m across. Shieling huts are also doughnut shape but are less wide than roundhouses. Small cairns are small mounds. We expect shieling huts to be a confusion object to round houses because they have the same ring structure but have a different size, and also to small cairns because they are both circular and have the same size. Other archaeological confusion objects that are known on Arran include burial cairns, burned mounds, enclosures, kilns, rectangular buildings, and horse platforms (Figure 4.4). There are also several modern confusion objects including cattle feed stances and sand bunkers and tees in golf courses. Also, there are geological confusion objects like glacial drumlins and peat erosion mounds that look like small cairns and shieling huts.

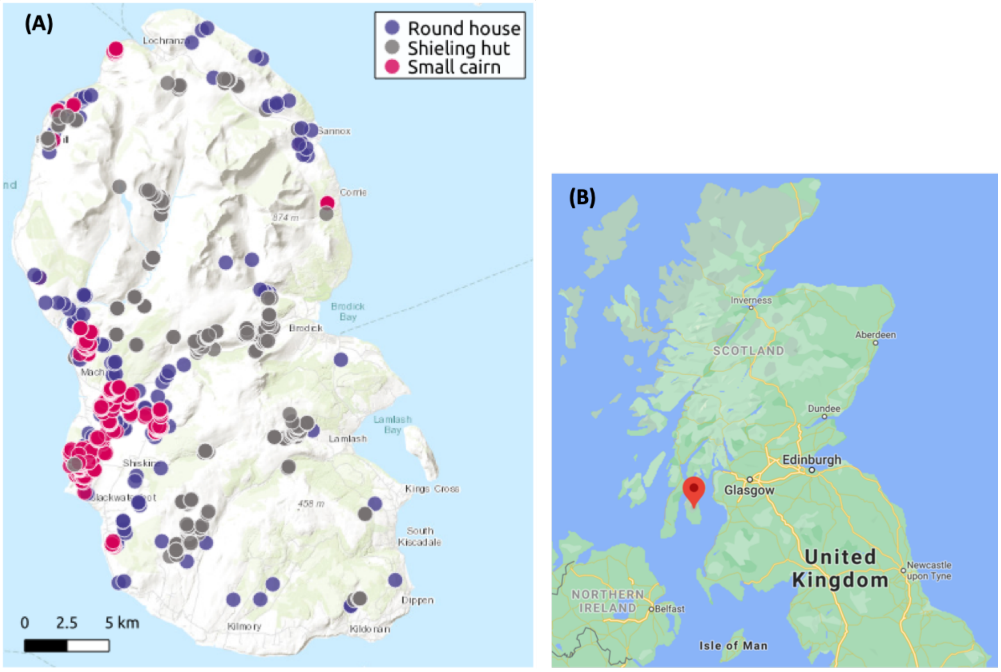


FIGURE 4.2: (A) a map of Arran with the site distribution of round houses, shieling huts and small cairns. ©Crown Copyright, ©Historic Environment Scotland. (B) shows the location of the Isle of Arran in the United Kingdom with a red arrow (GoogleMaps, 2020a).

4.1.2 Airborne LiDAR

The LiDAR data used in this case study are from the Scottish Remote Sensing Portal, a partnership between the Scottish Government and the Joint Nature Conservation Committee. The average 'ground' point density per square metre was 2.75, but varies from 0.43 to 7.44 depending on vegetation density and the presence of buildings. The LiDAR data was processed into a DTM at 0.5 m spatial resolution and is provided in 16-bit unsigned integer format. Along with the DTM Several pre-processed DTM visualisations were supplied (Figure 4.5). In some areas of Arran dense coniferous plantations obscured the ground which created gaps in the data (Figure 4.6) which for single trees can create circular confusion objects.

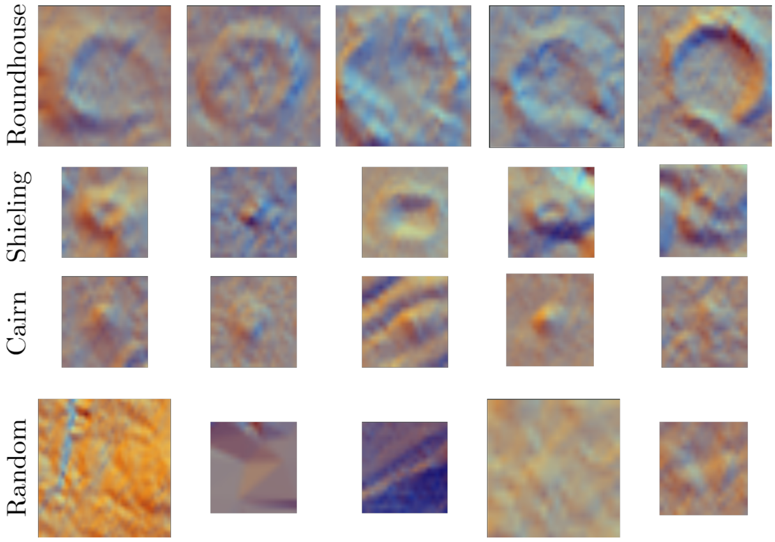


FIGURE 4.3: Examples of each class that is used in the Arran case study (shown on Multi-directional hillshade). Top row: Roundhouses. Second row: Shieling huts. Third row: Cairns. Fourth row: Random (background). ©Historic Environment Scotland.

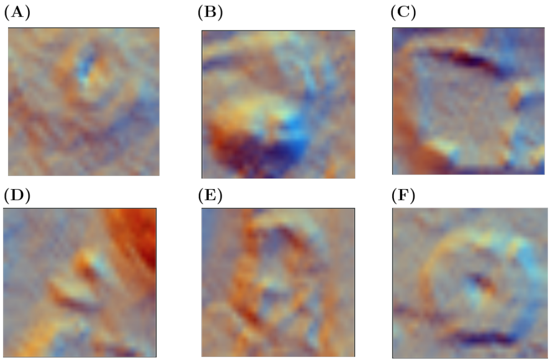


FIGURE 4.4: Archaeological confusion objects (shown on Multi-directional hillshade). (A) burial cairn, (B) burned mound, (C) enclosure, (D) kiln, (E) rectangular building, (F) horse platform (all 20x20 meter images). ©Historic Environment Scotland.

4.2 Image classification

The approach to image classification is the same as the New Forest case study. We found it very useful to apply image classification as a first step to better understand the dataset and make our hypothesis for the object detection stage. For this brief case study we used multi-directional hillshade.

4.2.1 Data pre-processing

Similar to the New Forest case study the data is provided as shapefiles with central points of every site which we have used to create images around each point. For the round houses we created 20x20 meter so 40x40pixel images, and for shieling huts and small cairns we created 10x10 meter so 20x20pixel images. We created 316 random

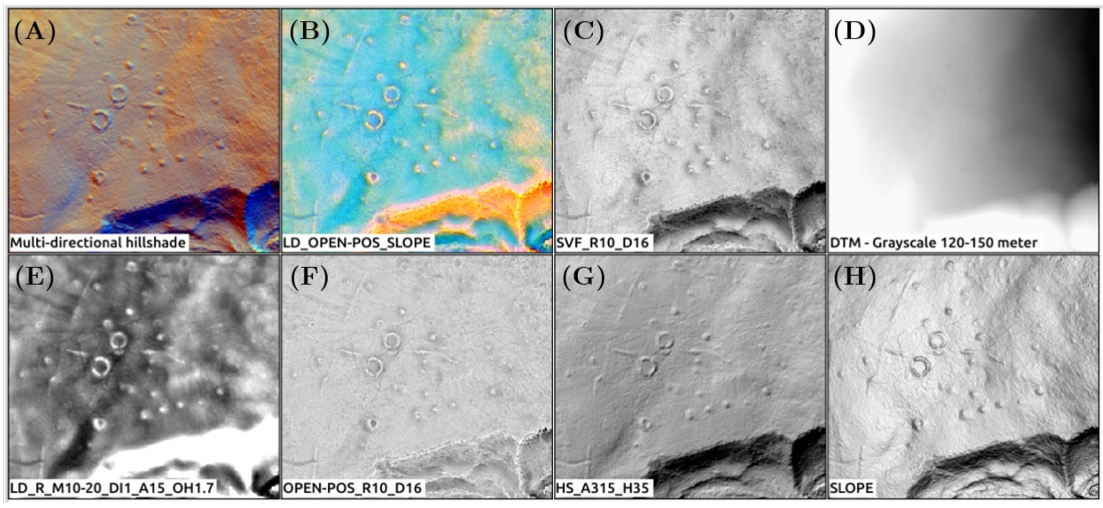


Figure 4.5: DTM visualisations supplied by HES. (A) Multi-Directional Hillshade, (B) RGB-combination of Local Dominance, Open-Positive and Slope, (C) Sky View Factor, (D) DTM, (E) Local Dominance, (F) Open Positive, (G) Hillshade, (H) Slope. The image also shows the parameters used to create each visualisation. ©Historic Environment Scotland.

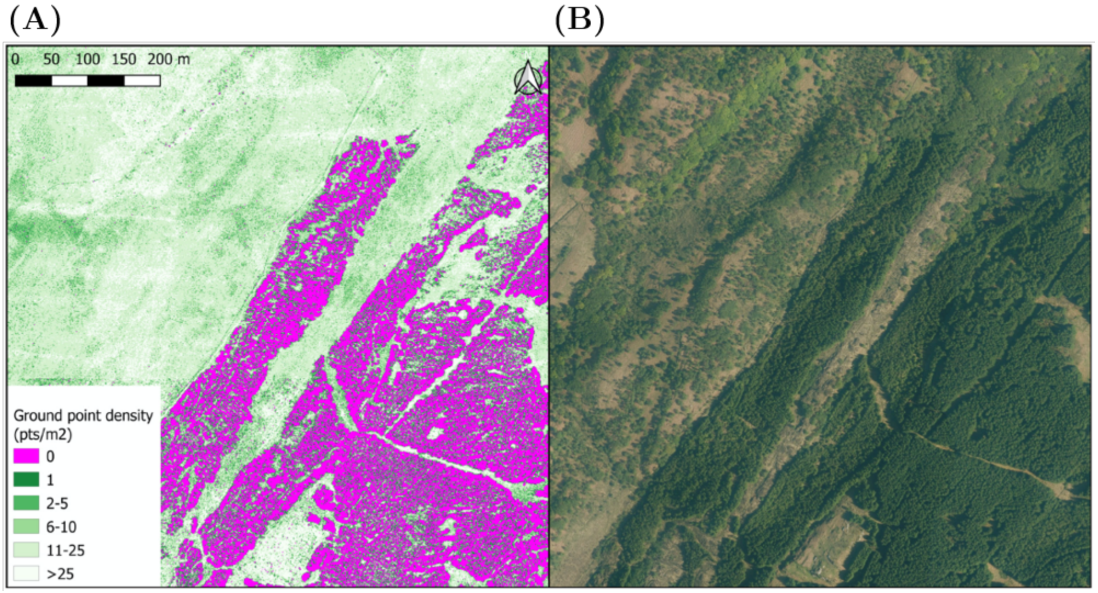


Figure 4.6: (A) ground point density and (B) aerial image of a coniferous plantation in Kilmartin Glen (Western Scotland mainland). The LiDAR data was captured at high point density but still wasn't able to penetrate the dense forest cover. This area shows similar forestation to Arran. Image reproduced from Cowley et al. (2020).

images that were randomly cropped at 40x40 and 20x20 pixels (Figure 4.3). The split between training and validation images in again made geographically, this time between with the training data in the south and validation data in the north. Different from the New Forest case study, augmentations in this research were applied in Keras. We only used random flips and 90 degree rotations.

4.2.2 Experiments

Based on the New Forest experiment and brief trial and error on our current dataset we found that a pre-trained ResNet50 was able to obtain high accuracy for all the classes. The main confusion has been between shieling hut and small cairns where it is also challenging to see the difference by eye (Figure 4.7). Also, the fainter objects have sometimes been classified as background/random class. Only very few objects that were random have been classified as objects, either because of a circular appearance or other mount structure.

We also experimented with the different image visualisations that were available to us. This showed that overall the Open-Positive and Local Dominance were best followed by Slope and Multi-directional hillshade (Table 4.4). Because Open-Positive, Local Dominance and Slope were the best performing single band visualisations we concatenated those images into and RGB-image called LD_OPEN-POS_SLOPE for further analysis (Figure 4.5 (B)).

Based on these results we feel comfortable that the dataset is sufficient. We required no further data cleaning or different LiDAR visualisations.

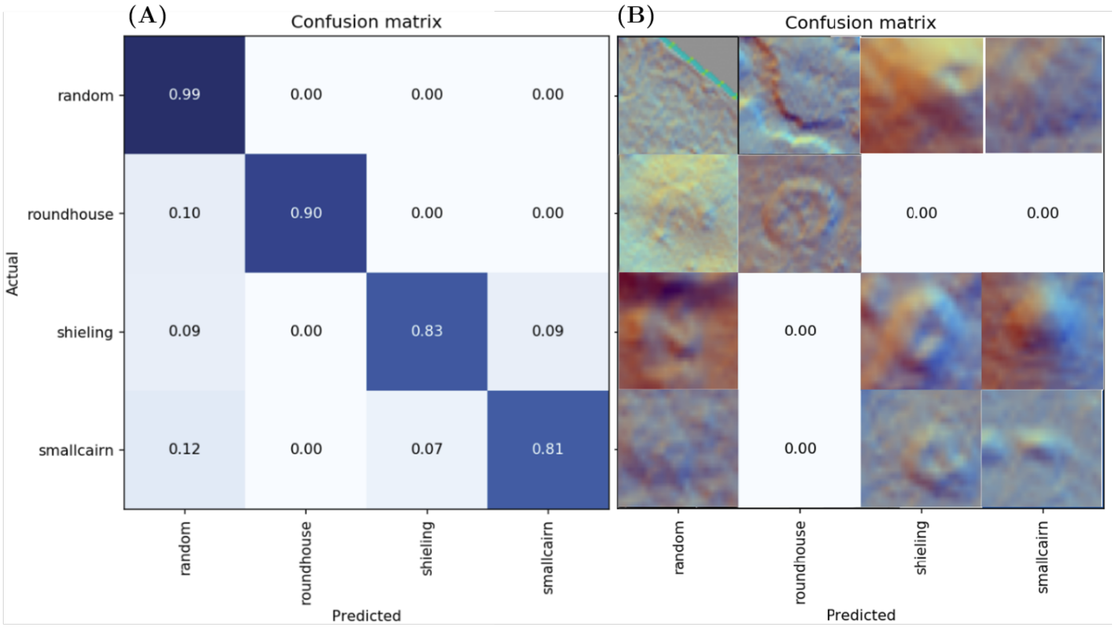


FIGURE 4.7: (A) confusion matrix from image classification depicted in percentages, (B) images that represent the confusion objects. ©Historic Environment Scotland.

4.3 Object detection

During our preparation for the second case study we reviewed the suitability of the SSD algorithm which we previously applied in section 3.3 and compared this with

Table 4.1: Image classification results showing the validation accuracy from different LiDAR visualisations.

LiDAR visualisation	**Validation Accuracy**
Slope	0.78
SVF	0.75
Open-positive	0.89
Local dominance	0.86
DTM	0.43
Hillshade	0.76
Multi-directional HS	0.78

new research on object detection. We found that one-stage detectors like SSD have a foreground-background class imbalance problem. These detectors evaluate hundreds of candidate locations per image but only a few locations contain objects which is especially problematic in our case study where we have a very sparse dataset with only a few foreground examples against a lot of background. The easy negative/background examples can overwhelm training and lead to worse performance of the models. Since our initial case study, a new approach in archaeology was presented by Verschoof-van der Vaart and Lambers (2019) who successfully applied a Faster Region-CNN (Faster R-CNN). In two-stage detectors such as Faster R-CNN, the first stage, region proposal network narrows down the number of candidate object locations which filters out most of the background. In the second stage, classification is performed for each candidate object location. At this stage the class-imbalance is further addressed with sampling heuristics which is implemented by a fixed foreground-to-background ratio (1:3) per minibatch. With further research into the issue we found that Focal Loss is another, improved approach to address the issue organically. The Focal loss function down-weights "easy" negative examples and thus focuses training on "hard" negatives, which improves the prediction accuracy. This concept was introduced by Lin et al. (2017) and is accompanied with a CNN called RetinaNet. This is a one-stage detector that uses ResNet and a Feature Pyramid Network (FPN) as backbone for feature extraction, plus two task-specific subnetworks for classification and bounding box regression (Figure 4.8). The hierarchical FPN merges information from different scales. This cross-scale learning is critical for archaeological objects that vary in size and where context is important.

