AGU ADVANCING EARTH AND SPACE SCIENCES

Earth and Space Science

RESEARCH ARTICLE
10.1029/2023EA003220

**Special Section:**
Advances in Machine Learning for Earth Science: Observation, Modeling, and Applications

# Leveraging Spatial Metadata in Machine Learning for Improved Objective Quantification of Geological Drill Core

**Lewis J. C. Grant[1], Miquel Massot-Campos[2], Rosalind M. Coggon[1], Blair Thornton[2,3], Francesca C. Rotondo[1], Michelle Harris[4], Aled D. Evans[1], and Damon A. H. Teagle[1]**

[1]School of Ocean and Earth Science, National Oceanography Centre Southampton, University of Southampton, Southampton, UK, [2]School of Engineering, University of Southampton, Southampton, UK, [3]Institute of Industrial Science, The University of Tokyo, Tokyo, Japan, [4]School of Geography, Earth and Environmental Sciences, Plymouth University, Plymouth, UK

**Key Points:**

- Use of spatial metadata improves state-of-the-art unsupervised feature extraction
- Semi-supervised methods using spatial metadata outperform supervised methods for the same expert labeling effort
- New methods described enable the standardization of core-logging processes and improve cross-dataset comparisons

**Correspondence to:**
L. J. C. Grant,
lg1e20@soton.ac.uk

**Citation:**
Grant, L. J. C., Massot-Campos, M., Coggon, R. M., Thornton, B., Rotondo, F., Harris, M., et al. (2024). Leveraging spatial metadata in machine learning for improved objective quantification of geological drill core. *Earth and Space Science*, *11*, e2023EA003220. https://doi.org/10.1029/2023EA003220

Received 5 OCT 2023
Accepted 5 FEB 2024

**Author Contributions:**
**Conceptualization:** Lewis J. C. Grant, Miquel Massot-Campos, Rosalind M. Coggon, Francesca C. Rotondo, Michelle Harris, Aled D. Evans, Damon A. H. Teagle
**Data curation:** Lewis J. C. Grant, Blair Thornton
**Formal analysis:** Lewis J. C. Grant
**Funding acquisition:** Rosalind M. Coggon, Damon A. H. Teagle
**Investigation:** Lewis J. C. Grant
**Methodology:** Lewis J. C. Grant, Miquel Massot-Campos, Rosalind



**Abstract** Here we present a method for using the spatial *x*–*y* coordinate of an image cropped from the cylindrical surface of digital 3D drill core images and demonstrate how this spatial metadata can be used to improve unsupervised machine learning performance. This approach is applicable to any data set with known spatial context, however, here it is used to classify 400 m of drillcore imagery into 12 distinct classes reflecting the dominant rock types and alteration features in the core. We modified two unsupervised learning models to incorporate spatial metadata and an average improvement of 25% was achieved over equivalent models that did not utilize metadata. Our semi-supervised workflow involves unsupervised network training followed by semi-supervised clustering where a support vector machine uses a subset of *M* expert labeled images to assign a pseudolabel to the entire data set. Fine-tuning of the best performing model showed an $f_1$ (macro average) of 90%, and its classifications were used to estimate bulk fresh and altered rock abundance downhole. Validation against the same information gathered manually by experts when the core was recovered during the Oman Drilling Project revealed that our automatically generated data sets have a significant positive correlation (Pearson's *r* of 0.65–0.72) to the expert generated equivalent, demonstrating that valuable geological information can be generated automatically for 400 m of core with only ~24 hr of domain expert effort.

**Plain Language Summary** This work presents a novel method for using the spatial context of digital core images to improve the descriptive accuracy of unsupervised machine learning algorithms. The addition of spatial metadata improves model performance by an average of 25%, with the best performing model in this study achieving an accuracy score of 90%. The output of this model was then used to estimate the amount of fresh and altered rock within a 400 m long drill core, which was shown to be of comparable quality to the same estimations made by geologists on the cores themselves.

## 1. Introduction

Drilling into the Earth to recover cores for geological analysis is an essential tool that provides valuable insight into otherwise inaccessible environments, yielding data sets utilized for mining, infrastructure planning, and reconstructing the history of the planet. The task of describing these cores falls to specialists who systematically work through the recovered material to produce a series of descriptive and quantitative logs (core-logging) as well as visual core descriptions (VCDs). The features documented may include, but are not limited to, changes in rock type, veins and alteration features, structural measurements, and variations in relative mineral abundance downhole. These tasks are time consuming and rely on subjective estimates of the abundance of key features within a core. Furthermore, human interpretation tends to overestimate the abundance of a given feature in a scene, causing estimates to vary widely between individuals (Finn et al., 2010; Olmstead et al., 2004) and objective automated methods could resolve this underlying bias. In addition to VCDs, cores are digitally imaged, and in the case of scientific drilling, their physical properties are measured prior to detailed petrographic and geochemical analyses (Coggon et al., 2024; Jarrard et al., 2003; Kelemen et al., 2020), but additional processing is needed to make these data sets machine readable, limiting their use in emerging machine learning applications. During drilling campaigns downhole wireline geophysical logs of the borehole wall may also be collected, providing useful continuous data sets for comparing borehole features with recovered core material to compensate for incomplete core recovery (Tominaga & Umino, 2010; Tominaga et al., 2009). Most attempts to automate the classification of rock-types downhole initially focused on applying artificial neural networks (ANN) to one-

M. Coggon, Blair Thornton, Damon A. H. Teagle
**Project administration:** Lewis J. C. Grant
**Resources:** Lewis J. C. Grant
**Software:** Lewis J. C. Grant, Miquel Massot-Campos
**Supervision:** Rosalind M. Coggon, Blair Thornton, Damon A. H. Teagle
**Validation:** Lewis J. C. Grant, Rosalind M. Coggon
**Visualization:** Lewis J. C. Grant, Miquel Massot-Campos, Michelle Harris
**Writing – original draft:** Lewis J. C. Grant, Miquel Massot-Campos, Rosalind M. Coggon, Blair Thornton, Francesca C. Rotondo, Damon A. H. Teagle
**Writing – review & editing:** Lewis J. C. Grant, Miquel Massot-Campos, Rosalind M. Coggon, Blair Thornton, Francesca C. Rotondo, Michelle Harris, Aled D. Evans, Damon A. H. Teagle

dimensional borehole data (Al-Mudhafar, 2017; J. He et al., 2019; Ma, 2011; Tominaga et al., 2009). However, using only numerical data has the limitation of providing less direct information about the rock when compared to core images (Chai et al., 2009; Thomas et al., 2011).

Most recent efforts to automatically classify rock-types using images of drill core have utilized convolutional neural networks (CNNs) as they are more suited to image analysis (LeCun & Bengio, 1995). When training a CNN to classify images, there are three main types of machine learning; supervised, unsupervised and semi-supervised, which involves a combination of unsupervised learning followed by a less intensive supervised step (Camps-Valls et al., 2007). The initial 'learning' stage of training is where a CNN determines which images it considers similar and dissimilar, however, additional steps are required to assign classifications or labels to the images. In supervised learning, each training image has been labeled to give the model a target output to work toward, however this requires significant effort on the part of the annotator. In contrast, unsupervised learning does not involve any labeling effort as the network extracts salient information from each image, referred to as a latent representation, and clustering techniques allow grouping of images based on these simplified representations. An expert then inspects these clusters and provides a label to each. When taking a semi-supervised approach, a subset of expert labeled images can be provided to an unsupervised model to allow it to both cluster and assign a label to all images. Images are labeled based on where their latent representations plot in the hyper-dimensional feature space relative to the expert labeled subset. To date, there have been numerous attempts to use neural networks to classify images of drill core, all of which have taken slightly different approaches.