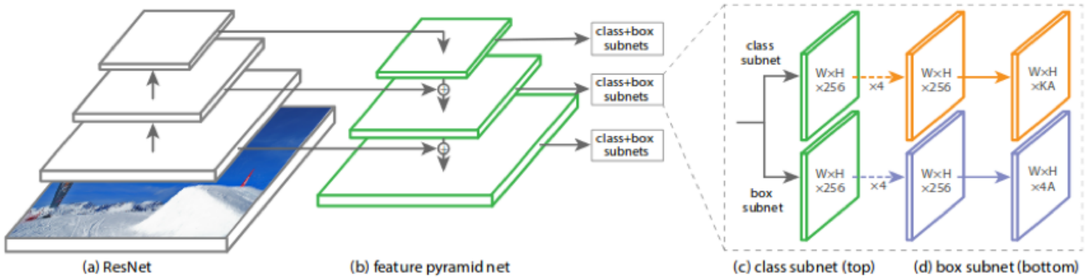


Figure 4.8: ResNet architecture with (A) ResNet, (B) Feature Pyramid Network, (C) the two task-specific sub-networks for classification and (D) bounding box regression. Image reproduced from Lin et al. (2017).

4.3.1 Approach

When we started this experiment there were not any specialised geographical approaches that implemented or utilised RetinaNet. Instead we used the Fizyr implementation of RetinaNet in Keras (Gaiser, 2019). This implementation is widely used because it has very good documentation, debugging tools and options to change hyperparameters. We used most of the implementation's default hyperparameters and will describe in the next sections where we altered the code for our domain.

4.3.2 Data pre-processing

The implementation we used inputs a csv-file with links to the images, box coordinates and class names (`path/to/image.jpg,x1,y1,x2,y2,class_name`). Each line contains only one box annotation and for images without objects the coordinates and class name remained blank. We wrote our own code to transfer our geographical coordinates into image coordinates and generate the csv-file with the required format.

To ensure that we did not mix training and validation data we divided the large DTM into areas that were 10 times the size of our input images. These area's were divided into our input images. Through experimentation we found that 500x500 pixels (250x250 meter) was the best input image size (Figure 4.9). With larger image size (1000x1000) our objects became too small (discovered through debugging option in the RetinaNet implementation), and with smaller image sizes (100x100, 200x200) the model struggled to detect our larger, roundhouse, objects. We further overlapped our input images by 10% to ensure that edges did not detract from detections. Another useful aspect of the debugging option in the RetinaNet implementation is the feedback on objects close to the edge of the image. Those objects would not be too small for a classification by the model which would reduce the accuracy. To overcome this, we updated our data preparation code to exclude all object annotations that were 10 pixels from the image edge.

We have split our data into 80% training and 20% validation areas. We experimented with North/South divide of training and validation areas and found that the best was South training and North validation (Figure 4.10). There is a large cluster of small cairns in the South West of the island which is difficult to break up into even training and and validation sets (Figure 4.2 (A)). When choosing the South side for our validation area only 14 small cairns would be included against 25 in the North side (Table 4.2). We also experimented with 75%/25% training/validation balance however we then found half of the small cairns to be in the validation area. To maintain enough small cairns in the validation data we selected the train/validation divide from Table 4.2 (A).

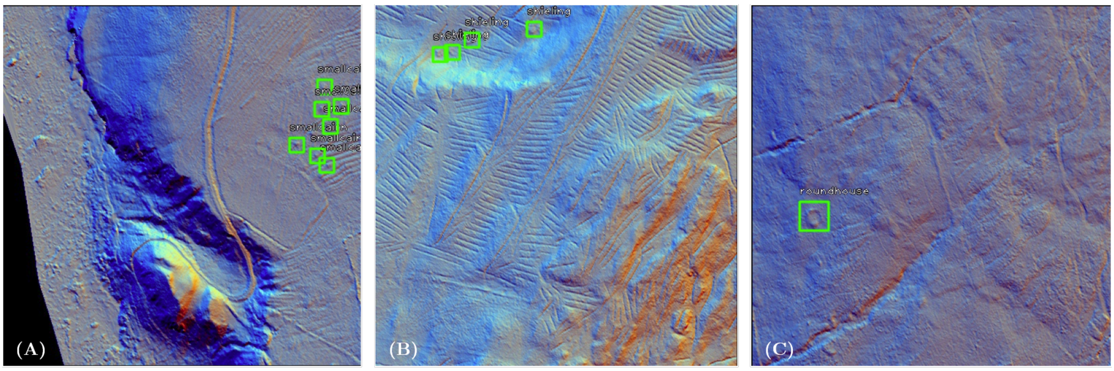


Figure 4.9: Images show the image pre-processing step from the Fizyr implementation of RetinaNet: debugging object detection input images (Gaiser, 2019). All boxes are green which means that the boxes are big enough for the image size and that they are not too close to the border.

Table 4.2: Separation of objects per class in Training and Validation sets for both North/South and South/North data splits. When using the South for validation data there are only 14 known cairn sites.

	Roundhouse	**Shieling**	**Small cairn**
South/North training	247	477	541
South/**North** validation	44	106	25
North/South training	252	472	552
North/**South** validation	39	111	14

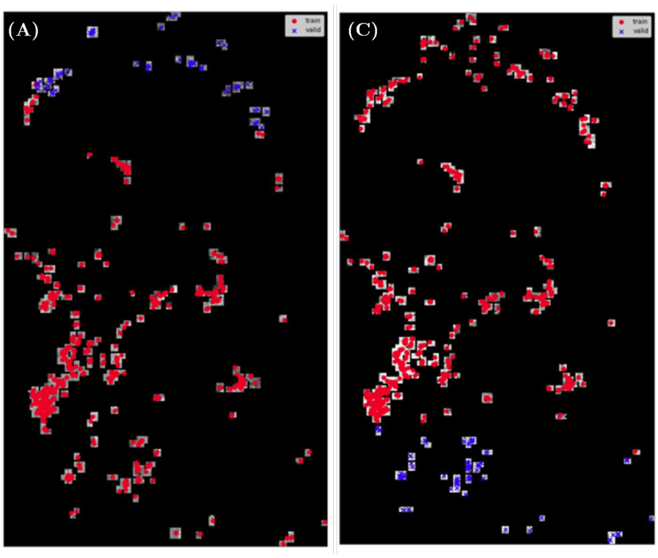


Figure 4.10: Visual representation of the North/South or South/North training/validation divide. (A) Shows the South/North divide of training and validation data. (B) Shows the North/South divide of training and validation data. ©Historic Environment Scotland.

4.3.3 Hyperparameter tuning

Even though we used most of the implementation's default hyperparameters, we did change some parameters and code for our domain adaptation.

Because many locations around an object will have high confidence many overlapping bounding boxes are created for each object. Non-Maximum Suppression (NMS) removes the boxes that overlap more than a given threshold (called Intersection over Union (IoU)), the box with the lower confidence value will be removed. The default IoU for the implementation is 0.5 which did not work well for us because of the relation between our object size and pixel resolution (Figure 4.11). Through experimentation we found that 0.20 was the optimal threshold for our case study.

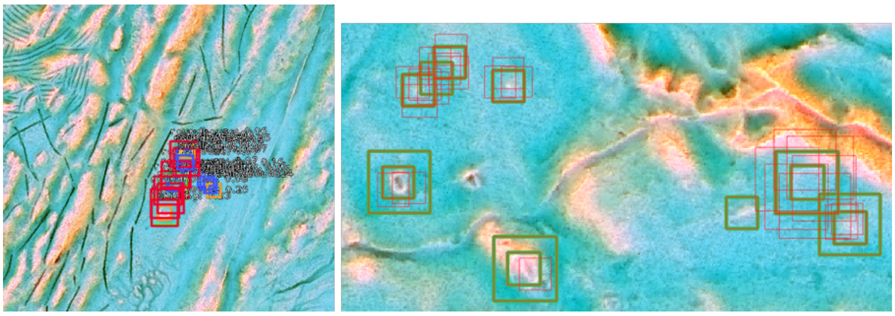


Figure 4.11: An evaluation of the NMS hyperparameter. (A) Shows the image result from RetinaNet implementation with the standard parameter of IoU 0.5. (B) Shows our review of different IoU values in GIS software with standard 0.5 (red) and our choice of 0.2 (green). ©Historic Environment Scotland.

The score threshold is a parameter that is often changed but we kept it at 0.05. We found that it was better to have more detections and accept a higher recall because the additional boxes which might represent new sites could be quickly verified by the experts.

At the evaluation stage the RetinaNet implementation generates bounding boxes with image coordinates so that those results can be reviewed per image. To view these results in our GIS system we transformed the image coordinates (created in `eval.py`) to geo-coordinates and saved those in a GeoJSON file. Because of the overlapping images we end up with multiple detections along the edges which are removed with an additional non-max suppression step. We reviewed the IoU threshold and found that 15% worked best. A lower IoU would have removed more of the overlapping boxes which could improve the precision but for objects that tend to cluster closely together like shieling huts it would reduce the recall. For a specialist it would be very simple to remove the extra boxes at the manual inspection stage and so we kept the IoU at 15%.

After prediction, we would generally import the GeoJSON file with the detections into our GIS project and analyse the results. Here we found that the boxes were again overlapping from the concatenation of the different images that were overlapped by 10% at the image pre-processing stage. We applied our NMS and reviewed those results again for precision and recall separately which were compared to a GeoJSON file that was created for the known sites with an attribute column showing their train/validation split.

Table 4.3: Results of object detection for different LiDAR visualisations. The results are shown for precision per class and the Mean average precision across the classes.

Validation Areas	Round House	Small Cairn	Shieling	MaP
Slope	**83**%	23%	51%	50%
SVF	79%	18%	36%	44%
Hillshade	73%	13%	40%	42%
Open-positive	78%	22%	**53**%	51%
Local dominance	78%	31%	15%	41%
Multi-directional hillshade	73%	**39**%	49%	**54**%

4.3.4 Experiment 1: Training on images that contain objects

In our first experiment we trained only on the images with objects in them. This allowed us to quickly experiment with hyperparameters and understand the result we could aim for when looking for new sites in the remaining images. This also reduced the chance of finding false positives from new/unknown sites. However, it also meant that the model was only trained on high quality data which wouldn't generalise well to areas where less or no sites are known.

Our first results are shown in Table 4.3. Similar to our image classification experiment we found that Slope and Open Positive overall perform really well. Interestingly Local Dominance is performing poorly shieling huts but exceptionally well on small cairns. The best overall performance was by the multi-directional hillshade so we based our further experiment on that visualisation.

For our next step we used the trained model to detect previously unknown sites on the images that were not used for training or validation. We visually inspected this data and requested feedback on our results from the HES expert. He confirmed that the approach caused many false positives but that they were reasonable and could also be made by an inexperienced image analyst. For example there were many small cairn detections made on areas where there is high peat depth. In such areas the false positives showed glacial drumlins and peat erosion mounds Figure B.2. To further understand these false positives we comparing this data to a map of peat depth that was supplied by HES (Figure B.1 (A)). The map is very generalized at 1:250,000 but still allowed us to analyse the pattern. We found that there aren't any known examples of small cairns at 1 and 1.5 peat depth but respectively 21 and 225 small cairns were detected (Table B.1). Another area of clustering false positives was found in modern built up areas such as urban areas, recreation areas (mainly golf courses) and coniferous plantations. To analyse the scale of the effect HES supplied us with a Historic Land-use Assessment (HLA) map (Millican et al. (2017), Figure B.1 (B)). We visually compared the false positives to these areas and indeed found that golf bunkers on a golf course were detected as round houses (Figure B.3). Again we also found that there aren't any known sites in the modern built up area which could be masked out in further analysis (Table B.2). We also found that there aren't many known sites in the “Agriculture and Settlement” class which does produce a lot of false positives such as a confusion between cattle feeders and round houses (Figure 4.12). An expert might want to exclude these areas in their analysis if they are interested in a very rapid mapping project. However, the areas close to the modern built up area are also most prone to destruction from development or land use. For our further analysis we didn't exclude those areas. Another suggested area to exclude were locations where low LiDAR point density caused by coniferous plantations created visible groups of points resembling "boulders" that were detected as round houses (Figure 4.13). We eventually didn't go through with this mask because the boulders weren't a significant issue and could be quickly disregarded by the expert.

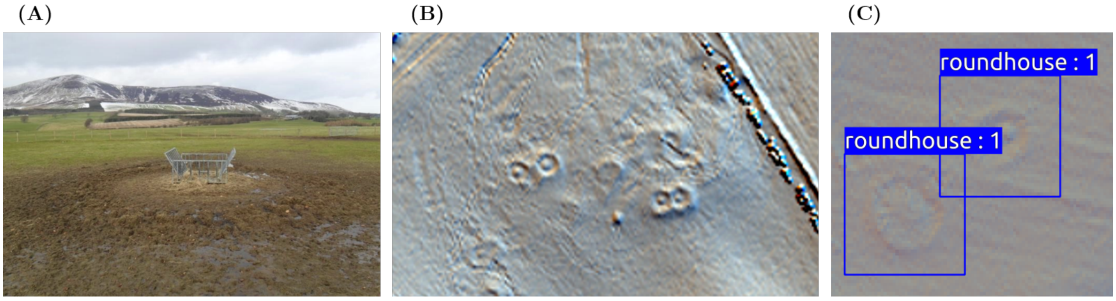


FIGURE 4.12: (A) image of a cattle feeder on the ground where cattle have been fed around a metal bin, (B) on LiDAR data where it shows the effect of the trampled soil and (C) as a false positive roundhouse detection. ©Historic Environment Scotland.

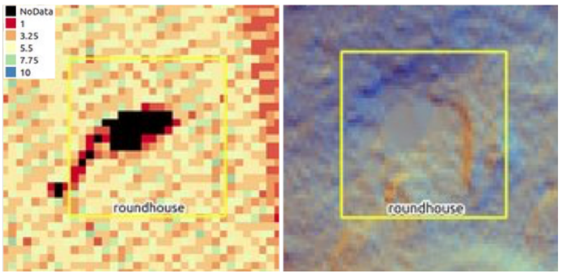


FIGURE 4.13: (A) Low ground point density caused by filtering the pointcloud to create a DTM. (B) This has selectively removed large boulders and produced a circular feature that looks like a round house. ©Historic Environment Scotland.

Based on this experiment we found that many false positives could be removed with a mask based on domain knowledge. Although a mask would be fine for rapid analysis, for accurate mapping we should improve the model or approach itself. Ideally the model would learn internally what makes a modern object different from an archaeological object. Another hypothesis we explored was to include common false positives as classes in our model (Figure 4.4). However, this made the approach more sensitive to finding these objects and confusion between the objects increased. As such we decided against continuing to develop that approach.

4.3.5 Experiment 2: Training on all the images

Our next experiment was to train and validate on all the available images. Even though the quick experiments reduced the chance of finding false positives from new/unknown

Table 4.4: Recall, precision and F1 results from object detection with a LiDAR visualisation combination of Local Dominance, Slope and Open Positive.

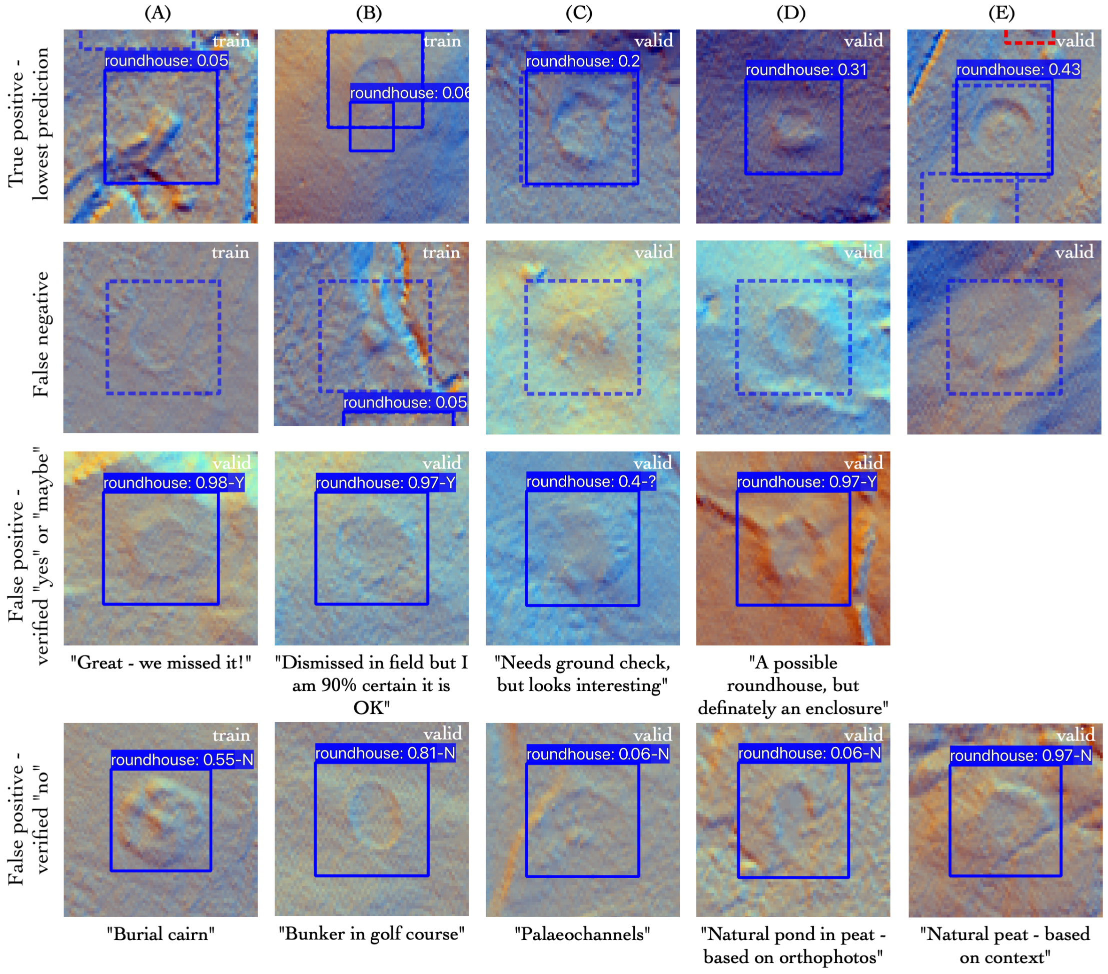
	Train			**Valid**		
	Roundhouse	*Shieling*	*Small cairn*	*Roundhouse*	*Shieling*	*Small cairn*
True positive	159	311	377	22	4	0
False positive	1	28	79	9	44	16
False negative	5	4	28	14	24	20
Precision	0.994	0.917	0.827	0.710	0.083	0
Recall	0.970	0.987	0.931	0.611	0.143	0
F1	0.981	0.951	0.876	0.657	0.105	0

sites, it also meant that the model was only trained on high quality data which didn't generalise well to areas where less or no sites are known. We considered the possibility that including all the images in the training would allow the model to learn to disregard modern built up areas and peat erosion mounds. We undertook this experiment using the LiDAR visualisation combination of Local Dominance, Slope and Open Positive (Figure 4.5). Through trial and error we also found that we got our best results when training each object class individually. The implementation of RetinaNet optimises the approach to Mean average Precision (MaP) which in practice meant that the model seemed to optimise to just one of the classes. Training each class individually increased the computation time but the trade-off was found to be worth it. As soon as we changed this element of approach we received visually impressive results in the validation area which we shared with HES for feedback. Unfortunately, we weren't able to retrieve the information from the best performing epoch (due to a power outage in the office) and so for each model we used the weights from the final epoch 50. Based only on the numbers in Table 4.4 it is clear that the model was overfitting on training data for the shieling huts and especially for the small cairns. Only 4 and 0 known sites were found respectively. Still, the visual results looked promising with possible new detections so feedback on the false positives would be useful to further tune the approach. In addition to the false positives in the validation we also ran the model in the training area to find out if any new sites could be found among the false positives.

Table 4.5: Recall, precision and F1 results from object detection adjusted by manual verification of the results. The manual verification increased the number of True Positives, especially for shieling huts and small cairns in both the training and validation areas.

	Train			**Valid**		
Manual verified	*Roundhouse*	*Shieling*	*Small cairn*	*Roundhouse*	*Shieling*	*Small cairn*
True positive	159	334	435	26	48	10
False positive	1	5	21	5	0	6
False negative	5	4	28	14	24	20
Precision	0.994	0.985	0.954	0.839	1.000	0.625
Recall	0.970	0.988	0.940	0.650	0.667	0.333
F1	0.981	0.987	0.947	0.732	0.800	0.435

The feedback from the HES expert was really useful; for each false positive he provided a basic interpretation of ‘yes’, ‘maybe’ and ‘no’ with further comments on why a detection was likely right or wrong (Figure B.4, Figure B.5). We updated the True Positives results from Table 4.4 by adding all the ‘yes’ and ‘maybe’ detections and removing those from False Positives and we then recalculated Precision, Recall and the F1 score (Table 4.5). The feedback from the HES expert has dramatically changed the initial results in both the training and the validation areas. To visualise the results we have created images for each class that show *True Positives with low predictions*, *False negatives*, *False positives that were verified ‘Yes’ or ‘Maybe’*, and *False positives that were verified ‘No’* (Figure 4.14, Figure 4.15, Figure 4.16).



"Great - we missed it!"

"Dismissed in field but I am 90% certain it is OK"

"Needs ground check, but looks interesting"

"A possible roundhouse, but definately an enclosure"

"Burial cairn"

"Bunker in golf course"

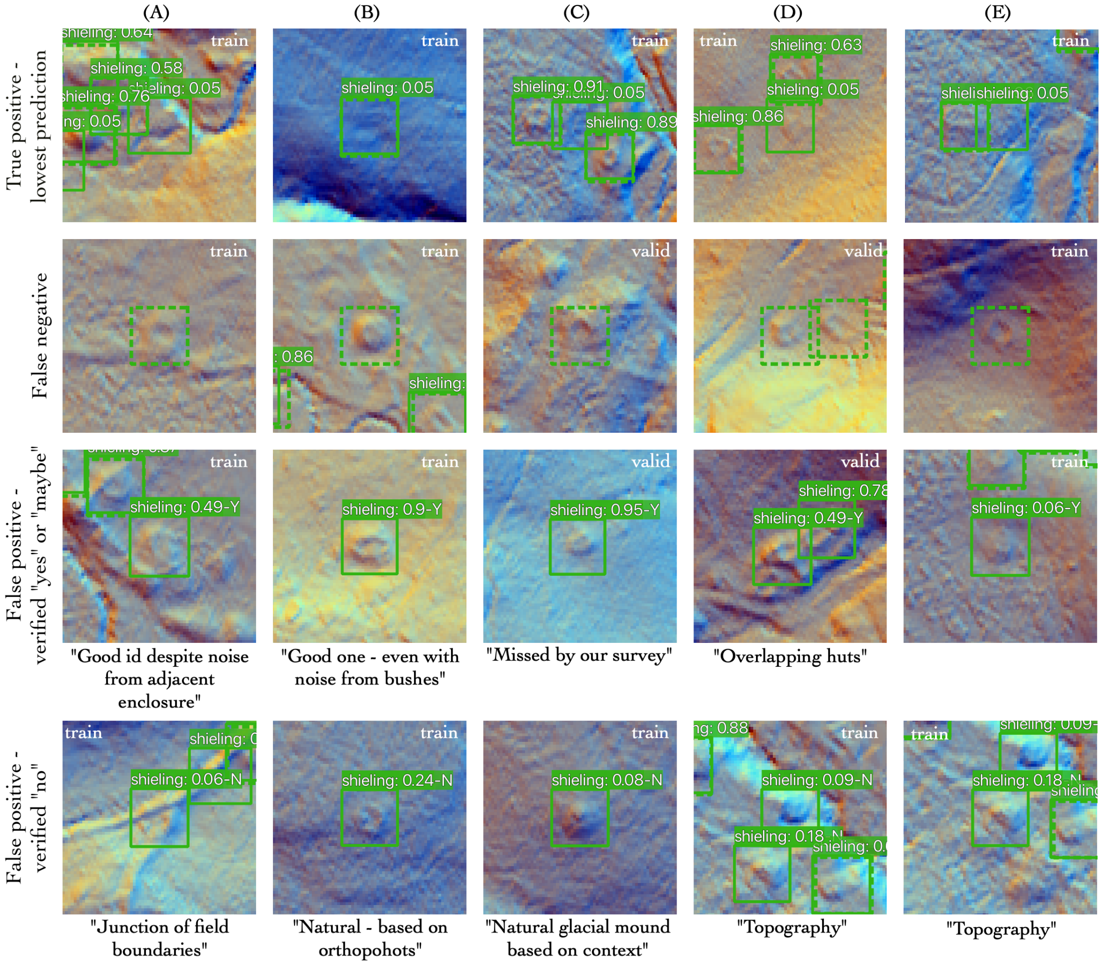
"Palaeochannels"

"Natural pond in peat - based on orthophotos"

"Natural peat - based on context"

Figure 4.14: Round houses detections: Visual examples of True Positive, False Negative and False positives that were verified as “Yes”, “Maybe”, and “No” accompanied with further explanation from the HES expert. ©Historic Environment Scotland.

Shieling huts now have the highest F1-scores. Shieling huts went from 28 down to 5 false positives in the training area and from 44 to 0 in the validation area. The *True Positives with low predictions* for shieling huts found mainly objects that were just outside of the known site and could be removed with a stronger threshold of IoU. Although shieling huts do tend to cluster together and overlap in that way as can be seen on the row *False positives that were verified ‘Yes’ or ‘Maybe’*. The *False Negative* row shows training



"Good id despite noise from adjacent enclosure"

"Good one - even with noise from bushes"

"Missed by our survey"

"Overlapping huts"

"Junction of field boundaries"

"Natural - based on orthopohots"

"Natural glacial mound based on context"

"Topography"

"Topography"

Figure 4.15: Shieling hut detections: Visual examples of True Positive, False Negative and False positives that were verified as “Yes”, “Maybe”, and “No” accompanied with further explanation from the HES expert. ©Historic Environment Scotland.

examples that are atypical but the validation examples seem like they should have been detected so the model must have overfitted to the training data.

Small cairns went from 79 down to 21 false positives in the training area and from 16 to 10 in the validation area. It seems that for small cairns the model was overfitting more than it was on shieling huts. There are two obvious cairn fields in the validation area that were missed by the approach (Figure B.5). The cairn field the west of the island was however found and makes up the majority of the new *True positives* (Figure B.5). An interesting observation is that in both the training and validation areas the small cairn detections in clusters have a seemingly higher prediction than stand-alone detections of cairns in the landscape. The model seems to have learned the clustering pattern in the training data which is a positive effect of the FPN structure and shows that it is considering the same geospatial pattern of context as an expert would do. We also noted that most of the false positives that were manually verified as ‘no’ were part of clusters. We expect that some of these objects would have been disregarded if they were found in an empty part of the landscape.

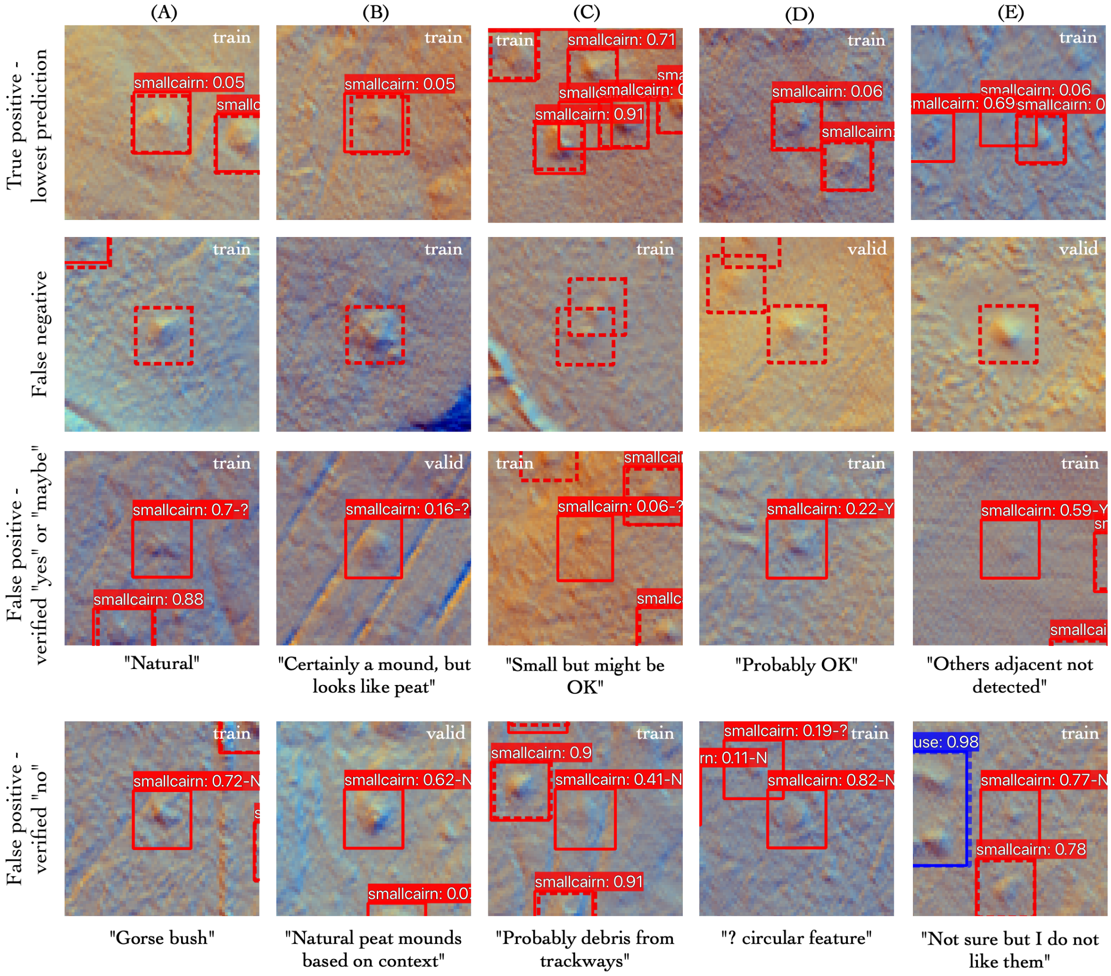


FIGURE 4.16: Small cairn detections: Visual examples of True Positive, False Negative and False positives that were verified as “Yes”, “Maybe”, and “No” accompanied with further explanation from the HES expert. ©Historic Environment Scotland.

4.3.6 Comparison with established results

The Isle of Arran case study has provided us with the opportunity to compare our automation approach to an exemplary manual approach. The dataset that we were able to use was free of noise because the site locations were manually tagged on the same LiDAR data. We were very surprised to have discovered 139 archaeological sites that were not in the training dataset. The expert from HES was most impressed with the new detections of round houses. Generally, those are simple to detect easy to verify just with the LiDAR data. One round house detection made by our AI was also identified as ‘enclosure’ on the LiDAR by one of the experts during the desk-based stage and it was given low confidence score so it was verified in the field (Figure 4.17). In the fieldwork stage it was dismissed as being ‘Natural’ but the main expert at HES now agrees with the model and is convinced this is an actual round house based on just the LiDAR. The False negatives for the round houses are all very faint or atypical.

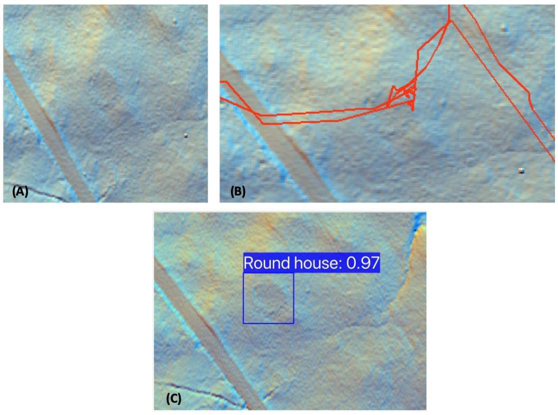


Figure 4.17: (A) RAMP Stage 1: Desk-based identification = 'Enclosure'. Level of confidence = 2. (B) RAMP Stage 2: Field visit = 'Natural'. (C) Object detection: Round house:0.97. The HES expert agrees with the round house detection. ©Historic Environment Scotland.