Zhang et al. (2017) used a supervised approach to train a CNN to classify a data set of 1500 2D grayscale borehole wall resistivity images into three texturally distinct sedimentary rock types (sandstone, shale, and conglomerate). Their number of training images was class imbalanced with an order of magnitude more sandstone images used in an attempt to improve their model's ability to identify potential hydrocarbon reservoirs. Similarly, Alzubaidi et al. (2021) used a supervised workflow to compare the performance of several CNN model architectures in identifying three sedimentary rock types in photos of boxed core sections (box photos) with the ResNeXt-50 CNN architecture out-performing other networks. Their training data set consisted of 76,500 (25,500 per class) 2 $cm^2$ patches cropped from the box photos and all models were trained to identify non-core artifacts in the images to avoid them being labeled as classes of geological interest. Although this work showed promising results, such models are only capable of classifying a few distinct classes of rock and consequently have only limited applicability to more complex image data sets that display greater variability of geological features. Most recently, Fu et al. (2022) demonstrated a supervised workflow based on fine-tuning CNNs to identify 10 rock types commonly encountered during subsurface engineering projects. Their work showed ResNeSt-50 produced the best prediction accuracy of 99.6%. Supervised training of models requires careful preparation of the input data by an expert to ensure each desired class is well represented. For this reason, Fu et al. (2022) trained their models using 15,000 3 $cm^2$ labeled images of best-case examples of each rock type having first discarded images not of interest, such as crushing structures and crayon marks. Images removed from the training data set were also defined based on what the authors believed would confuse the CNNs and cause them to mis-classify features of interest.

A concerted effort to label a large database of images of all known rock types would provide a widely applicable training data set, however, unlike in satellite imagery and object recognition research, there are no publicly available training data sets for classifying common rock types in drill core (Deng et al., 2009; Van Etten et al., 2018). This is partly because resources are rarely put toward labeling such data sets, but also because it is difficult to combine individual data sets with variable resolution and quality, often stored in different media and file formats, into a single database. In response to these limitations, this study is intended to provide researchers with a means of analyzing large numbers of images on a per-dataset basis with minimal effort in the hope that widely applicable training data sets of rock images can begin to emerge. Furthermore, use of spatial information alongside numerical data sets have been shown to improve the automatic classification of geological information stored in the data (Hill et al., 2015, 2021; Yamada et al., 2021), and here we make a first attempt at leveraging spatial information when classifying digital geological core imagery.

In this study we modify two unsupervised learning frameworks originally designed to use 3D geolocational metadata for improved semantic interpretation of seafloor imagery (Yamada et al., 2021; Yamada, Massot-Campos, et al., 2022; Yamada, Prügel-Bennett, et al., 2022) to instead use the x-y coordinate of where an image lies on the surface of a 3D drill core image. The first framework uses an autoencoder that was trained both

**Table 1**
*List of the Machine Learning Models Used in This Study*

| CNN framework | Feature extraction | Spatial metadata | Reference |
|---|---|---|---|
| Autoencoder (AE) | Unsupervised (autoencoder) | N | Yamada et al. (2021) |
| Location-guided autoencoder (LGA) | Unsupervised (autoencoder) | Y | Yamada et al. (2021) |
| SimCLR | Unsupervised (contrastive learning) | N | Chen et al. (2020) |
| GeoCLR | Unsupervised (contrastive learning) | Y | Yamada, Prügel-Bennett, et al. (2022) |
| ResNet18 | Supervised | N | K. He et al. (2016) |

*Note.* The feature extraction column identifies whether the model learns with (supervised) or without (unsupervised) domain expert input and the spatial metadata column identifies models which utilize spatial information accompanying images during training (Y = yes, N = no). The references provided are those that outline the original development of each model.

with and without the addition of this spatial metadata, whereas the second uses two contrastive learning methods, one that makes use of metadata, and another that does not (Table 1). The performance of each framework is reviewed to determine which is most accurate and we present a novel semi-supervised workflow for training CNNs using images accompanied by spatial metadata. The output of the best performing model is then used to automatically generate a downhole log of hydrothermal alteration extent, which is bench-marked against expert generated alteration logs.

## 2. Methods

### 2.1. Background

#### 2.1.1. Artificial Neural Networks

An ANN is a computer model inspired by the structure of the human brain and consists of multiple layers of stacked artificial neurons, also referred to as perceptrons or nodes (Rosenblatt, 1962). Each artificial neuron is a mathematical model that takes multiple binary inputs ($x$) and gives a binary output determined by whether the weighted sum of the inputs meet some threshold value ($t$). The weight ($w$) assigned to a given input expresses its importance to the output, and the weight and threshold parameters can be adjusted to customize a model to a particular task. To exert control on how easily a neuron will give a 1, the threshold is often replaced by a bias ($b \equiv -t$) and the neuron's activation function is expressed using the following dot product:

$$\text{output} = \begin{cases} 0, & \text{if } w \cdot x + b < 0, \\ 1, & \text{if } w \cdot x + b > 0. \end{cases} \quad (1)$$

The layers of stacked neurons between the input and output layers of an ANN are called hidden layers and each neuron in a hidden layer receives its input from every neuron of the previous layer. Therefore, each neuron in an ANN is fully connected to each neuron in the adjacent layers (Figure 1a) (Krogh, 2008). By using weights, biases and activation functions, each hidden layer extracts features within its input, and multiple hidden layers make a flexible model capable of identifying complex patterns within a data set. The final output layer of an ANN provides a prediction for the information passed through the hidden layers, and the number of neurons in this layer depends on the application. In the case of a binary classification model the last layer would contain only two nodes, but for more complex cases the number of nodes will be equal to the number of potential classes in the input data. One drawback of using ANNs for image processing is that each neuron possesses a unique weight and bias, requiring great processing power due to the large number of parameters handled by the model. This, as well as the fact that ANNs do not achieve spatial invariance, that is they are unable to recognize features regardless of their specific location in an image, limit their use in computer vision applications.

#### 2.1.2. Convolutional Neural Networks

A CNN is a type of deep ANN developed in the early 1990s that can account for the spatial structure of input data (LeCun & Bengio, 1995). Innovations over the last decade have made CNNs increasingly popular for computer vision tasks, as their architecture is particularly suited for image analysis (Krizhevsky et al., 2012; Russakovsky

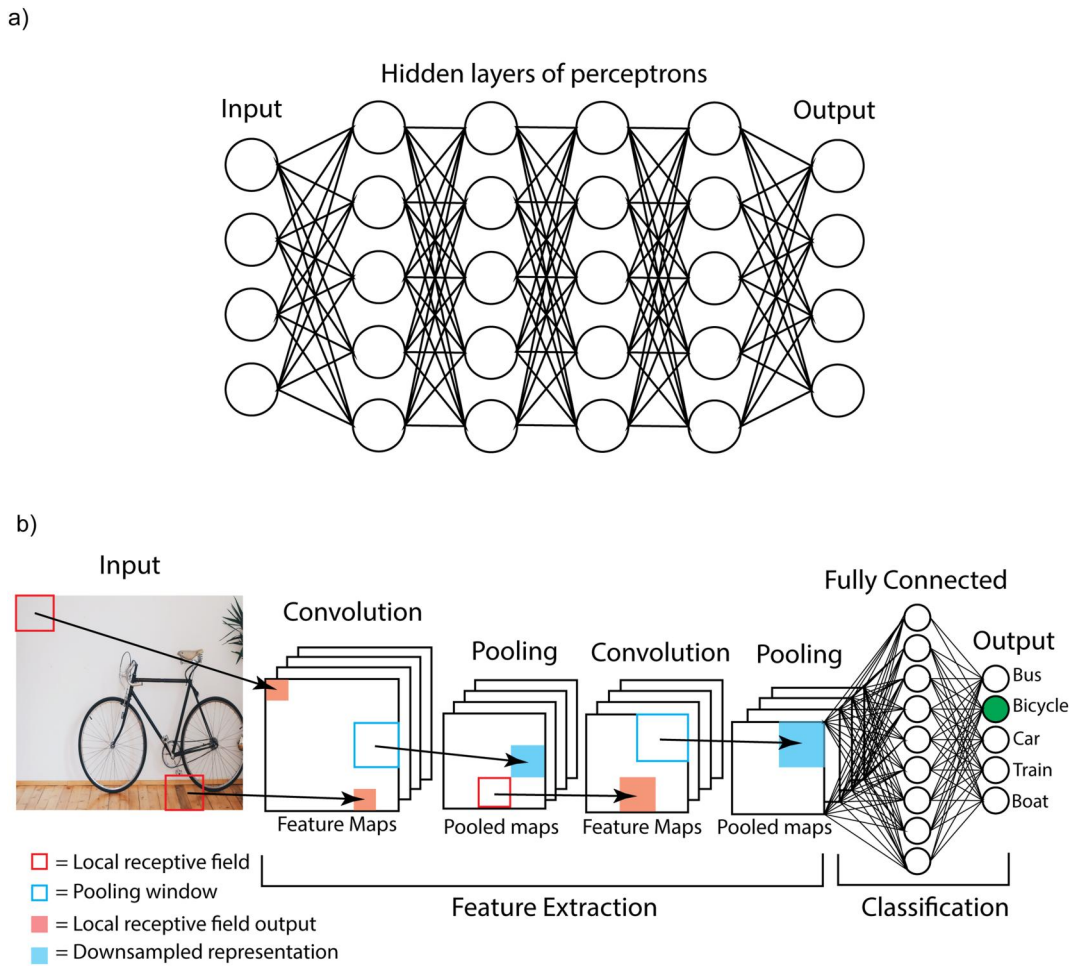


**Figure 1.** Diagrams showing (a) the basic structure of an artificial neural network and (b) a convoluted neural network.

et al., 2015). Unlike the fully connected layers of an ANN, each neuron in a CNN's first hidden layer corresponds to a rectangular region of a defined size and location in the input image (Figure 1b). This rectangular region is processed by a convolutional filter or kernel with a weight and bias and is known as a "local receptive field." The local receptive field then moves across the input neurons while keeping the same weight and bias when mapping information to its corresponding neuron in the hidden layer. Iterating this process across an image (convolution) creates a hidden layer (feature map) of neurons capable of detecting the same feature anywhere in the image. Each hidden layer of a CNN can contain multiple feature maps, and at shallow levels they can detect simple features, such as lines and shapes, whereas at deeper layers increasingly more complex features become identifiable. Convolution layers are then followed by pooling layers that simplify the output of each feature map by summarizing a specified sub-region into a condensed feature map (Figure 1b). Pooling in a CNN is a downsampling operation that reduces spatial dimensions while retaining important features, aiding in computational efficiency, and promoting robustness and generalization of the network. The major benefit of using local receptive fields and pooling layers, is that they make CNNs well adapted to handle spatial invariance within an image. The feature maps created by convolution-pooling layers are multi-dimensional arrays (tensors), which make them suitable for identifying complex features, but not for assigning class scores or probabilities. Therefore, the tensors produced by the last convolution-pooling layer are flattened into a one-dimensional vector that is fed into a fully connected layer of neurons. This fully connected layer transforms its input into high-level features that can be used for classification and regression. The output layer of a CNN is also a fully connected layer consisting of as many neurons as possible classes, and the input image is classified depending on which of these neurons is triggered by its activation function (Figure 1b).

### 2.1.3. Unsupervised Machine Learning

In computer vision, there are many publicly available databases of labeled images, such as ImageNet, MS COCO, and CIFAR-100, that can be used to train CNNs to classify common objects. However, a supervised approach cannot be used when the classes within these data sets have no relevance to the application domain. In fields such as geology, there are no large labeled data sets of rock images available to pre-train a model and the labeling effort required to generate enough training images for supervised approaches would be too time consuming, particularly in the context of a drilling campaign. This is because a given campaign often involves drilling numerous holes that may yield hundreds or thousands of meters of complex drill core, all of which needs describing by an expert. A solution to this is to utilize unsupervised CNN frameworks capable of extracting salient information from geological images without any prior labeling effort, and two such frameworks include autoencoders and contrastive learning.

### 2.1.4. Autoencoders

An autoencoder (AE) is a form of neural network architecture used for unsupervised learning and dimensionality reduction that consists of two elements. Firstly, an encoder ($f$) that takes an input ($x$) and compresses it into a lower dimensional representation, or latent representation ($h = f_\phi(x)$) (Figure 2Bii). Secondly, a decoder ($g$) which uses the latent representation to re-construct the input to give $x_r = g_\theta(h)$ (Figure 2Biii), where $\phi$ and $\theta$ are the parameters of the encoder and decoder, respectively. Where the input data is continuous $\left(\{x\}_{i=1}^{n}\right)$, the difference between $x$ and $x_r$ (reconstruction loss) can be calculated using the mean square error, making the optimizing objective (loss function) of the AE:

$$\min_{\phi,\theta} L_{rec} = \min \frac{1}{n} \sum_{i=1}^{n} \|x_i - x_{ri}\|^2 \tag{2}$$

The major objective of network training in machine learning is to find the minimum loss. Clustering techniques are often used to improve the grouping of similar datapoints in latent space by using both the reconstruction loss ($L_{rec}$) and clustering loss ($L_{clust}$) (Aljalbout et al., 2018; Min et al., 2018). The purpose of $L_{rec}$ is to learn realistic features, whereas $L_{clust}$ promotes discrimination and grouping of feature points within the latent space (Min et al., 2018). When using deep clustering, the loss function becomes:

$$L_{all} = (1 - \lambda) L_{rec} + \lambda L_{clust} \tag{3}$$

where $\lambda \in \{0,1\}$ is a hyperparameter that balances $L_{rec}$ and $L_{clust}$ and should be set to prevent over/under fitting of the model for a given data set. If set too low, over-fitting will occur as the model has learned too much about the noise in the data, limiting its ability to identify characteristic features of each class. In contrast, if set too high under-fitting occurs as the model becomes too simplistic and overlooks key patterns in the data. $L_{clust}$ can be obtained by calculating the Kullback-Leibler (KL) divergence loss between the soft assignment probability of sample $i$ belonging to cluster $j$ with an auxiliary target distribution using the following equation (Xie et al., 2016):

$$L_{clust} = KL(P||Q) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}} \tag{4}$$

where $p_{ik}$ and $q_{ik}$ are the $i$th sample of the $k$th cluster of the target ($P$) and soft ($Q$) probability distributions (Van der Maaten & Hinton, 2008). Calculating all soft assignments for a sample produces probability distribution $Q$, whereas the target probabilistic distributions ($P$) are derived by squaring $q_{ik}$ and normalizing by the sum of its soft cluster frequencies:

$$q_{ik} = \frac{\left(1 + \|h_i - \mu_k\|^2\right)^{-1}}{\sum_{k'} \left(1 + \|h_i - \mu_{k'}\|^2\right)^{-1}} \tag{5}$$