In chapter 2, subsection 2.5.1, we introduced the automation approach of Trier et al. (2019) on the Isle of Arran. The Norwegian Computing Center was commissioned by HES to preform an automation approach which they had previously presented in Trier et al. (2018). They trained three individual ResNet18 CNNs on round houses, shieling huts and small cairns. To infer new sites in a large area, they used a sliding window approach where each patch gets a probability from model. The combined result of the 3 heatmaps is shown in Figure 4.18-A. To visually compare the results with our approach we show our results side-by-side in Figure 4.18-B. Based only on this comparison we see that our approach detected 3 shieling huts that were not in the training set, 2 of these were verified true and 1 maybe. The detections of Trier et al. (2018) are more difficult to interpret. Our object detection approach creates a vector file that can be used with GIS to quickly iterate through the detected locations. With a vector file the heritage manager can also query the result as we have shown by comparing the locations of detections with land use and peat depth maps. In the workflow of a heritage manager this creates a more efficient approach than scanning the raster map for the entire case study area. It is probably more convenient and efficient to manually analyse the DTM visualisation than it is to interpret the raster heatmap. However, by thresholding the pixel locations that have a high class confidence in the the raster map, a polygon vector file could be created. This file can then be iteratively reviewed and queried. Based on a visually inspection of Trier et al. (2019) result for Glen Shuring this would probably result in many *False positives* and *False negatives*. Overall a HES expert compared our results to the Trier et al. (2019) outputs saying: "your data seems much cleaner – less noise for certain – and (to me) a clearer rationale for the false positives". We consider that our object detection approach performed better because we trained on more true negative locations which improved the generalisability of the trained CNN.

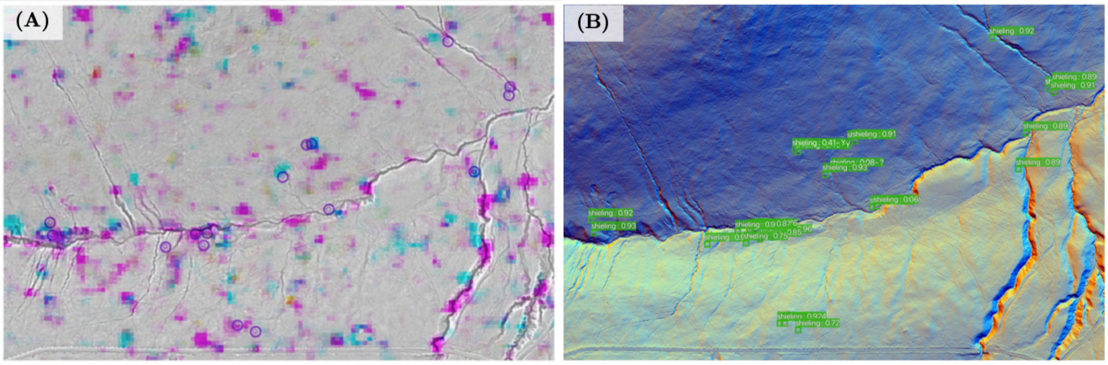


FIGURE 4.18: Comparison of our approach with previous automation work shown at Glen Shurig. (A) shows the result from Trier et al. (2019) with a heatmap of the probability of roundhouses (cyan), shielings (magenta) and small cairns (yellow) (verified sites are depicted as circles). (B) shows our object detection results. ©Historic Environment Scotland

4.4 Summary

In this chapter we streamlined our approach from experimenting with image classification to using optimised parameters in our object detection approach. We found in our image classification stage that Local Dominance, Slope and Open Positive DTM visualisations performed best on small cairn, round house and shieling hut respectively. We used a combination of these bands in our object detection approach. Different than chapter 3 we applied RetinaNet for object detection which generated better detections with less false positives. In the object detection we experimented with training on only images that had objects. This was a useful stage for quick iteration of different hyperparameters. We did however find that this trained model did not generalize well across the island. The majority of sites in Arran are on moorland and so the training data did not include images of the modern built-up area and only a few images of agriculture and woodland areas. Rather than removing potential false positives by masking them out we decided to re-train the model with all the images and this resulted in much cleaner detections. These result was shared with HES who verified our false positives and were pleased to find many sites that weren't previously known. In total 139 sites that were classed as a false positives turned out to be actual archaeological sites. We expect that retraining the experiment with the newly verified detections would again increase the number of new sites. We didn't pursue this because the value of the technique had been shown with this experiment alone. However, if we would improve this case study we would try to improve the balance of the training/validation data. We do expect that there are be more sites to be discovered in the dataset. By publicly sharing the dataset along with our RetinaNet benchmark we encourage our results to be improved and new detections to be made (Kramer and Hare, 2020). We envision this to be a learning resource and a testing ground for new techniques. To the best of our knowledge this is the first large-scale publicly available dataset of archaeological sites for benchmarking automation approaches.

Chapter 5

Discussion

The aim of this thesis is to find an optimal automation approach for archaeologists who are new to deep learning. Based on our experiments and extensive literature we have gathered deep insights on best practise for successful workflows. In section 5.1 we review our experiences on gathering and labelling data for deep learning approaches. In section 5.2 we review our most effective approach to using deep learning, and overcoming challenges that we have found while working with remote sensor data to detect archaeology. In section 5.3 we discuss future work topics that we think will further solve our identified problems.

5.1 Creating a deep learning dataset

The most important element of any deep learning approach is a high quality dataset. Most researchers spend the majority of their time improving or expanding the dataset because it is often the best way to improve accuracy. So, rather than tweaking network parameters for small percent improvements, most researchers should review and improve their dataset quality and quantity and results will improve.

5.1.1 Evaluating data quality

In chapter 3 we used archaeological site locations that were available through local archives (HER) that are gathered in a similar approach across the country. This allowed us to test the usability of such a dataset for automation and analyse how our approach could be scaled up across the country. Through our experience we found that this data source required some manual improvement. For example, the dataset contained legacy data which meant that some sites no longer excised due to modern development. We also found that some of the sites were not visible on one or both of the remote sensor datasets we used. Finally we also had to adjust the box size for each barrow. Because there was a lot of size variation between the barrows, our initial approach of using the largest barrow size as a guide to crop all barrows resulted in low detections rates of the smallest barrows. We were able to improve the data quality over a couple of days. However, we are not local experts so mistakes could have induced some noise. Overall the dataset was very good to work with because the New Forest has been extensively researched over the years. Our experience of this dataset will probably transfer well to similar case study areas but we expect that less studied areas might require more expert manual interpretation of the area. Alternatively one could speed up the manual process by iteratively applying deep learning and reviewing results to update the training dataset and retrain the algorithm.

In chapter 4 we were able to use a manually improved version of the national database which was very accurate and thus quick to experiment with. The case study covered 432km^2 which provided enough objects for training our deep learning approach. The clustering of known objects in the west of the island made it difficult to separate areas for training and validation whilst maintaining a good balance of object classes in the respective areas. In our approach we divided the training and validation areas based on their km^2 size and accepted the training/validation object class imbalance. We experimented with increasing the validation objects by using a larger validation area but we did not find a good balance. This was mainly because we choose to create the same training/validation area split for all classes and maintain geographical separation. If we relaxed those requirements, an improved approach might be to divide the areas based on the number of known sites per class. Our result evaluation yielded many false positives which turned our to be previously unknown archaeological sites. We expect this will be the same for most cases studies if the area has had long periods of occupation.

On balance we would recommend that very simple case studies can use national datasets albeit with the caveat of the known issues and suggested quick improvements. For an in-depth research we would recommend the approach of collaborating with local experts that have resources to provide a high-quality dataset, and time to provide feedback and discussions on the results. This allowed us to find many new sites among the false positives and come up with potential improvements. Without this feedback, we were at danger of optimising our approach to a dataset that is too noisy.

5.1.2 Expert labelling tools

At the start of this PhD there were no labelling tools available that could be used for a geographical purpose. We therefore created our own code to automatically create training data based on the known site locations. In both case studies we were provided with a Shapefile that contained the centre points of known sites. We also asked for the maximum width of the objects which we then used to crop our images for image classification or to create a file with bounding boxes for object detection. This approach worked well on Arran but in the New Forest there was a lot of variance of barrow width which reduced general accuracy and manual adjustment of boxes was required to gain further improvements. Our tools are shared in our GitHub repository and although they work for points and maximum object width, the full approach works best with polygons that indicate the width of each object (Kramer and Hare, 2020).

There are several opensource labelling tools used in archaeology; Verschoof-van der Vaart and Lambers (2019) used a general image labelling tool called LabelImg for object detection and Soroush et al. (2020) used 3D Slicer for segmentation which is an open source software platform widely used in medical image processing and annotation. Both these tools do not support geocoordinates and are not able to automatically transfer known sites locations to a machine learning format. Because they require manual tagging the labelling process can be time consuming. These tools are also only useful to create labels which means that either the interpretation is can only be done locally with resulting images from the deep learning package used. If they would want to analyse their resulting detections in a GIS software then they require additional tools to transfer the local detection coordinates into geocoordinates. For us it was more convenient to write our own code.

There are several commercial labelling tools that have geocoordinate options. GroundWork is the first annotation tool designed for geospatial data. One can upload any remote sensor dataset to label and selected parameters to create overlapping tiles that can be exported to use directly for machine learning. ArcGIS Pro is a commercial GIS software that has recently expanded its software with tools to transfer GIS datasets into a deep learning format and tools to review results. This was used in archaeology by Gallwey et al. (2019) to automatically convert their labelled vector file containing geographical site locations and raster data containing the DTM into deep learning training datasets. Although ArcGIS Pro is expensive, the tools seem well suited for aerial archaeology because archaeologists are often already familiar with the software and it provides good user support and guides. Another labelling tool that has some geospatial options is Amazon Sagemaker which could be useful when using the full AWS machine learning pipeline. A similar one-stop platform is Google Earth Engine which has limited labelling tools but accepts GIS vector and raster data without further processing requirements (Gorelick et al., 2017).

5.1.3 Crowd sourced labelling

Alternative to expert interpretation, labels could be created with the help of crowd sourcing. The ImageNet dataset is also crowd sourced and approach and showed the impressive nature of deep learning. The ImageNet crowd sourcing was run on a commercial platform called Amazon Mechanical Turk where users are paid per label. In archaeology all crowd sourcing projects have been run as citizen science experiments where citizens contribute to scientific discovery and are upskilled in both archaeology and remote sensor interpretation. Although most experiments discussed below were not created for the purpose of deep learning, many can be used for this.

The successful project of the search for Genghis Khan's tomb by National Geographic attracted 10,000 volunteers who contributed 30,000 hours (3.4 years), and together examined 6000 km^2 of high-resolution satellite images in Mongolia (Lin et al., 2014). The volunteers were asked to provide centre locations of potential heritage and this generated 2.3 million points that included burial mounds, megaliths, and city fortifications. Their approach of allowing any site type to be added to the database does require extensive post-processing which is very time consuming.

In early 2017 another approach was launched by Sarah Parcak called GlobalXplorer. The platform used binary image classification to identify and quantify looting and encroachment to archaeological sites. This in a sense is more efficient because it is a simple task that can be quickly learned so the resulting data is likely of good quality. The images are shown to multiple users and those that are consistently marked as showing looting are further analysed by project staff. This type of labelled data could be used to train a deep learning approach.

The previously mentioned deep learning approach by Verschoof-van der Vaart and Lambers (2019) found around 1000 new sites during the manual tagging of their research area. The extrapolated potential of the entire research area was found significant and so they explored the opportunity to use crowd sourcing for image labelling (Lambers et al., 2019). They launch their crowd sourcing campaign Heritage Quest on the Zooniverse platform. The tasks were to identify the centre point of barrows, segmenting Celtic field systems and segment cart tracks. With this approach they detected many potential archaeological sites. In addition to the traditional online crowd sourcing they are also experimenting with volunteer field verification.

Stewart et al. (2020) also used crowd sourcing for the purpose of training a machine learning algorithm. In their pilot study they used binary image classification to identify crop marking on Satellite Imagery using the Pybossa platform. While only 28% of the tasks are completed they already identified many new cropmarks which can be used at the next stage of their research in which they will apply machine learning. However, they also realise that the detection of cropmarks is very challenging because the patterns are very different between crop types and growth stages.

Whereas Lin et al. (2014) allowed a wide range of archaeological detection, most approaches implemented simplified tasks that require limited training and post-processing work. This reduces the ability to detect unusual archaeological sites and so most projects have added the opportunity for expert users to notify tiles that include potential archaeological sites. Such unique detections are unlikely to be found with deep learning because of their limited occurrence, making crowd sourcing the superior the superior choice.

5.2 Effective workflow

5.2.1 Iterative workflow

We have found that it is challenging to gain intuition without experience. We have therefore built our approach on experiments that take a baseline and improve that step by step. This both builds intuition and provides a justification for each tool that has been applied which together result in improved outcomes.

The most important step of our process is to start with image classification to get to know your dataset and to test your hypothesis. Our first step was to establish a simple deep learning baseline. We iteratively added more complexity to ensure that each step was improving our deep learning workflow. This included data augmentation, pre-trained networks, data visualisations, data fusion and different CNNs. Based on the outcome of the image classification we set our expectations for object detection.

Our best workflow for object detection was developed during the Arran case study. Here we also set a simple baseline from which we would test potential improvements. Whilst we initially trained our approach on all classes we eventually chose to train on individual classes to ensure that the model could optimise the outcome for each class individually. The baseline included only the tiles that had known sites to enable quick iteration and avoid overwhelming noise from false positives. We then used the optimised trained model to validate on the remaining tiles. This resulted in a lot of noise from false positives because the model was not trained on modern land cover. To omit the noise, we considered masking the areas that caused the majority of false positives. Although we decided against this because it could potentially remove actual sites. For our final approach we divided the whole island in train and validation areas so that the model could learn to disregard false positives in modern areas. This resulted in very high performing model that drastically reduced the false positives although it still had a much lower precision than recall for shieling hut and small cairn, even in the training area. We shared the resulting detections of both training and validation areas with our local expert and the feedback was essential to discover new sites and further hypothesise improvements.

5.2.2 Problem reduction

In this thesis we have discussed several reasons why using deep learning for the detection of archaeology on remote sensor data is much more challenging than the generally used ImageNet datasets. The main challenges include the non-conventional data format, low contrast and the small size of our datasets. In order to make deep learning work for our problem we have had to reduce our problem which means to modify our data to a known problem such that it can be easily solved using existing techniques.

For example, LiDAR data is captured in a point cloud and at the start of this PhD there were no deep learning solutions to this problem. We therefore experimented with raw DTMs which worked but they were underperforming when compared to the aerial photography. Large parts of the case study area were covered with forest canopy which precludes detections on aerial photography. We found that the raw DTM input required normalizing which we did with the means of the training data. Kazimi et al. (2019) addressed the same problem and found that applying min-max normalization on a per-image basis worked better than on the whole dataset. This process emphasises the importance of the local pixels without losing detail of the DTM. To maintain this high level of detail, it is important not to convert the image into 8 bit before feeding it to the CNN as previously discussed in subsection 2.5.3.

Even though per-image normalization could have improved our outcome, we consider that in this approach the small dataset will remain challenging to work with. The most powerful tool for small datasets is the use of pre-trained networks which is why most approaches in archaeology, including ours, focussed on DTM visualisations. The DTM visualisations are essential to human interpretation of LiDAR data which makes them an excellent fit for LiDARs that are pre-trained on the natural image scenes of ImageNet. Most researchers including Trier et al. (2019) and Verschoof-van der Vaart and Lambers (2019) used a SLRM which is commonly used for flat terrain but is discouraged to use in a more dynamic landscape. To overcome this Somrak et al. (2020) used a blend of analytical hill shading, slope, positive openness and sky-view factor into a single greyscale image called VAT. We contributed to this debate by using the 3 visualisations that worked best for our case study and combined them as the Red, Green and Blue bands of an image. This both allowed us to use a pre-trained network and to add more detail for the CNN to learn from. Despite this innovative approach the visualisation still reduces the detail that is available in the DTM. We are therefore very excited by the Lunar LiDAR pre-trained approach from Gallwey et al. (2019) because it maintains the detail of the DTM.

5.2.3 Choosing a deep learning algorithm

We have found that most deep CNNs will provide a similar result on our datasets and have therefore not focussed extensively on creating new CNNs but rather explore auxiliary techniques to alleviate specific problems we identified with the data. We found that deep CNNs worked better than shallow networks although the deep CNNs required transfer learning to perform well. We also found that RetinaNet worked best because it addresses the class imbalance and scale issues with focal loss and the FPN.