$$p_{ik} = \frac{q_{ik}^2 / f_k}{\sum_{k'} q_{ik'}^2 / f_{k'}} \tag{6}$$

10.1029/2023EA003220

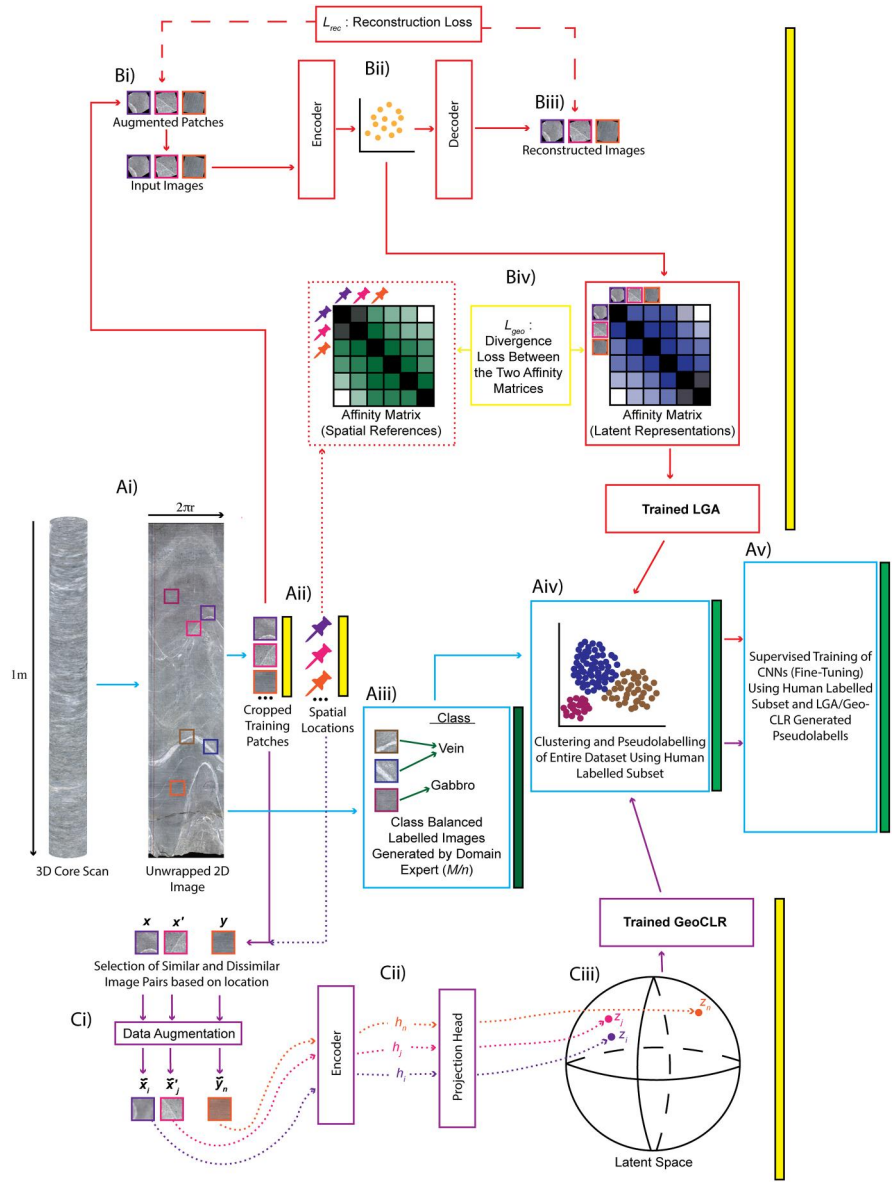


**Figure 2.** Diagram of the semi-supervised workflow used in this study. Unsupervised feature extraction used both the location-guided autoencoder (LGA) (red path) and GeoCLR (purple path) to create latent representations of the data set, and the workflow involves image processing, sampling, labeling and clustering prior to fine tuning of the pre-trained convolutional neural networks (CNNs) generated by LGA and GeoCLR (blue path). Yellow bars indicate completely automated steps, whereas dark and light green bars indicate supervised and semi-supervised steps, respectively. Modified after Yamada et al. (2021, 2022).

where $h_i = f_\phi(x_i)$, $\mu_k$ is the centroid of cluster $k$, and $f_k = \sum_i q_{ik}$ is the soft cluster frequency. Making use of $h$, which is a compact version of the original input, allows auto-encoders to pick out only the most salient features in the data.

### 2.1.5. Location Guided Autoencoder

Spatial information is important in many applications, and while CNNs can find patterns within an image, many spatial patterns are larger than the footprint of a single image cropped from a larger scene and CNNs cannot correlate these patterns. In response, Yamada et al. (2021) developed a novel location-guided autoencoder (LGA) for automated semantic interpretation of seafloor images that utilizes 3D geolocational metadata. Their base autoencoder for feature extraction uses AlexNet (Krizhevsky et al., 2012), where the encoder is AlexNet's original

architecture and the decoder is an inverted version of the encoder (Figure 2Bii). The LGA was designed with the assumption that "two images captured close together look more similar than those far apart." Using this assumption, the position of data in the latent space ($h_i$ and $h_j$) is modified by accounting for the distance between the locations ($y_i$ and $y_j$) of the original images ($x_i$ and $x_j$) (Figure 2Aii). The assumption can then be applied by using a Gaussian distribution as a kernel to quantify the affinity between $h$ and geographical space ($y$) (Figure 2Biv):

$$q'_{ij} = \frac{(1 + \|h_i - h_j\|)^{-1}}{\sum_{i'} \sum_{j'} (1 + \|h_{i'} - h_{j'}\|^2)^{-1}} \tag{7}$$

$$p'_{ij} = \frac{(1 + d(y_i\, y_j))^{-1}}{\sum_{i'} \sum_{j'} (1 + d(y_{i'}\, y_{j'}))^{-1}} \tag{8}$$

where $q'_{ij}$ and $p'_{ij}$ are the values of the affinity matrices at index $(i, j)$ in the latent space ($Q'$) and physical space ($P'$) respectively, and $d(y_i, y_j) = \min\| y_i, y_j\|^2 d^2_{\max}$. In this context $d_{\max}$ is the user-defined maximum distance between two locations that will be corrected and will vary on the application domain and scale of the image scene. The LGA is trained to minimize the KL divergence between $Q'$ and $P'$ using the following loss function:

$$L_{all} = L_{rec} + \lambda L_{geo} = L_{rec} + KL(P' \| Q') \tag{9}$$

This approach results in $h_i$ and $h_j$ being moved closer together in feature space if they are close in physical space.

### 2.1.6. Contrastive Learning

Contrastive learning is an unsupervised machine learning technique that attempts to learn features in an image by comparing similar pairs of images close together in $h$ to a random dissimilar pair embedded far apart in $h$. The aim of this comparison is to maximize the similarity between positive pairs (images that look similar) and minimize the similarity between negative pairs (images that look dissimilar). An issue with contrastive learning is that you must confirm that the positive pair of images are indeed similar. In response, Chen et al. (2020) developed a framework for self-supervised contrastive learning of visual representations (SimCLR) that attempts to improve agreement between variably augmented images ($x_i$ and $x_j$) derived from the same original image ($x$). At each training iteration, a minibatch (i.e., a small subset) of $N$ images is taken for augmentation. During augmentation, random cropping, color distortion and Gaussian blur are applied before a CNN is used as a base encoder ($f()$) that extracts representations, known as feature vectors ($h_i$), from the augmented images ($h_i = f(x_i)$) (Figure 2Cii). These vectors then act as the input for a projection head ($g()$) consisting of a two-layer multi-layer perceptron (MLP), which produces an embedding ($z_i = g(h_i)$) that is mapped to a latent space (Figure 2Ciii) where the following loss function is applied to compute the contrastive loss ($\ell$):

$$\ell_{i,j} = -\log \left( \frac{\exp\left(\frac{\text{sim}(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\frac{\text{sim}(z_i, z_j)}{\tau}\right)} \right) \tag{10}$$

where sim () is the cosine similarity; $\tau$ is a temperature parameter that controls the penalty given to hard negative samples, which controls the smoothness of the probability distribution (Kumar & Chauhan, 2022; Wang & Liu, 2021); and $\mathbb{1}_{[k \neq 1]} \in 0,1$ is the indicator function, which is set to 1 when $k \neq 1$. The total loss ($L$) for the minibatch can then be calculated as:

$$L = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)] \tag{11}$$

At each training iteration a stochastic gradient descent (SGD) optimizer with linear rate scaling is used to update the base encoder and projection head parameters toward the fastest training loss (Goyal et al., 2017). Fine-tuning

of a CNN trained using SimCLR also showed improved accuracy even with two orders of magnitude fewer hand labeled images provided (Chen et al., 2020).

#### 2.1.7. GeoCLR

Although the method proposed in SimCLR works well to present individual similar and dissimilar images, it does not account for spatial patterns with footprints larger than a single image. To overcome this limitation, Yamada, Prügel-Bennett, et al. (2022) developed "georeference contrastive learning of visual representation" (GeoCLR) to efficiently train CNNs by leveraging georeferenced metadata. Their data set consisted of 86,772 seafloor images collected by an autonomous underwater vehicle (AUV) from a single locality, and each image had an associated depth, northing and easting. In summary, GeoCLR generates a similar image pair ($\tilde{x}_i$ and $\tilde{x}'_j$) from two different images ($x$ and $x'$) that are close together in physical 3D space (Figure 2Ci). Image $x$ possesses a unique geolocation ($g_{east}$, $g_{north}$, $g_{depth}$) and image $x'$ is then selected from a batch of images with a 3D geolocation $\left(g'_{east}, g'_{north}, g'_{depth}\right)$ within a given distance ($r$) of image $x$ provided it meets the following criteria:

$$\sqrt{\left(g'_{east} - g_{east}\right)^2 + \left(g'_{north} - g_{north}\right)^2 + \lambda\left(g'_{depth} - g_{depth}\right)^2} \leq r \tag{12}$$

A scaling factor ($\lambda$) is used to include or exclude images that are close but at different depth. Once image pairs are selected the same augmentations are applied as SimCLR to generate the similar image pair ($x_i$ and $\tilde{x}'_j$) (Yamada, Prügel-Bennett, et al., 2022). Using a semi-supervised framework, the average classification accuracy of GeoCLR was 10.2% higher than an identical CNN trained using SimCLR alone, highlighting the value of utilizing geolocational metadata when using a latent space for feature extraction.

### 2.2. Adapting Spatial Machine Learning for Drill Core Imagery

Here we present a modification of LGA and GeoCLR that involves calculating 2D cylindrical ($x$–y) coordinates, instead of 3D Cartesian coordinates, to guide semantic interpretation of a 2D core image (Figure 2Aii). Typically, images taken during scientific coring operations include: 2D scans of a cut surface of a core section half, 2D images of core sections (either cut or uncut) in a core box, or 3D line scans taken on a 360° core scanner that images the outer surface of the uncut core. As 2D images are more common, and when a 3D image is unwrapped it is also 2D (Figure 2Ai), spatial metadata accompanying a given cropped patch from a core image is a 2D $x$–$y$ coordinate. All these image formats capture visual information about the rocks in the form of a three-channel (RGB) 2D array where the top and bottom of the image have an associated depth down hole. Cores also have different diameters depending on the drill bit used to collect them and this information can be used to calculate the horizontal position of a given patch ($s_i$) as a function of the minimum ($m_i$) and maximum ($M_i$) width of the original image:

$$s_i = m_i + \frac{f_i}{\left(\frac{M_i}{n}\right)}\left(M_i - m_i\right) \tag{13}$$

where $n$ is the number of adjacent patches that fit horizontally into $M_i$ and depends on the image resolution and user defined patch size, and $f_i \in 0, \ldots, \left(\frac{M_i}{n}\right)$ is the horizontal patch index. Similarly, the vertical position ($s_j$) of each patch can be calculated in the same fashion:

$$s_j = m_j + \frac{f_j}{\left(\frac{M_j}{n}\right)}\left(M_j - m_j\right) \tag{14}$$

where $m_j$ and $M_j$ are the minimum and maximum depth of the original image and $f_i \in 0, \ldots, \left(\frac{M_j}{n}\right)$ is the vertical patch index. Our proposed workflow calculates a horizontal 2D spatial location, or polar coordinate, for a given patch and combines this with the depth downhole the patch is from to give an $x$–$y$ coordinate ($s_i$, $s_j$) which is used to determine how close patches are in physical space. Following the methods described above for GeoCLR, our polar coordinate system is used to select $\tilde{x}'$ from a batch of images with a spatial location ($s'_i$, $s'_j$) that meets the following criteria:

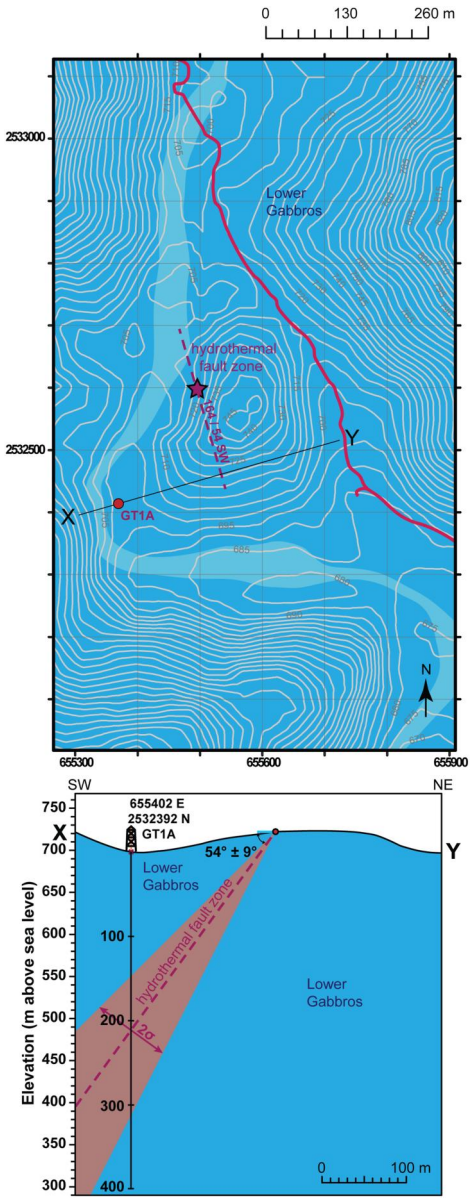


**Figure 3.** Location and cross section of the Hole GT1A drill site, Oman. From Kelemen et al. (2020).

$$\sqrt{(s_i' - s_i)^2 + (s_j' - s_j)^2} \le r \tag{15}$$

Patch pairs ($x_i$ and $\tilde{x}_j'$) then go through the same augmentations used by SimCLR and GeoCLR to extract features from the input data. In contrast, the LGA was modified to use ($s_i$, $s_j$) when quantifying the affinity between $h$ and $y$ using a Gaussian distribution as a kernel, where sigma is set to $d_{\max}$.

### 2.3. Experiment and Workflow

In this study the performance of frameworks that utilize the spatial context of training images (GeoCLR and LGA) are compared to equivalent methods that do not use this context (SimCLR and AE) (Table 1). Additionally, a 4800 (400/class) image subset was used for supervised training of ResNet18 to benchmark against the performance of unsupervised learning results. A summary of the models tested in this study can be seen in Table 1.

#### 2.3.1. Data Set

All images used in this study are of core recovered from Oman Drilling Project (OmanDP) Hole GT1A drilled into gabbroic rocks from the Semail ophiolite (Figure 3), an ancient slab of ocean crust preserved on the Arabian margin (Kelemen et al., 2020). All cores were imaged using a DMT CoreScan3 digital line scanner which rotated them about their cylindrical axis as the DMT incrementally imaged the full length of the core exterior (Grant et al., 2024). Cores were imaged one section at a time, and each section was no longer than 1 m, as this was the maximum length the scanner could fit. Each section had a blue and red crayon line drawn along its length to indicate way up and as a guide for where it was to be cut into an archive (preserved for future reference) and working (for sampling) half. When orientated to its original vertical position, the blue line is to the left of the red. The total depth of Hole GT1A is 403.4 m; cores collected from the upper 254.2 m were drilled with an HQ diamond bit yielding core with a diameter of 63.5 mm (1995 pixels). Below this depth, coring used a narrower PQ bit and cores are 47.8 mm in diameter (1493 pixels) (Kelemen et al., 2020). All images were taken at a 10 pixel/mm resolution and stored as bitmap files.

Core exterior images collected during the OmanDP were an excellent candidate for this study due to the large amount of data accompanying them in the form of VCDs and detailed core logs generated by expert geologists. Therefore, all labeling of training and validation images in this study were cross referenced and groundtruthed to these data, as well as confirmed by the geologists involved in the description of these cores.