The choice between object detection or segmentation should depend on the case study aims. We have seen that Verschoof-van der Vaart and Lambers (2019) used object detection for the discovery of Celtic fields which can take irregular shapes and do not fit into a bounding box. At first sight the choice of object detection over segmentation might seem odd but their objective was to detect new sites not to perfectly segment them. If pixel accuracy is not the most important metric than using loss-functions for per-pixel optimisation is not the right approach.

5.2.4 Choosing an evaluation metric

Most researchers evaluate their approach with the false positive rate and the final F1 score. To improve their false positive rate Verschoof-van der Vaart et al. (2020) published an updated version of their approach using Location Based Ranking to mask built-up areas, and areas with drift-sand that were known to have low likelihood of archaeology but a high number of false positives. Ultimately the success of an approach is not dependent on one metric, it depends on what is most suited for a specific task (Soroush et al., 2020). In the medical profession classifying a sick person as healthy has a different cost than the opposite case and so doctors prefer to review more false positives and accept a higher recall with lower precision. In the case of Verschoof-van der Vaart et al. (2020), their focus was on large scale mapping where it was acceptable to miss a few objects for a higher precision to increase the overall success measured in the F1 score. Automation in archaeology is still at an early stage where researchers are trying to locally optimise an approach. In the future we foresee that a heritage manager may accept high recall with lower precision when it only takes them a short while to shift through the detections. The same is apparent in commercial archaeology where high recall is the most important metric.

5.2.5 Choosing an implementation

For most archaeologists it will be challenging to recreate a deep learning approach based on a paper alone. Luckily, many papers in deep learning are published with code that is available through GitHub folders. This was the case for the RetinaNet implementation we used and also for the Lunar LiDAR approach that Gallwey et al. (2019) used. These implementations make it a lot easier to apply deep learning but users still require skills in coding to use this. In our case we had to write code to generate images and matching label-files that could be interpreted by our RetinaNet implementation. We further wrote code that transformed the outcome back into geocoordinates and applied NMS on all the detections to remove duplicate detections caused by overlapping tiles. Gallwey et al. (2019) was able to find a ArcGIS Pro solution for most of these coding problems which made it easier to use their implementation. Yet this approach requires multiple pieces of software and is not flexible when experimenting with other CNN implementations.

We extensively researched alternative options and found two open-source end-to-end implementations, Avezea Raster Vision and Solaris, that are specifically designed for deep learning on remote sensing. They both offer a wide array of deep learning implementations written in Python with PyTorch and have good documentation and support. Their implementations offer image classification, object detection and object segmentation and multiple choices of CNNs. They are both part of commercial companies and have large teams that maintain the platform so we are confident that they will be maintained for a long period throughout which they will update the approach with the latest deep learning research.

5.3 Future works

During the course of the PhD the field of deep learning expanded rapidly with many new approaches published on non-conventional data sources. Work on self-driving cars has especially pushed research using LiDAR sensors. This has resulted in many networks resigned for convolutions directly on the point cloud (Özdemir and Remondino, 2019; Qi et al.). It also resulted in networks that combined both RGB images and LiDAR point cloud or LiDAR derived depth maps (Hazirbas et al., 2016; Qin et al., 2018). Progress on each of these elements has also been shown on remote sensor data which can be very helpful for archaeology. For example, Rudner et al. (2019) segmented flooded buildings based on change detection in two satellite images before and after the flood. This could also be used for archaeology to aide the detection of site destruction because of modern development, climate change or looting.

We also inspired by the crop stress dataset that was published by Chiu et al. (2020) and so we embarked on our own research to detect crop marks. In our first attempt we hoped that we could use high resolution aerial imagery from the Ordnance Survey that was captured during a drought in the summer of 2018. For site locations we were able to get crop mark locations from the national database (https://canmore.org.uk/) from HES. Unfortunately there weren't enough matches between the known cropmarks of the past and the images provided by the Ordnance Survey to train a CNN. Our second attempt included the national database (https://canmore.org.uk/) from HES which includes all the historic cropmarks that have been captured by HES. This approach provided a reasonable match between the objects and the images. Many objects were captured multiple times which increased the number of training examples. Although not every image included a the same/all cropmarks that were digitised (due to crop rotation or other growth differences in the field). Yet, we still persevered to see what result could be achieved. The object information was provided in lines which created a challenge because for our desired segmentation approach we required polygons. To create polygons from the lines we experimented with the buffer and polygonise options in QGIS (Figure 5.1). We discovered that polygonise worked well for enclosed lines and that buffer worked well for stand-alone lines. Although this worked most of the time we noticed that many polygon-type features were not enclosed such as the paleochannels in Figure 5.2. Rather than focusing on all cropmarks we reduced the problem to only detecting round houses. We had 5654 individual instances of round house in the dataset we would be a good dataset to train on. To further simplify the approach we chose not to continue with segmentation but experiment with our already verified RetinaNet approach (subsection 4.3.5). We created bounding boxes around each round house entry and created our training data (Figure 5.3). The results of a quick training routine were not good and after further inspection we found that many objects were in a white background that had value 255 rather than no data (Figure 5.4). We experimented with removing the objects that contained values of 255 but this also removed many good objects. In the end we decided that there were too many problems with the data and this project should be continued in a dedicated project where resources are available to manually improve the data.

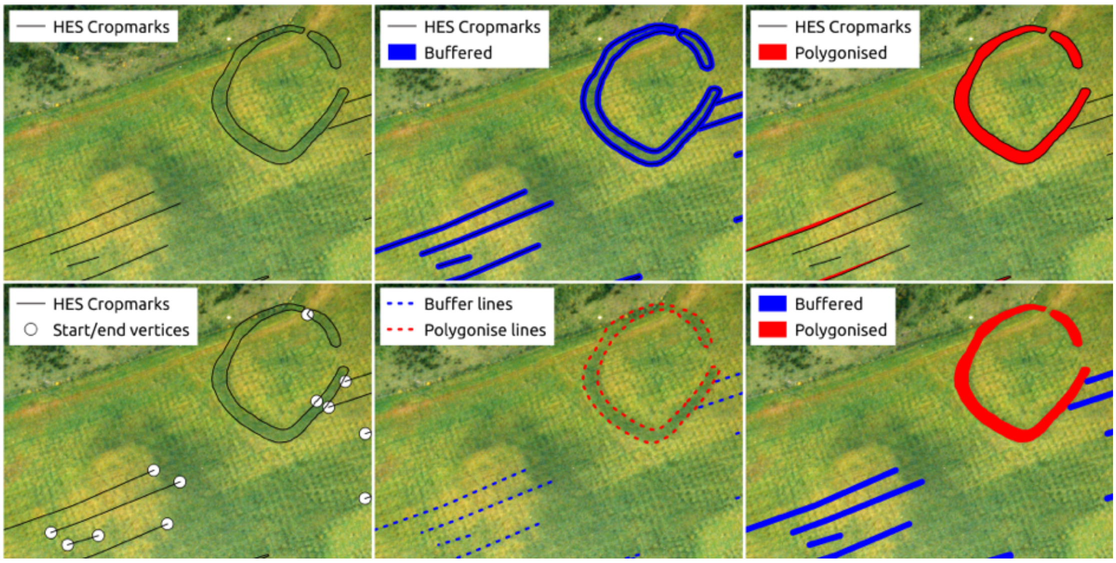


Figure 5.1: From top left to bottom right the image shows the process of creating polygons from lines. We found that the GIS buffering option didn't enclose the circular feature. We also found that polygonising didn't represent the actual lines properly. We found that the start/end vertices of the polygon object were on the same location for the circular feature which separated them from the lines. Using this insight we were able to separately polygonise and buffer the different object types. ©Historic Environment Scotland. Licensor canmore.org.uk

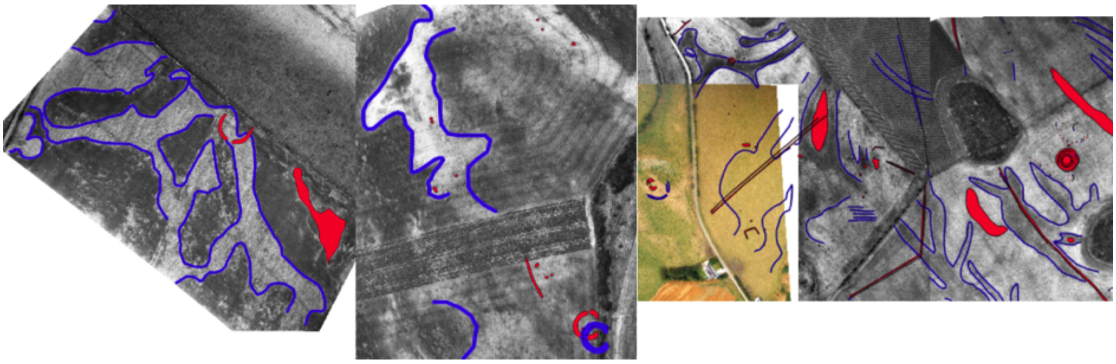


Figure 5.2: Image shows that our separate polygonise and buffer approach did not work for all objects. ©Historic Environment Scotland. Licensor canmore.org.uk

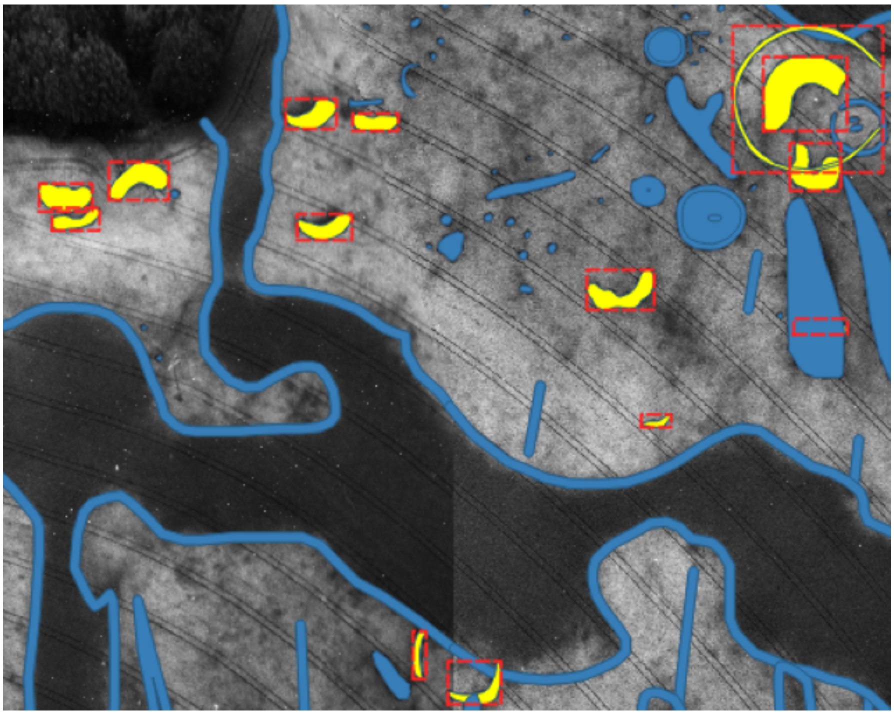


FIGURE 5.3: Image showing the created bounding boxes for object detection on aerial images. We used the extent of the round houses to create bounding boxes. ©Historic Environment Scotland. Licensor canmore.org.uk

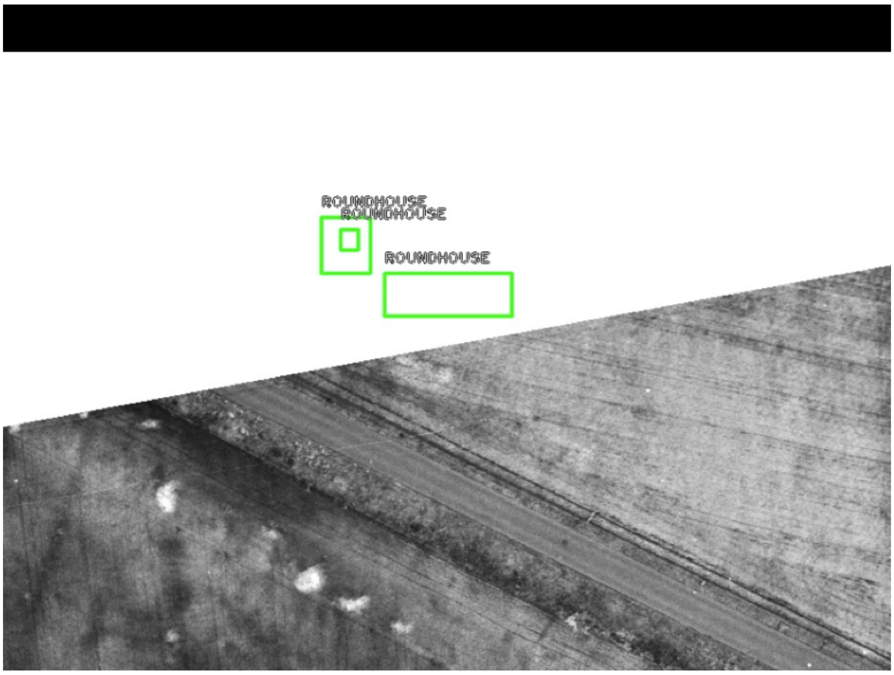


FIGURE 5.4: Image showing bounding boxes that are stored where the aerial image is cut off. ©Historic Environment Scotland. Licensor canmore.org.uk

Chapter 6

Conclusions

This thesis presented an in-depth research of deep learning approaches and the challenges that are presented by the datasets of archaeological sites and the pattern they leave on remote sensing data. In chapter 2 we highlighted several reasons why the detection of archaeology on aerial imagery is highly challenging. In the following case study chapters we addressed most challenges in extensive experimentation and created a workflow that addresses these central issues.

In chapter 3 we focussed on the use of multiple sensors, comparing results from multispectral imagery with LiDAR derived DTMs. The case study focussed only on barrows in the New Forest National Park and we used the known sites primarily from the local archives HER for training locations. Through experimentation we found the best results with DTMs derived visualisations that highlight the archaeological earthworks. This showed that for small datasets the problem should be simplified to attain high accuracy. Although we do expect that with larger datasets or pretraining with DTMs (or similar dataset) will eventually surpass the accuracy that can be obtained with simplified visualisations. We also discovered that noise in the dataset was trailing our accuracy and that extensive manual improvement of datasets is required for deep learning use cases. In this chapter we also experimented with a multitude of networks and hyperparameters. Eventually we concluded that the SOTA networks also work best for our datasets and that network tweaks are less important than improving/increasing training data.

In chapter 4 we focussed only on LiDAR data but diversified with 3 different archaeological object types; round houses, shieling huts and small cairns. The dataset from the Isle of Arran was provided by HES and had been gathered through extensive desk-based and field verification. In this case study we were able to perfect our workflow that encourages feedback and critical evaluation of the dataset and results by starting with image classification and using those learnings/ optimised hyperparameters to apply object detection. We shared our approach and dataset on GitHub as a benchmark which is the first in it's field and we hope it will encourage comparison with new research.

In chapter 5 we discussed the best practise for the workflow that we created based on the literature review and our own experiments. The discussion can be read as general advice for new researchers in the field and ranges from the creation of a deep learning dataset to model selection and model evaluation. We finally concluded the chapter with a review of the latest research in deep learning that could be used in archaeology to improve the approaches that have been published so far.

In chapter 2 we listed the key challenges of archaeological label datasets and the remote sensing datasets. Below we have summarised our most important experience to overcome each of the challenges:

- Small datasets; We used several tools such as data augmentation, transfer learning to improve outcomes for small datasets and tested these hyperparameters during the image classification stage.
- Class imbalance; We addressed the issue of class imbalance only at the object detection stage with a RetinaNet that implements focal loss.
- Noise; We addressed the possibility of noisy labels during the image classification stage by reviewing the most extreme "right" and "wrong" predictions from the CNN. We also highlighted the importance of improving the dataset before moving into object detection. We discovered many unknown sites at the object detection stage by critically analysing the relatively high number of false positives in both our training and validation areas. We concluded that the algorithm should be iteratively updated with new verified detections to gain the optimal result.
- Scale; We addressed the issue of scale at the object detection stage with a feature pyramid network that analyses objects at different scales and was able to detect very small objects. We also suspect the FPN learned to increase probability if certain objects were found in clusters.
- Low contrast; We used DTM visualisations to reduce the data complexity and improve visual interpretation with distinct lines and edges. These patterns are learned in CNNs that are pretrained on ImageNet. Using a pretrained CNN was essential for solving the previously mentioned small data problem.
- Non-conventional data format; We learned that using the raw DTM was not as successful as using a visualisation in the experiment of chapter 3. Using a CNN that was pretrained on 16 bit DTM or similar dataset like Gallwey et al. (2019) should improve the results because the raw DTM contains more details than the visualisations.
- Changing appearance; We addressed this issue by training and validating the approach in geographically distinct areas. On our scale this effect was not significant.