#### 2.3.2. Training Image Preparation

Raw bitmap images were prepared for training by: (a) transposing to the correct vertical orientation, (b) cropping any valueless pixel columns from image borders, (c) "rotating" the image horizontally until the blue cutting line was at 100 pixels from the left of the image (Figure 2Ai). Many of the images had been rotated more than 360° during scanning, making the apparent resolution of 10 pixels/mm inaccurate. However, this only duplicates ~20 pixels either side of the vertically rotated raw image. In some cases, images were over-rolled (>540°), which was resolved by cropping them to the correct width of 1995 or 1493 pixels, depending on core diameter. Uneven surfaces appear as visual interference, particularly at either end of a section with angular contacts with the sections above or below it. Spurious reflections are also present where tape was used to hold fractured core together during scanning, or where foam was used as a spacer in some cases where material was too fragmented to scan. Once prepared, all section images were segmented to produce 722,157 100 × 100 pixel (1 $cm^2$) patches that were used

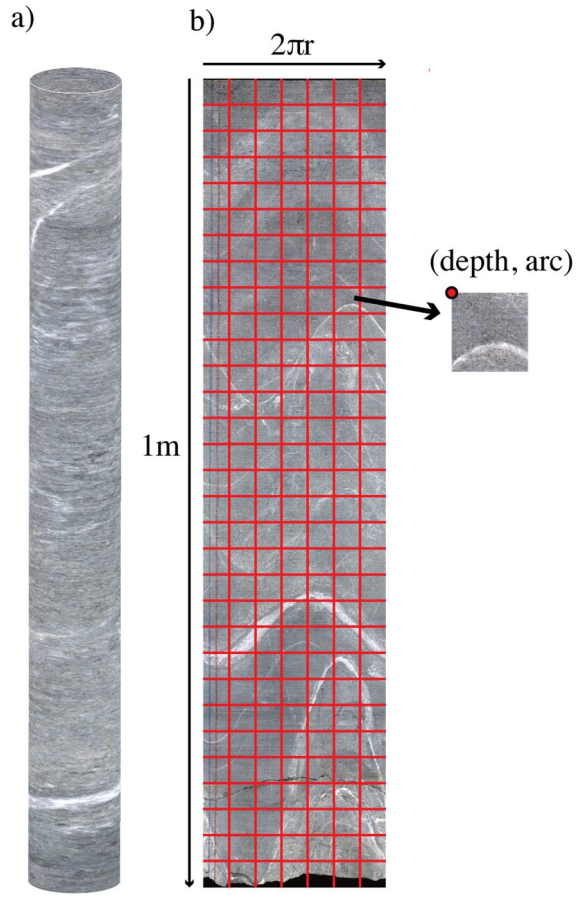


**Figure 4.** (a) Diagram of a 3D scan of a section of core, and (b) the unrolled 2D version of panel (a) with an example of the segmentation style used to generate training patches used in this study (red grid not to scale). The top left corner of each training patch is the location of the patch's depth and arc position on the core surface (right).

to train the machine learning models (Figure 2Aii). Patch size was chosen to be small enough to avoid multiple classes occurring in a single image, but large enough to be labeled by an expert (Figure 4).

### 2.3.3. Self-Supervised Learning Configuration

Configurations for all models are set to those deemed optimal by Yamada et al. (2021) and Yamada, Prügel-Bennett, et al. (2022) during their development of the LGA and GeoCLR methods, except for threshold closeness ($d_{\max}$ and $r$) and number of training cycles. All training patches were expanded to 227 × 227 pixels during feature extraction for the AE and LGA, as this was the size required by the AlexNet-based autoencoder. In contrast, SimCLR and GeoCLR methods re-scale each patch to a resolution of 2 mm/pixel and randomly crop out a 224 × 224 region for use during training (Yamada et al., 2021; Yamada, Prügel-Bennett, et al., 2022). The number of dimensions in latent space ($h$) for the autoencoders is set to 16, whereas for SimCLR and GeoCLR it is set to 128. For all frameworks, the number of images fed into the model at each training iteration (mini-batch) was set to 256 and training ran for 200 iterations (epochs). Patches physically adjacent in all directions to $x_i$ were deemed close enough spatially to assume they will look similar, therefore $d_{\max}$ and $r$ were set to 1.5 cm. Hyperparameters such as learning rate and weight decay for all models were set to the optimal values determined during their development (see reference in Table 1).

### 2.3.4. Geologically Constrained Semi-Supervised Clustering

A total of 12 classes were defined to be representative of the most common rock types and features that occur downhole within Hole GT1A (Kelemen et al., 2020) (Figure 5). All 722,157 image patches were used during self-supervised learning, and two subsets of 100 and 300 per class were expert labeled for validation and training, respectively (Figure 2Aiii). Several classes are not of geological interest so to avoid these features being incorrectly labeled, they were treated as distinct classes. These include spurious noise from tape and foam, as well as crayon lines and dark empty space. Gabbros in Hole GT1A were subdivided based on their color, with light gray, more felsic, patches being termed simply "gabbro." Gabbro with ~1%–5% darker minerals was termed "olivine-bearing gabbro," whereas patches with ~6%–50% dark minerals were referred to as "olivine gabbro," and patches containing ≥50% dark minerals were labeled as "mela-olivine gabbro." Dark minerals in Hole GT1A are primarily a mix of olivine and clinopyroxene and distinguishing between the two in the training images was not always possible. Therefore, all expert labels given to patches were groundtruthed to the lithology and modal abundances recorded for the appropriate interval in the OmanDP VCDs. Other classes considered of interest for alteration logging included: veins composed of white minerals (vein type A), veins that contain a mix of prehnite and chlorite (vein type B), "fracture" and "alteration zone," which were also groundtruthed using the OmanDP vein and alteration logs (Kelemen et al., 2020). Here alteration refers to parts of the core where primary igneous minerals have been replaced by secondary phases due to hydrothermal alteration and/or deformation, which occurs in Hole GT1A mostly as patches, halos and densely spaced vein networks. Within Hole GT1A there is variability in the dominant secondary minerals present in an alteration zone (Greenberger et al., 2021; Kelemen et al., 2020), however, all were placed in a single class to capture zones of focused alteration. In many cases, patches labeled as alteration zone could be confused as a type of vein if the annotator only looks at the 1 × 1 cm patch. However, when the spatial context of a patch revealed that it sits within an altered interval, and is not part of a single linear vein, it was labeled as "alteration zone."

For all experimental configurations a class-balanced approach was used where an equal number of representative expert annotations per class ($M/n$) were manually generated. A class-balanced approach can be time consuming when compared to other selection methods (Yamada, Prügel-Bennett, et al., 2022). However, it ensures all labels provided are representative of the high intra-class variation at the cm-scale in the rocks. Each