- Fuzzy site definitions; Our approach in Arran contained round houses and shieling huts which have the same function. We saw many detections of both classes on the same location. In all cases the NMS removed the least confident class which was the confused classification.

We have addressed each of our main concerns with the archaeological datasets but there is still much to be improved. We hope to that the availability of our new datasets and benchmark will both facilitate more comparisons of existing methods and lead to increased interest in the detection of archaeology on remote sensing data in the machine learning and computer vision communities.

Like we said in chapter 1, archaeology is under constant threat of destruction, and it of utmost importance that sites are located so that they can be monitored and protected. The potential saving that automation can provide is huge. In addition, we have shown that more training data creates better results. We envision a continuous loop with deep learning detections and manual verification which in turn is fed back into the model for retraining. In time the model will be able to detect all the archaeological sites and monitor any changes. We therefore encourage bold initiatives for large scale mapping.

Appendix A

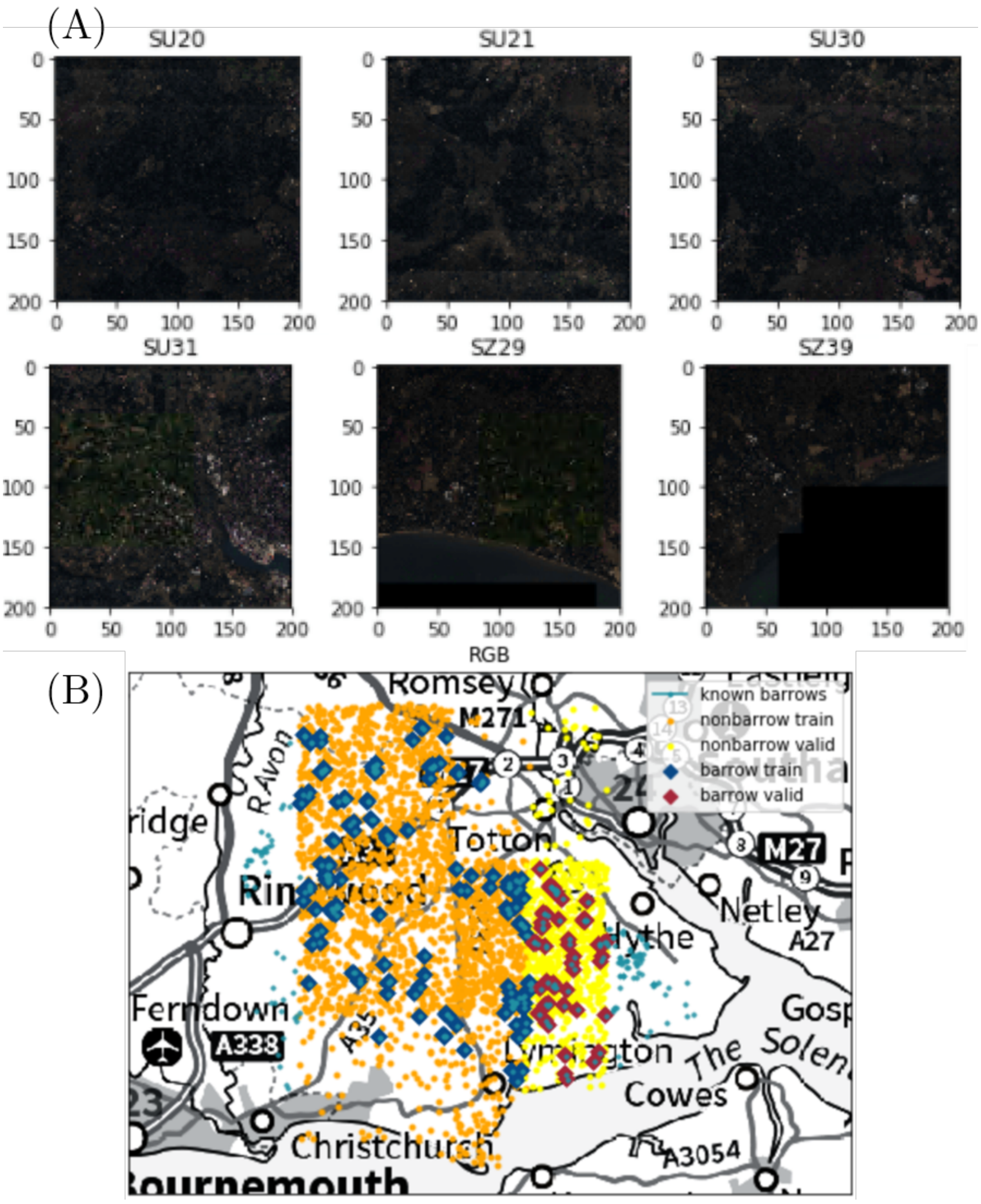


FIGURE A.1: (A)RGBN data as it was provided by the Ordnance Survey and (B) Spatial spread of RGBN positive/negative images.

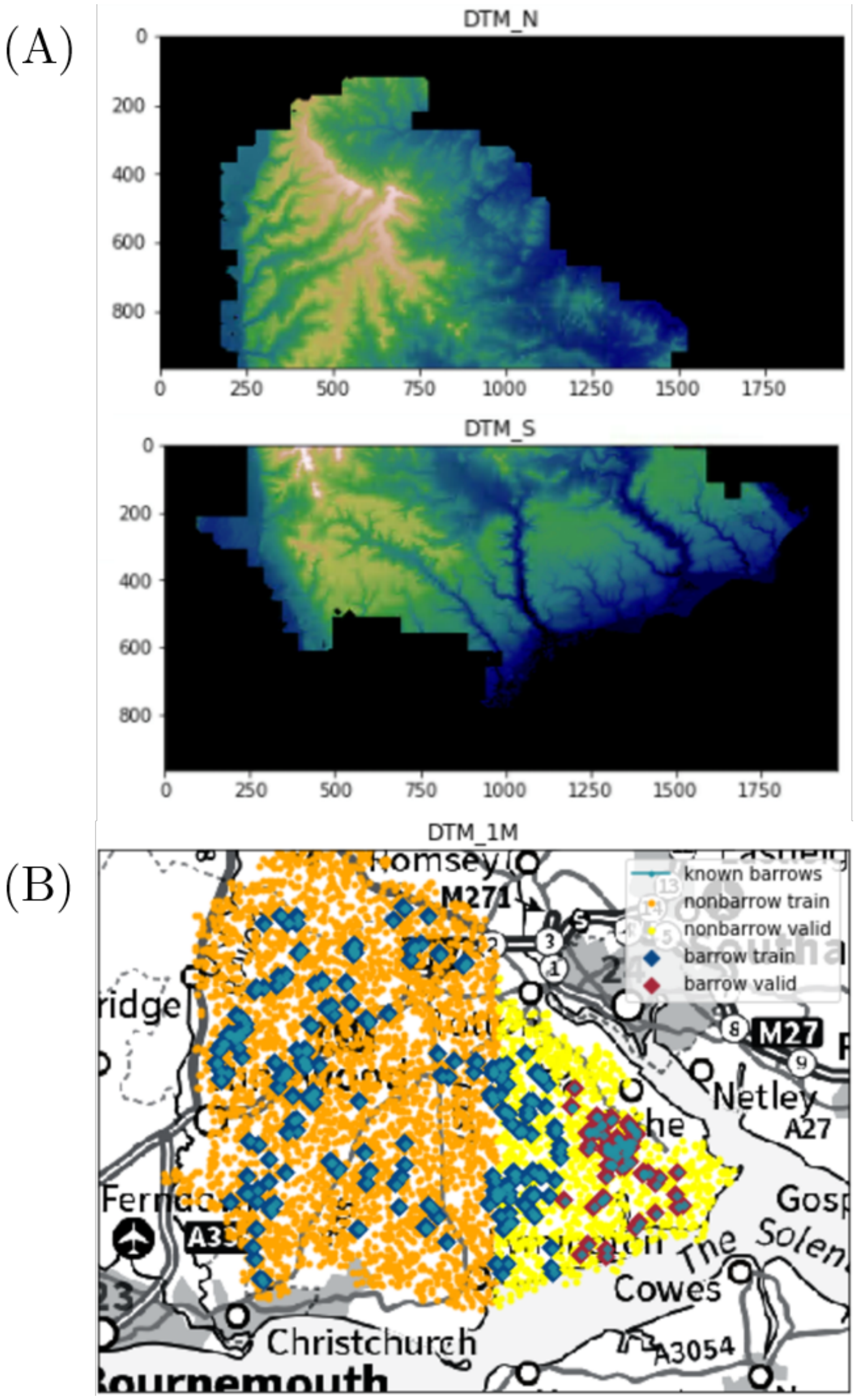


Figure A.2: Spatial spread of RGBN positive/negative images.

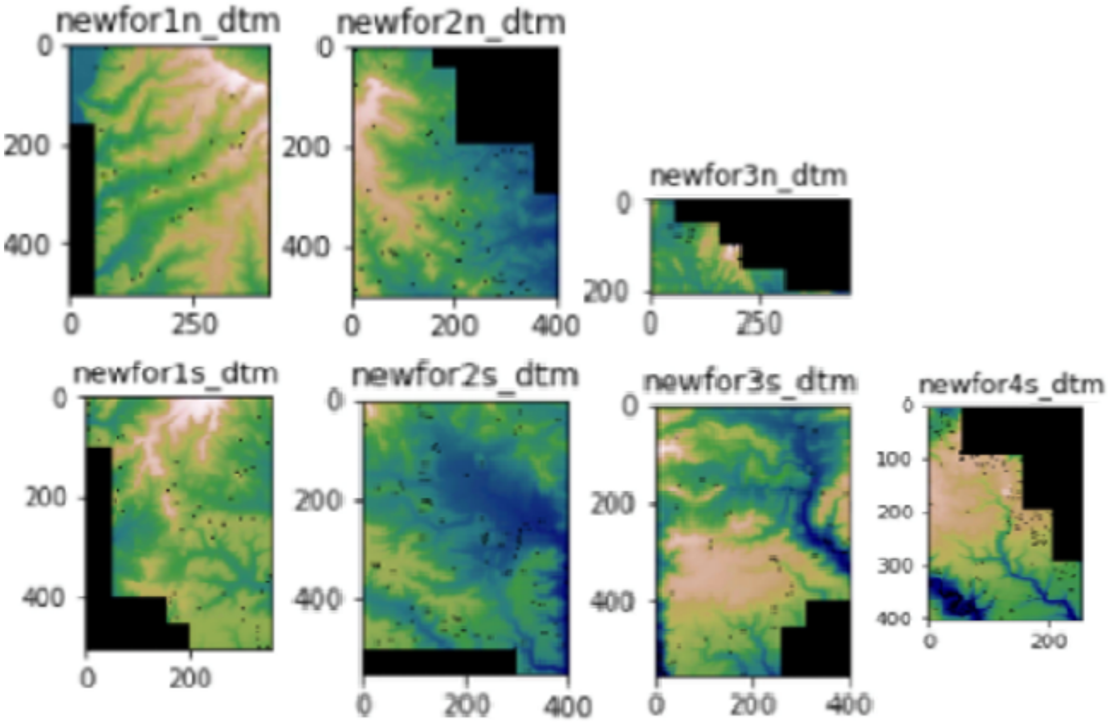


Figure A.3: DTM-0.5 m data as it was captured and processed by University of Cambridge (2011) and provided by the New Forest Knowledge project.

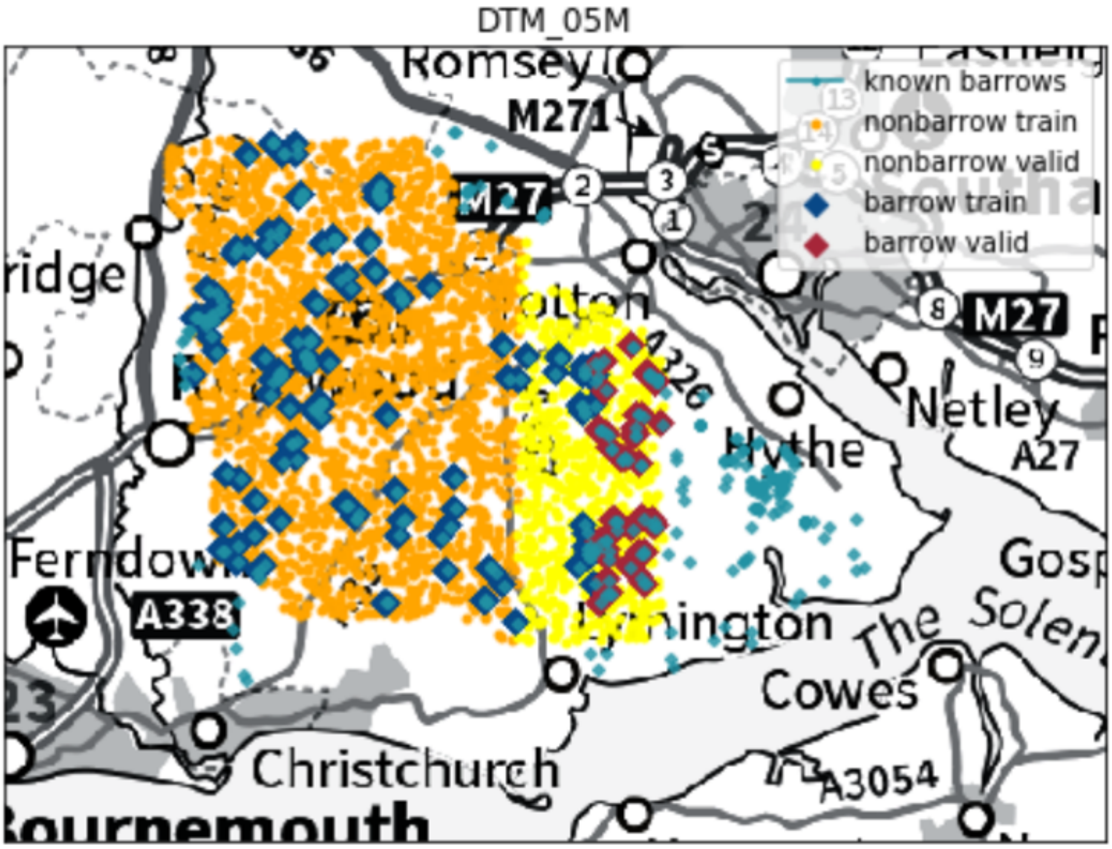


Figure A.4: Spatial spread of DTM-0.5 m positive/negative images.