10.1029/2023EA003220

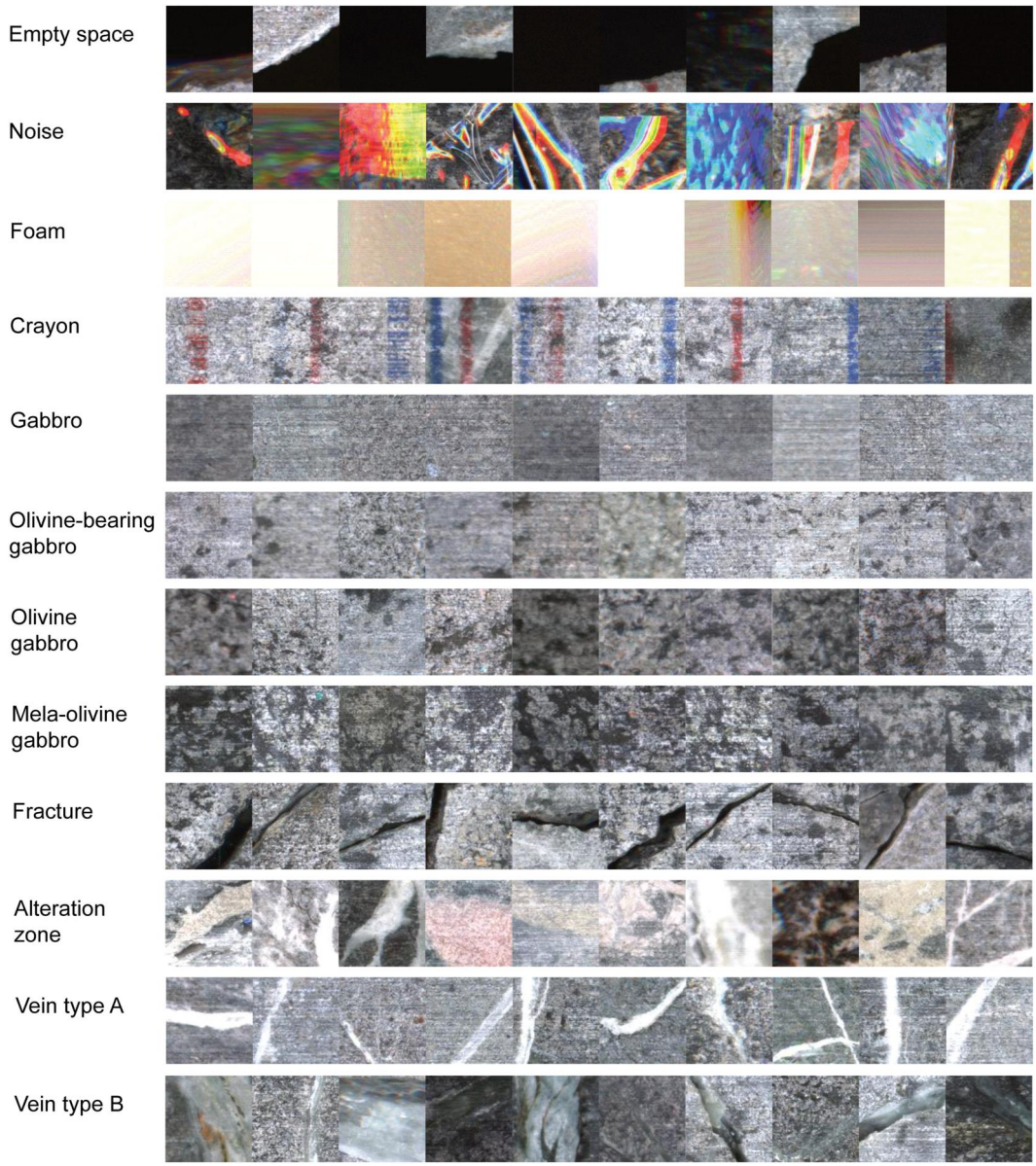


**Figure 5.** Example images of expert defined classes that each model was trained to identify. These were chosen to represent the most common rock types in Hole GT1A as well as highlight areas of intensified hydrothermal alteration and fracturing. A total of 400 images were expert labeled per class, with 300 used for training and 100 for validation.

model was trained multiple times, varying $M/n$ to find its optimal value. Labeling 100 images for each of the 12 classes in this study took ∼16–24 hr. Therefore, a maximum of $M/n = 300$ was chosen because the time taken to manually label more images would be inefficient in the context of real-time core analysis during a geological coring project. For a given $M/n$, self-supervised training produced a latent representation of the data set before a support vector machine with a radial basis function as a kernel (R-SVM) was used to classify the data based on the expert annotated subset (Figure 2Aiv). The outcome of this classification is all images are assigned a computer-generated pseudolabel, which were then compared to the expert labeled validation subset to quantify the accuracy of each model. The best performing configuration for each model was then fine-tuned with the pseudolabels generated by the R-SVM, and in all cases ResNet18 was used as the fine-tuning classifier (Figure 2Av).

**Table 2**
*Results of Each Model Trained Using the Semi-Supervised Workflow Presented in This Paper, As Well As Supervised Learning Results for ResNet18*

| | | *M*/*n* | | | | | |
|---|---|---|---|---|---|---|---|
| CNN framework | Training type | 3 | 10 | 30 | 100 | 200 | 300 |
| AE | Semi-supervised | 0.08 ± 0.04 | 0.08 ± 0.04 | 0.07 ± 0.04 | 0.09 ± 0.04 | 0.10 ± 0.05 | 0.10 ± 0.03 |
| LGA | Semi-supervised | 0.33 ± 0.25 | 0.38 ± 0.27 | 0.48 ± 0.25 | 0.59 ± 0.21 | 0.60 ± 0.22 | 0.62 ± 0.21 |
| SimCLR | Semi-supervised | 0.34 ± 0.23 | 0.49 ± 0.20 | 0.60 ± 0.19 | 0.69 ± 0.16 | 0.73 ± 0.15 | 0.74 ± 0.14 |
| GeoCLR | Semi-supervised | **0.40 ± 0.18** | **0.60 ± 0.14** | **0.74 ± 0.13** | **0.84 ± 0.07** | **0.86 ± 0.07** | **0.86 ± 0.07** |
| ResNet18 | Supervised | 0.33 ± 0.34 | 0.34 ± 0.35 | 0.52 ± 0.31 | 0.67 ± 0.21 | 0.82 ± 0.13 | 0.84 ± 0.11 |

*Note.* All models were trained using an increasing number of training images per class (*M*/*n*) and all results are $f_1$ scores (macro average) ± 1SD. The best performing model for each training configuration is shown in bold.

### 2.3.5. Supervised Training Configuration

Supervised learning methods use labeled data that have corresponding target labels or outputs, whereas unsupervised learning networks extract the underlying structure of the data with no target output. Unsupervised approaches are used in this study to generate a latent space before *M*/*n* expert labeled images are provided for the automatic assignment of computer-generated pseudolabels to the entire data set, which then allow for fine tuning. Fine tuning of a neural network takes the initial pre-trained network as a starting point before adjusting its parameters by re-training using a labeled subset of the data set in a supervised fashion. All semi-supervised frameworks trained with *M*/*n* = 100 and *M*/*n* = 300 were fine-tuned by feeding the entire pseudolabeled data set into ResNet18 with a minibatch size of 128, learning rate and weight decay of $1 \times 10^{-5}$ and Adam optimizer (Kingma & Ba, 2014) (Figure 2Aiv–v). Models trained with other values of M/*n* were not fine-tuned as it would have been computationally burdensome.

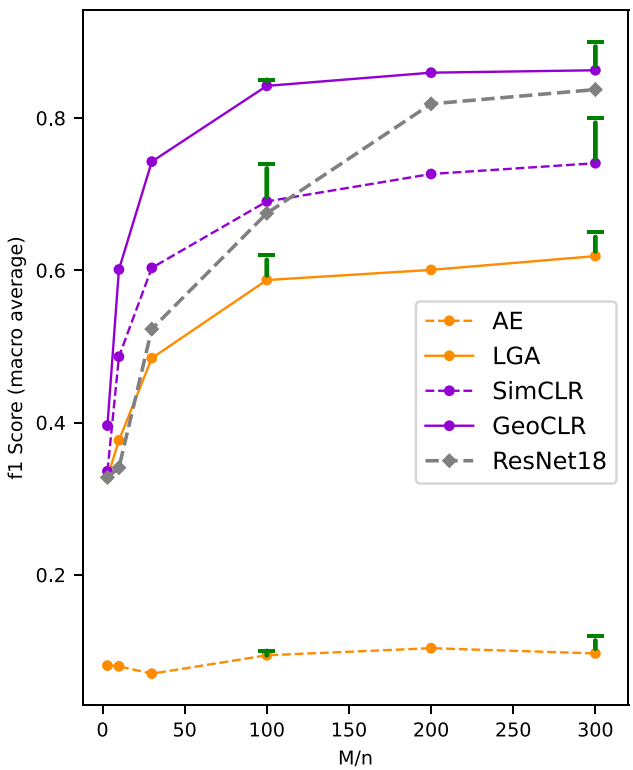


**Figure 6.** $f_1$ (macro average) scores of all models when 3, 10, 30, 100, 200, and 300 expert labeled images per class (*M*/*n*) were used for training. Results of the contrastive learning methods are shown in purple, the results of the autoencoder methods are in orange, and the results of the supervised model are in gray. Solid lines indicate models that make use of spatial metadata and dashed lines are those that do not, whereas circles represent the unsupervised model results and diamonds supervised model results. Green lines indicate the performance increase gained by fine tuning.

To quantify the improvement of using unsupervised feature extraction prior to fine tuning over a simple supervised approach that would require the same expert labeling effort, the expert labeled training (300/class) and validation (100/class) images were also used for supervised training of ResNet18. ResNet18 was pre-trained using ImageNet (Deng et al., 2009), and its hyperparameters were set to the same as those used during fine-tuning of the unsupervised frameworks. Training ran for 200 epochs with a batch size of 128 and the last layer of the network was set to the number of classes in the ImageNet database (1000). This is because the last layer of the pre-trained ResNet18 model used for fine-tuning is also 1000, due to the number of classes in the ImageNet database, which matches the approach Yamada et al. (2021) and Yamada, Prügel-Bennett, et al. (2022) took when fine tuning CNNs trained using their LGA and GeoCLR methods.