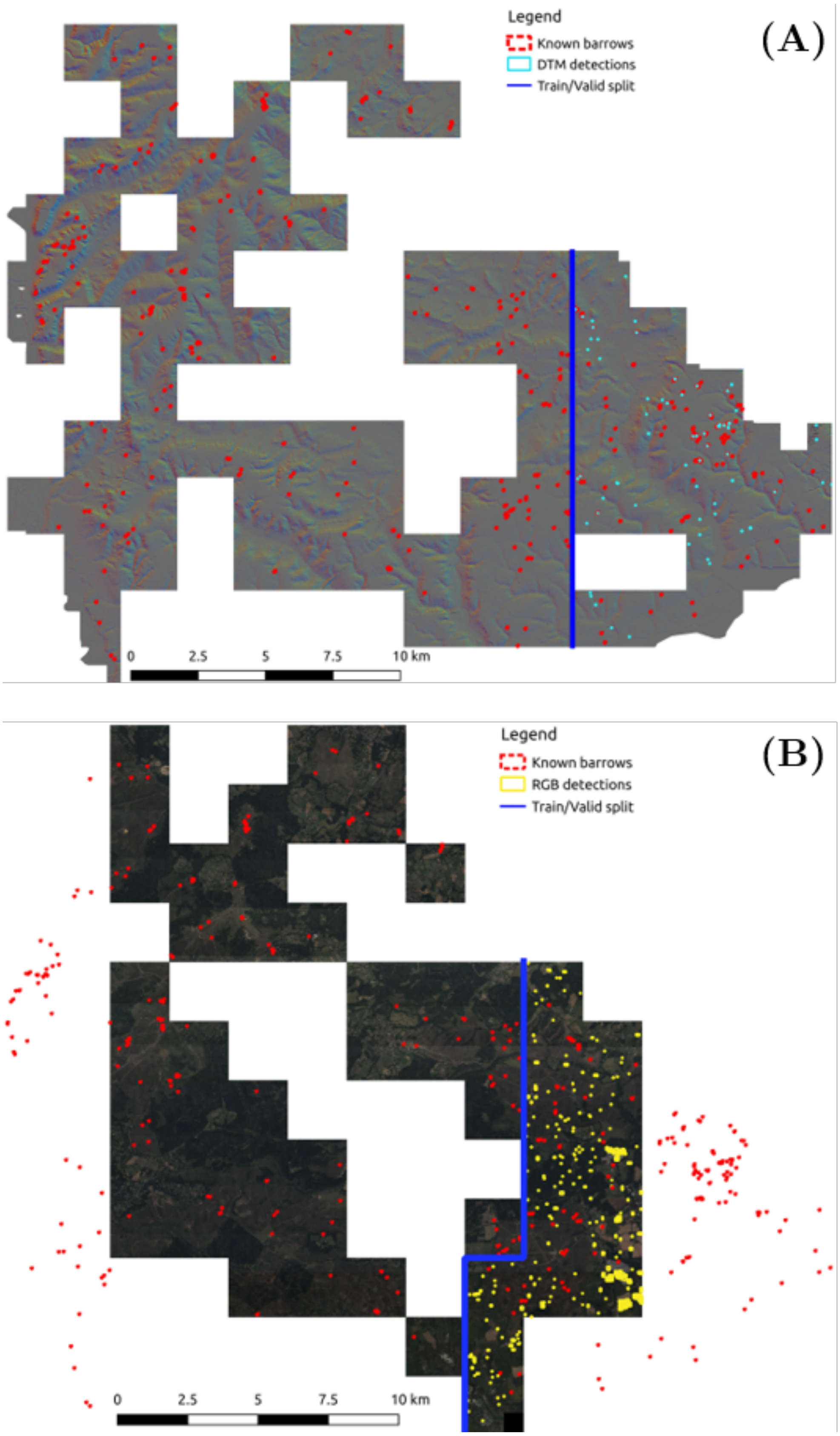


FIGURE A.5: Training and validation regions for (A) RGB and (B) DTM overlaid with known barrows and new detections.

Appendix B

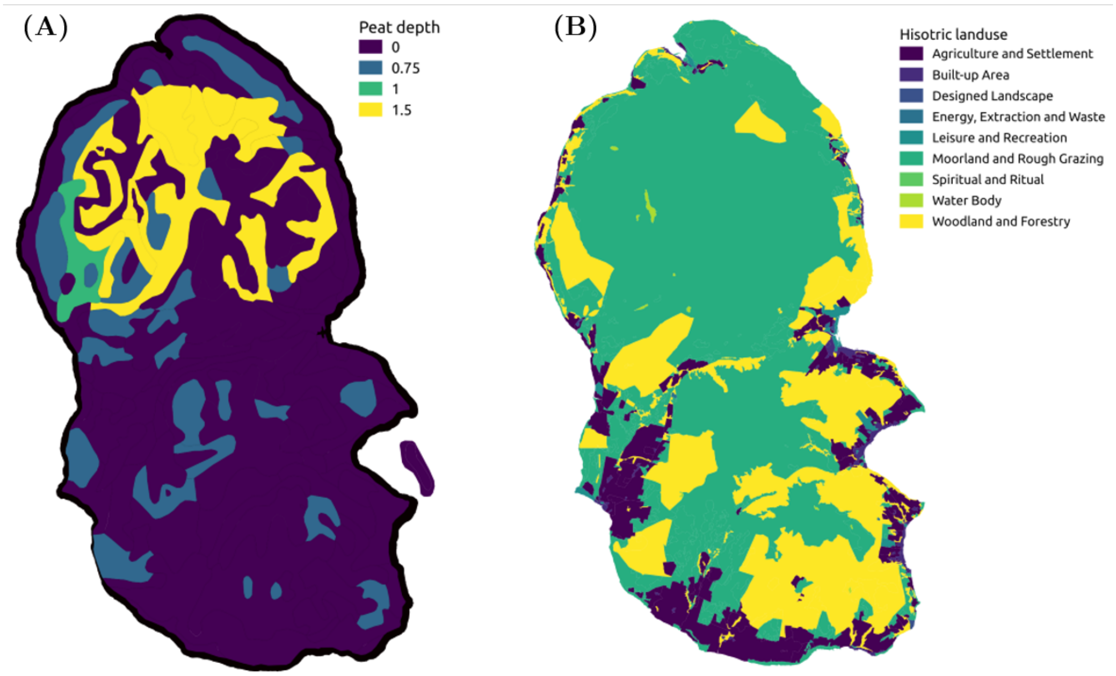


FIGURE B.1: Map of Arran with peat depth (A) and historic landuse (B) distribution. ©Historic Environment Scotland.

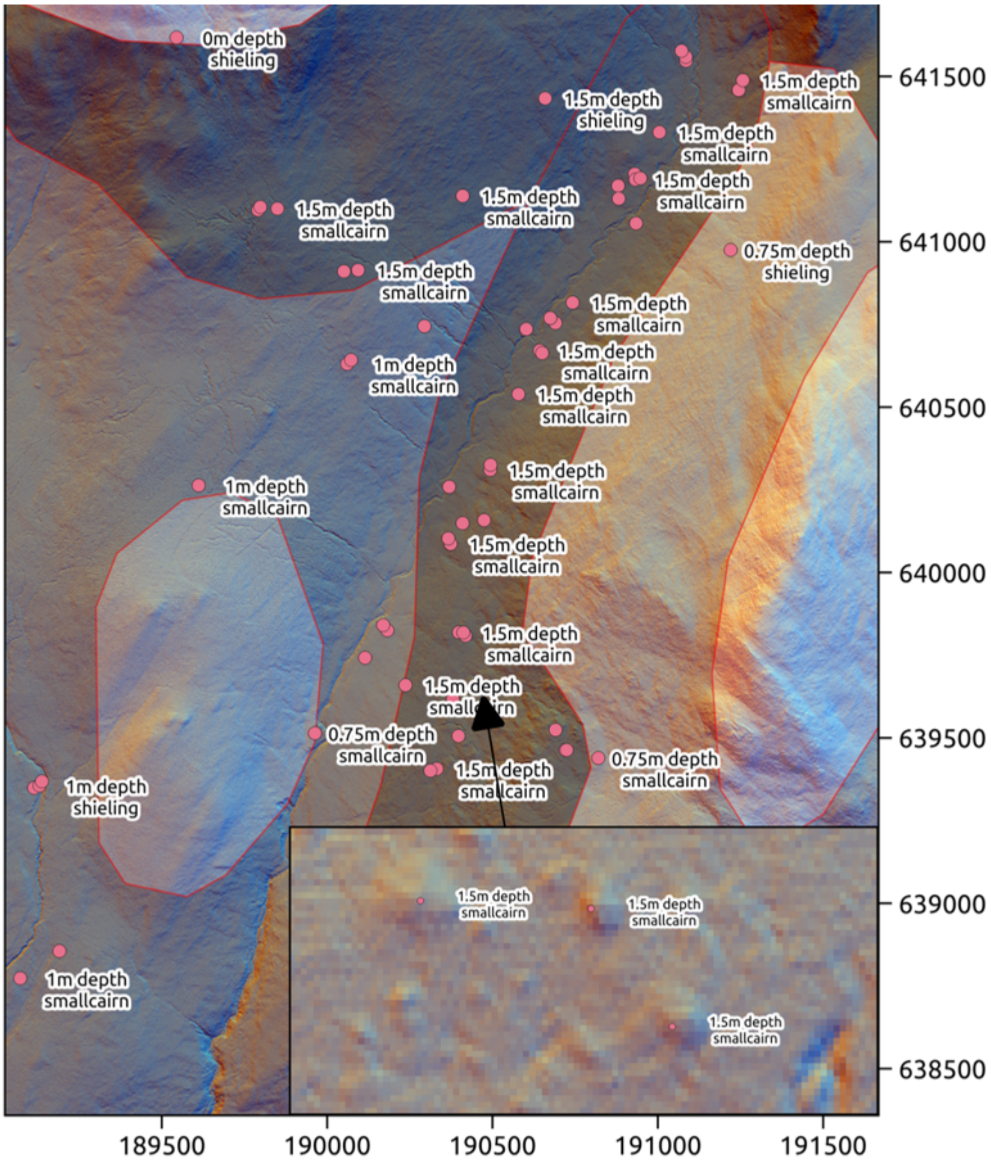


FIGURE B.2: Map showing a cluster of false positives from object detection using Multi-directional hillshade. The HES experts identified these sites as peat erosion mounds. The image is overlaid with a map of peat depth where darker colours are deeper peat. Labels show the peat depth and the class. The insert shows that the objects visually look like small cairns. ©Historic Environment Scotland.

TABLE B.1: Comparison of the known sites and detections in areas with different peat depth. The small cairn detections in red show that no sites were known at 1-1.5 meter depth but 225-21 objects respectively were detected. This confirms the pattern of false positives from peat erosion mounds.

Peat Depth	**Area (m^2)**	**Known roundhouse**	**Known shieling**	**Known smallcairn**	**Detections roundhouse**	**Detections shieling**	**Detections smallcairn**
1.5	592589818	4	71	0	57	120	225
1	7226899	7	0	0	4	8	21
0.75	62683531	50	6	155	62	52	51
0	301201636	118	141	52	386	344	370

Table B.2: Comparison of the known sites and detections in areas with different landuse. The detections in red show that no sites were known modern built up areas where many detections were made.

Historic Landuse	**Area (m^2)**	**Known roundhouse**	**Known shieling**	**Known smallcairn**	**Detections roundhouse**	**Detections shieling**	**Detections smallcairn**
Energy, Extraction and Waste	176288	0	0	0	0	6	3
Spiritual and Ritual	11817	0	0	0	0	0	0
Water Body	678577	0	0	0	1	0	0
Built-up Area	3595744	0	0	0	46	4	27
Designed Landscape	215318	0	0	0	0	0	0
Leisure and Recreation	2565834	0	0	0	30	4	38
Agriculture and Settlement	48728243	12	0	1	83	28	104
Woodland and Forestry	110507165	18	10	0	70	88	104
Moorland and Rough Grazing	262236380	149	208	205	294	398	448

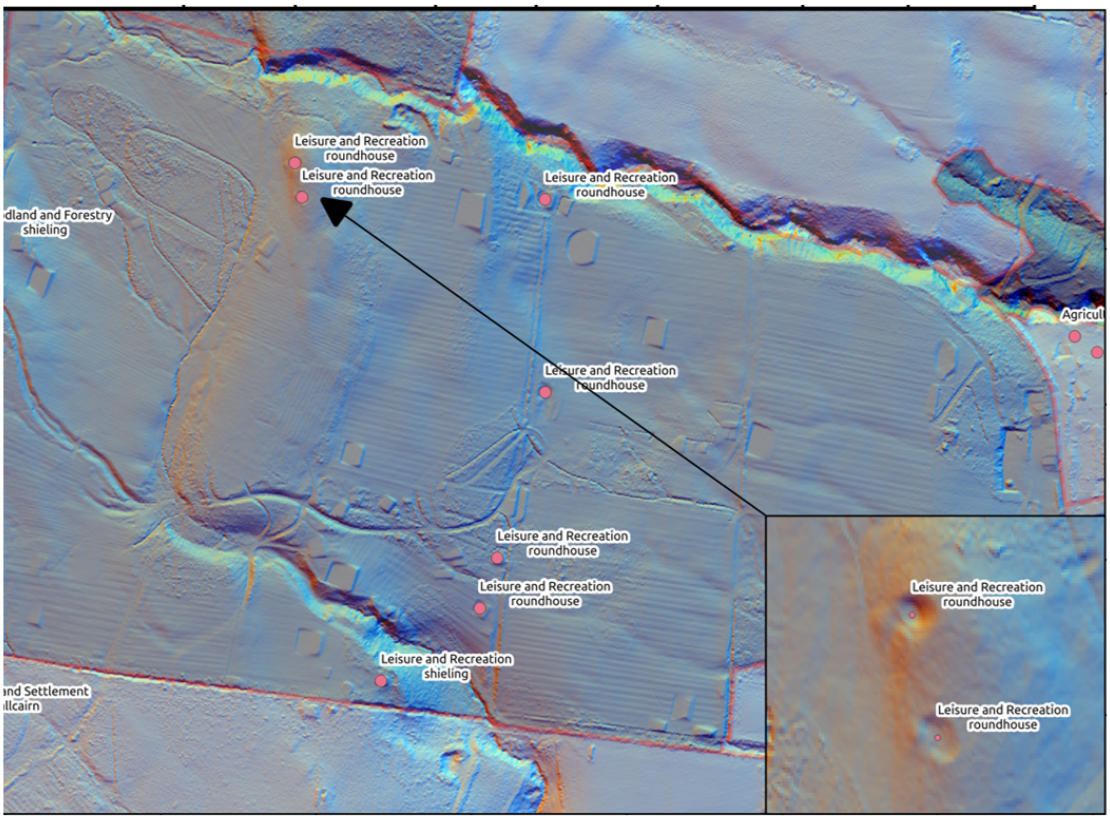

Leisure and Recreation roundhouse
Leisure and Recreation roundhouse
dland and Forestry shieling
Leisure and Recreation roundhouse
Agricul
Leisure and Recreation roundhouse
Leisure and Recreation roundhouse
Leisure and Recreation roundhouse
Leisure and Recreation shieling
nd Settlement llcairn
Leisure and Recreation roundhouse
Leisure and Recreation roundhouse

Figure B.3: Map showing a cluster of false positives from object detection using Multi-directional hillshade. The HES experts identified these sites as objects on a golf course. Labels show the Historic Landuse and the object class. The insert shows that some golf course objects have a similar size/shape to round houses but are clearly not round houses. ©Historic Environment Scotland.

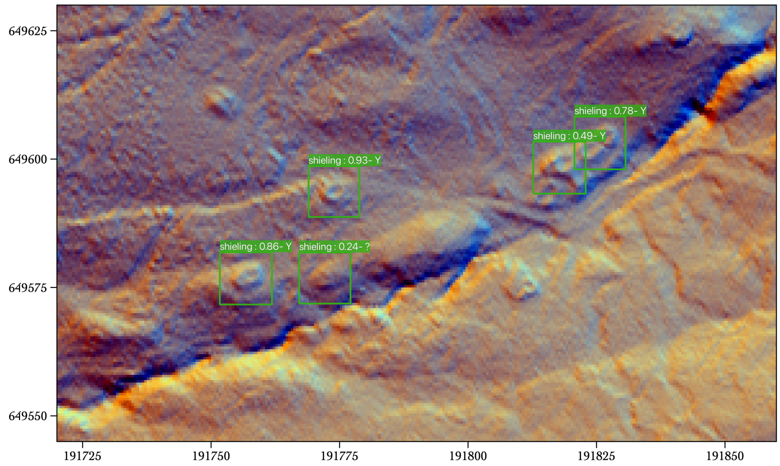


Figure B.4: Detected cluster of shieling huts in the validation area. The boxes show associated labels, predictions and the manual verification ("Y": Yes, "?": Maybe, "N": No). ©Historic Environment Scotland.

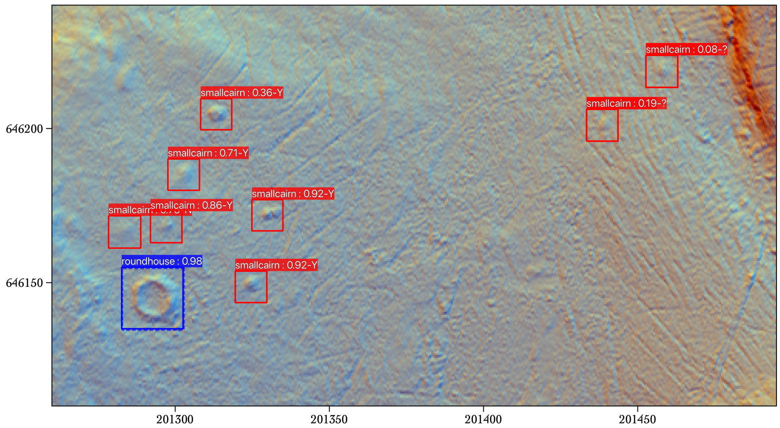


Figure B.5: Detected cairn field in the validation area. The boxes show associated labels, predictions and the manual verification ("Y": Yes, "?": Maybe, "N": No). The roundhouse was previously known and the small cairns were known in the NRHE as a cairnfield but the individual objects were not tagged. ©Historic Environment Scotland.

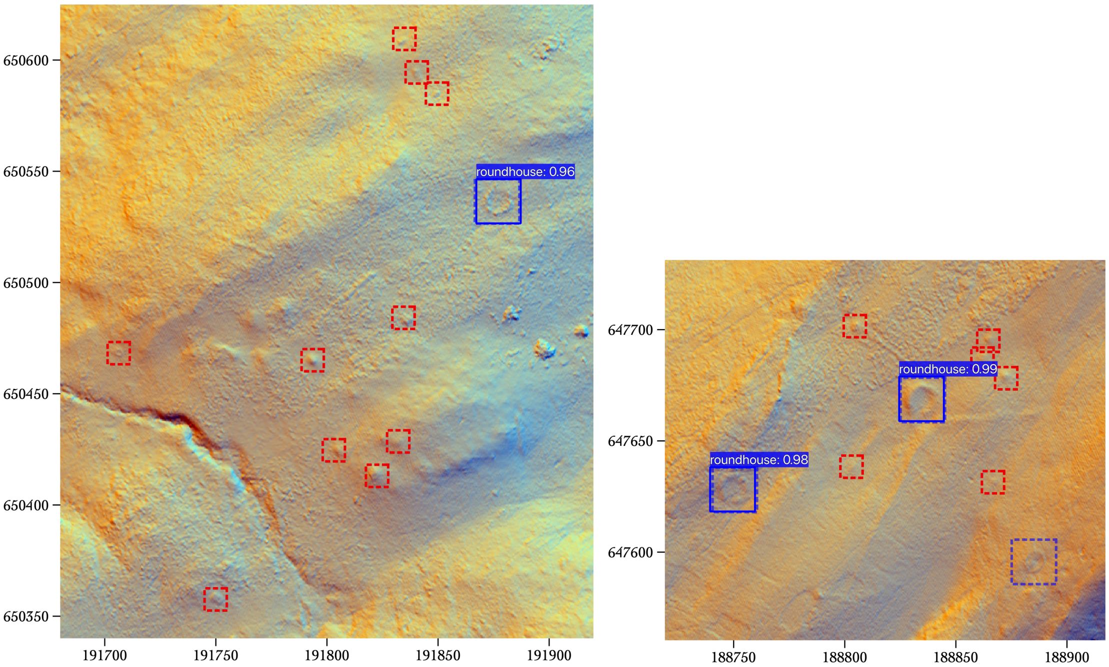


FIGURE B.6: Clusters of false negatives on cairn fields in the validation area. ©Historic Environment Scotland.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint*, page Software available from tensorflow.org., 2016.

Azavea. Raster vision. `https://github.com/azavea/raster-vision`, 2018. Accessed: 18/08/2018.

Łukasz Banaszek, Dave C Cowley, and Mike Middleton. Towards national archaeological mapping. assessing source data and methodology—a case study from scotland. *Geosciences*, 8(8):272, 2018.

Martyn Barber. *A History of Aerial Photography and Archaeology: Mata Hari's Glass Eye and Other Stories.* English Heritage Series. English Heritage, 2011. ISBN 9781848020368.

R. Bennett, D. Cowley, and V. De Laet. The data explosion: tackling the taboo of automatic feature recognition in airborne survey data. *Antiquity*, 88(341):896–905, 2014. ISSN 0003-598X.

Robert H Bewley. Aerial survey for archaeology. *The Photogrammetric Record*, 18(104): 273–292, 2003.

Peter Campbell, Christopher Stewart, and Iris Kramer. Artificial intelligence, machine learning, and deep learning in archaeology. In *Artificial Intelligence, Machine Learning, and Deep Learning in Archaeology.* British School at Rome, 2019.

Jesse Casana. Regional-scale archaeological remote sensing in the age of big data: Automated site discovery vs. brute force methods. *Advances in Archaeological Practice*, 2(3):222–233, 2014.

Jesse Casana. Global-scale archaeological prospection using corona satellite imagery: Automated, crowd-sourced, and expert-led approaches. *Journal of Field Archaeology*, 45(sup1):S89–S100, 2020.

Gino Caspari and Pablo Crespo. Convolutional neural networks for archaeological site detection–finding "princely" tombs. *Journal of archaeological science*, 110:104998, 2019.

Champion. Longdown (new forest) tumuli. `http://www.megalithic.co.uk/a558/a312/gallery/England/Hampshire/longdown_tumuli06.jpg`, 2006. Accessed: 29/10/2017.

Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020.

F. Chollet. Keras. `https://github.com/fchollet/keras`, 2015. Accessed: 01/03/2017.

F. Chollet. Building powerful image classification models using very little data. `https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html`, 2016. Accessed: 09/10/2017.

Dave Cowley, Łukasz Banaszek, George Geddes, Angela Gannon, Mike Middleton, and Kirsty Millican. 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), 2020.

Dave Cowley and Adara López-López. Developing an approach to national mapping—preliminary work on scotland in miniature. *AARGnews*, 55:19–25, 2017.

David C Cowley. What do the patterns mean? archaeological distributions and bias in survey data. In *Digital Methods and Remote Sensing in Archaeology*, pages 147–170. Springer, 2016.

O Crawford. Air survey and archaeology. *Geographical Journal*, 61(5):342–360, 1923.

Arjan de Boer. Using pattern recognition to search lidar data for archeological sites. In *The world is in your eyes. CAA 2005. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 33rd Conference, Tomar, March 2005. CAA Portugal, Tomar*, 2007.

V. de Laet, E. Paulissen, and M. Waelkens. Methods for the extraction of archaeological features from very high-resolution ikonos-2 remote sensing imagery, hisar (southwest turkey). *Journal of Archaeological Science*, 34(5):830–841, 2007. ISSN 0305-4403.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2009, Miami, FL, USA, 20-25 June 2009. IEEE, 2009. ISBN 1424439922.

M. Doneus. Openness as visualization technique for interpretative mapping of airborne lidar derived digital terrain models. *Remote Sensing*, 5(12):6427–6442, 2013. ISSN 2072-4292.

Michael Doneus, Christian Briese, Martin Fera, and Martin Janner. Archaeological prospection of forested areas using full-waveform airborne laser scanning. *Journal of Archaeological Science*, 35(4):882–893, 2008.

Dave Field. Prehistoric barrows and burial mounds. Report, 2011.

H Gaiser. Github keras retinanet. `https://github.com/fizyr/keras-retinanet`, 2019. Accessed: 01/12/2019.

Jane Gallwey, Matthew Eyre, Matthew Tonkins, and John Coggan. Bringing lunar lidar back down to earth: Mapping our industrial heritage through deep transfer learning. *Remote Sensing*, 11(17):1994, 2019.

GlobalXplorer. `https://www.globalxplorer.org`. Accessed: 01/12/2020.

GoogleMaps. Map of the united kingdom with arrow showing the isle of arran. `https://www.google.co.uk/maps`, 2020a. Accessed: 01/12/2020.

GoogleMaps. Map of the united kingdom with arrow showing the new forest. `https://www.google.co.uk/maps`, 2020b. Accessed: 01/12/2020.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017.

GroundWork. "`https://groundwork.azavea.com`. Accessed: 01/12/2020.

Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016.

R. Hesse. Lidar-derived local relief models - a new tool for archaeological prospection. *Archaeological Prospection*, 17(2):67–72, 2010. ISSN 1075-2196.

Ralf Hesse. The changing picture of archaeological landscapes: lidar prospection over very large areas as part of a cultural heritage strategy. *Interpreting Archaeological Topography: 3D Data, Visualisation and Observation; Opitz, RS, Cowley, DC, Eds*, pages 171–183, 2013.

HistoricEngland. `https://historicengland.org.uk/whats-new/news/hot-dry-summer-reveals-hidden-archaeological-sites`, 2018. Accessed: 10/12/2019.

HistoricEngland. Lidar (light detection and ranging). `https://historicengland.org.uk/research/methods/airborne-remote-sensing/lidar/`, 2018b. Accessed: 03/02/2018.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, and Sergio Guadarrama. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, volume 4, 2017.

Juha Hyyppa, Hannu Hyyppa, Xiaowei Yu, Harri Kaartinen, Antero Kukko, and Markus Holopainen. *Forest Inventory Using Small-Footprint Airborne LiDAR*, book section 12. CRC Press/Taylor and Francis Group, Boca Raton, 1st edition, 2009.

Matthew Johnson. *Ideas of landscape.* John Wiley & Sons, 2007.

Karpathy. Neural networks part 1: Setting up the architecture. `http://cs231n.github.io/neural-networks-1/`, 2018. Accessed: 03/08/2018.

B Kazimi, F Thiemann, and M Sester. Semantic segmentation of manmade landscape structures in digital terrain models. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2019.

Žiga Kokalj and Maja Somrak. Why not a single image? combining visualizations to facilitate fieldwork and on-screen mapping. *Remote Sensing*, 11(7):747, 2019.

Iris Kramer. *An archaeological reaction to the remote sensing data explosion. Reviewing the research on semi-automated pattern recognition and assessing the potential to integrate artificial intelligence.* Thesis, 2015.

Iris Kramer and Jonathon Hare. Arran. `https://github.com/ickramer/Arran`, 2020. Accessed: 09/01/2020.

Iris Caroline Kramer. Using ecognition to improve feature recognition. In *CProceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2016)*, 2016.

Iris Caroline Kramer. A future perspective for automation in large mapping projects with feature learning. In *CProceedings of the 45th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2017)*, 2017.

Iris Caroline Kramer. The basics of deep learning for archaeological site detection on remote sensor data. In *Proceedings of the 46th Conference on Computer Applications* *and Quantitative Methods in Archaeology (CAA 2018)*. Computer Applications and Quantitative Methods in Archaeology, 2018a.

Iris Caroline Kramer. Tackling the small data problem in deep learning with multi-sensor approaches" presented at track. In *CProceedings of the 46th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2018)*, 2018b.

Iris Caroline Kramer, Wouter Verschoof-Van Der Vaart, and Alex Brandsen. Challenges and opportunities of machine learning in archaeological research. In *Proceedings of the 47th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2019)*. Computer Applications and Quantitative Methods in Archaeology, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems, NIPS 2012, Lake Tahoe, NV,USA, December 3-6, 2012*, 2012.

Karsten Lambers, Wouter B Verschoof-van der Vaart, and Quentin PJ Bourgeois. Integrating remote sensing, machine learning, and citizen science in dutch archaeological prospection. *Remote Sensing*, 11(7):794, 2019.

Albert Yu-Min Lin, Andrew Huynh, Gert Lanckriet, and Luke Barrington. Crowdsourcing the unknown: The satellite search for genghis khan. *PloS one*, 9(12):e114046, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision, ECCV 2016 , Amsterdam, The Netherlands, October 11-14, 2016*. Springer, 2016.

Kirsty Millican, Piers Dixon, Lesley Macinnes, and Mike Middleton. Mapping the historic landscape: Historic land-use assessment in scotland. *Landscapes*, 18(1):71–87, 2017.

Volodymyr Mnih. *Machine learning for aerial image labeling*. Citeseer, 2013.

E Özdemir and F Remondino. Classification of aerial point clouds with deep learning. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

Sarah H Parcak. *Satellite remote sensing for archaeology*. Routledge, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Otávio AB Penatti, Keiller Nogueira, and Jefersson A dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2015, Boston, MA, USA, June 7-12, 2015*, 2015.

QGIS Development Team. *QGIS Geographic Information System.* QGIS Association, 2020.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.*

Nannan Qin, Xiangyun Hu, and Hengming Dai. Deep fusion of multi-view and multimodal representation of als point cloud for 3d terrain scene recognition. *ISPRS journal of photogrammetry and remote sensing*, 143:205–212, 2018.

Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*, 2014.

Melanie A. Riley. *Automated Detection of Prehistoric Conical Burial Mounds: From LIDAR Bare-Earth Digital Elevation Models.* Thesis, 2009.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, Munich, Germany, October 5-9, 2015.* Springer, 2015.

Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 702–709, 2019.

Christopher Sevara, Michael Pregesbauer, Michael Doneus, Geert Verhoeven, and Immo Trinks. Pixel versus object—a comparison of strategies for the semi-automated mapping of archaeological features using airborne laser scanning data. *Journal of Archaeological Science: Reports*, 5:485–498, 2016. ISSN 2352-409X.

Benoit Sittler. Revealing historical landscapes by using airborne laser scanning. *Laser-Scanners for Forest and Landscape Assessment*, pages 258–261, 2004.

Maja Somrak, Sašo Džeroski, and Žiga Kokalj. Learning to classify structures in als-derived visualizations of ancient maya settlements with cnn. *Remote Sensing*, 12(14): 2215, 2020.

Mehrnoush Soroush, Alireza Mehrtash, Emad Khazraee, and Jason A Ur. Deep learning in archaeological remote sensing: Automated qanat detection in the kurdistan region of iraq. *Remote Sensing*, 12(3):500, 2020.

Christopher Stewart, Georges Labrèche, and Daniel Lombraña González. A pilot study on remote sensing and citizen science for archaeological prospection. *Remote Sensing*, 12(17):2795, 2020.

T Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

Arianna Traviglia et al. Archaeological usability of hyperspectral images: Successes and failures of image processing techniques. *BAR International Series*, 1568:123, 2006.

Arianna Traviglia and Karsten Lambers. Computer vision vs human perception in remote sensing image analysis: Time to move on. In *Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2016)*. Computer Applications and Quantitative Methods in Archaeology, 2016.

Arianna Traviglia and Karsten Lambers. Automation is here to stay! the hitch-hiker's guide to automated object detection and image processing in remote sensing. In *Proceedings of the 45th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA2017)*. Computer Applications and Quantitative Methods in Archaeology, 2017.

Arianna Traviglia and Karsten Lambers. Setting the automation agenda for remote sensing: learning to see through a computer. In *Proceedings of the 46th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA 2018)*. Computer Applications and Quantitative Methods in Archaeology, 2018.

ØD Trier, AB Salberg, and LH Pilø. Semi-automatic mapping of charcoal kilns from airborne laser scanning data using deep learning. In *CAA2016: Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology*, pages 219–231, 2018.

Øivind Due Trier, David C Cowley, and Anders Ueland Waldeland. 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, 2019.

Øivind Due Trier and L. H. Pilo. Automatic detection of pit structures in airborne laser scanning data. *Archaeological Prospection*, 19(2):103–121, 2012. ISSN 1075-2196.

Øivind Due Trier, Arnt-Børre Salberg, Lars Holger Pilø, Christer Tonning, Hans Marius Johansen, and Dagrun Aarsten. Semi-automatic mapping of cultural heritage from airborne laser scanning using deep learning. In *EGU General Assembly Conference*, volume 18, 2016.

Øivind Due Trier, Maciel Zortea, and Christer Tonning. Automatic detection of mound structures in airborne laser scanning data. *Journal of Archaeological Science: Reports*, 2(0):69–79, 2015. ISSN 2352-409X.

Geert J Verhoeven. Are we there yet? a review and assessment of archaeological passive airborne optical imaging approaches in the light of landscape archaeology. *Geosciences*, 7(3):86, 2017.

Wouter B Verschoof-van der Vaart, Karsten Lambers, Wojtek Kowalczyk, and Quentin PJ Bourgeois. 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, 2020.

Wouter Baernd Verschoof-van der Vaart and Karsten Lambers. 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), 2019.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems, NIPS 2014, Montreal, Quebec, Canada, December 8-13 2014*, 2014.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. ISSN 2168-6831.

Igor Zingman, Dietmar Saupe, Otávio AB Penatti, and Karsten Lambers. Detection of fragmented rectangular enclosures in very high resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4580–4593, 2016. ISSN 0196-2892.