## 2.4. Validation

When quantifying the performance of machine learning algorithms there are a number of commonly used performance metrics, such as accuracy, precision and recall. Previous attempts to use machine learning to classify core images have primarily reported model performance using only accuracy. However, when the proportions of each class within the training data set are imbalanced accuracy can be inflated in cases where the model does particularly well at classifying the most abundant classes. In the case of using unsupervised learning approaches, the relative abundance of each expected class in the data set is not known. Therefore, in this study we use the $f_1$ score for each class to

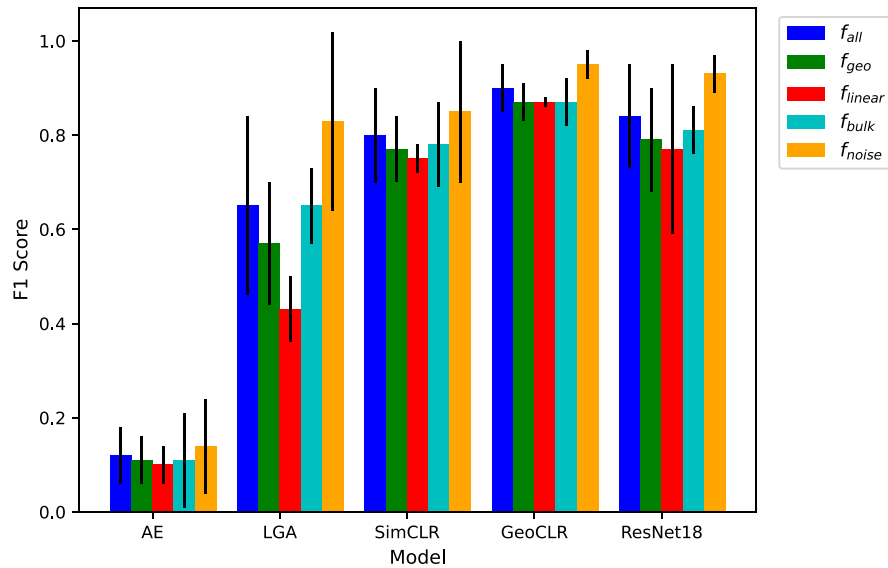


**Figure 7.** Class averaged $f_1$ scores for each model; $f_{all}$ is the macro average across all classes; $f_{geo}$ is the average score for geological classes only (gabbro, olive bearing gabbro, olivine gabbro, mela olivine gabbro, alteration zone, fracture, vein type A, vein type B); $f_{linear}$ is the average score for linear classes (fracture, vein type A, and vein type B); $f_{bulk}$ is the average score for all bulk rock classes (gabbro, olivine bearing gabbro, olivine gabbro, and mela olivine gabbro). $f_{noise}$ is the average score for all non-geological classes. All errors are shown as 1SD.

quantify model performance as it accounts for both the model's ability to correctly identify positive instances (precision) and to capture all positive instances (recall):

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$f_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{18}$$

where true positive (TP), false positive (FP), true negative (TN), and false negative (FN) results were generated by comparing the machine learning classifications given to the 1200 validation images labeled by domain experts. The overall performance of each model is then presented in this paper as the class-averaged $f_1$ score (macro average) where $f_{1i}$ is the $f_1$ score of class $i$ and $n$ is the total number of classes identified in the data set:

$$f_{1(\text{macro average})} = \frac{\sum_{i=1}^{n} f_{1i}}{n} \tag{19}$$

**Table 3**
*Fine Tuning Results for Each Model Given as* $f_1$ *Scores (Macro Average) ± 1SD*

| *M/n* | 100 | 300 |
|---|---|---|
| AE | 0.09 ± 0.07 | 0.12 ± 0.06 |
| LGA | 0.62 ± 0.20 | 0.65 ± 0.19 |
| SimCLR | 0.74 ± 0.13 | 0.80 ± 0.10 |
| GeoCLR | 0.85 ± 0.08 | **0.90 ± 0.05** |

*Note.* The best performance was acheived by GeoCLR when 300 expert labeled images per class (*M/n*) were provided for fine tuning (bold).

## 3. Results

### 3.1. Training Evaluation

In all cases GeoCLR showed best performance, and the $f_1$ scores (macro average) ± 1$\sigma$ for all model configurations can be seen in Table 2. At all values of *M/n*, the AE performed the worst, demonstrating that, without incorporation of spatial metadata, auto-encoders are not suitable for classification of core images. Excluding the AE, increasing *M/n* improved classification for all other models. The AE only showed minor improvement up to *M/n* = 200 before

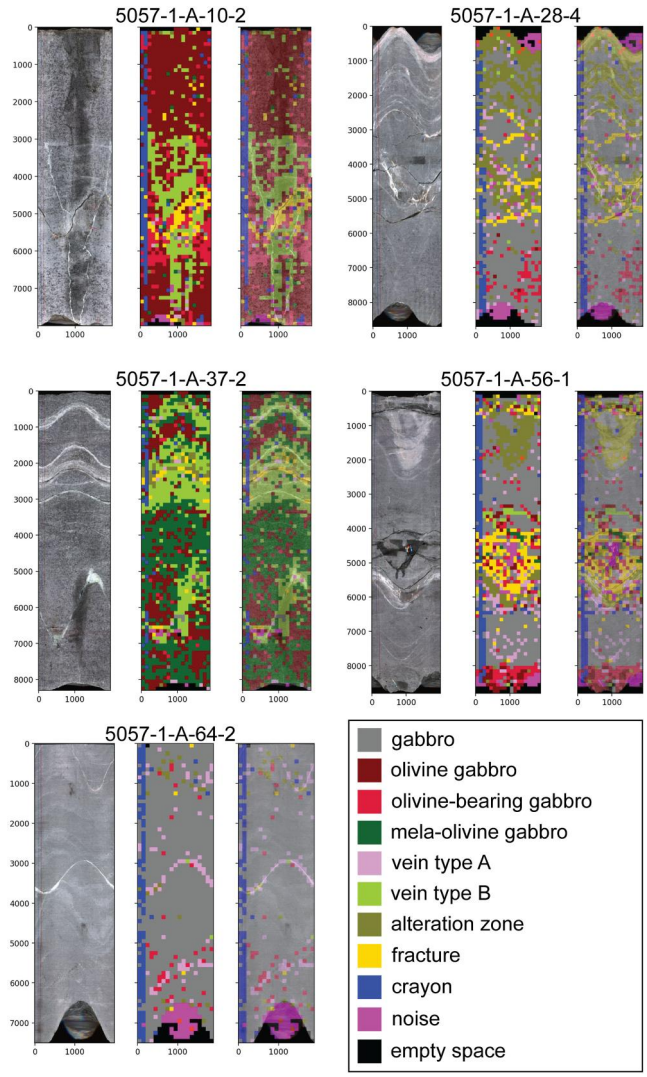


**Figure 8.** Visual comparison of original Hole GT1A core-section images to the classifications given to patches taken from each section. For each section above, the left image is the original 360° DMT image, the middle image is constructed using the patches from the original section where color corresponds to the class label generated using GeoCLR, the right image is the classified patch image overlaid above the original DMT image.

accuracy began to decrease (Figure 6). For the remaining models, sigmoidal growth is seen where most of the improvement in accuracy occurs at the lower end between $M/n = 3$ and $M/n = 100$ (Figure 6). Both the LGA and SimCLR show relatively large increases in accuracy at $M/n > 100$ compared to GeoCLR, suggesting they would have further improved with $M/n > 300$. At no point does the LGA outperform contrastive learning or supervised methods, however it consistently outperforms the AE with a maximum of 52% improvement. This indicates that introduction of spatial metadata when training auto-encoders drastically improves performance.

Peak performance of GeoCLR is achieved with $M/n = \sim 100$, as performance only increases by ~2% before plateauing with increased $M/n$. Furthermore, with only $M/n = 30$, GeoCLR was able to outperform SimCLR and ResNet18 trained with an order of magnitude more annotations by 7% and 5%, respectively. Both contrastive learning methods outperform ResNet18 at lower values of $M/n$, but at $M/n > 100$ ResNet18 is more accurate than SimCLR and begins to achieve comparable performance to GeoCLR with increasing $M/n$. However, GeoCLR requires less domain expert effort to produce higher accuracy image classification than supervised (ResNet18) and black box (SimCLR) models.

### 3.2. Class Identification

At lower values of $M/n$, the LGA outperforms the contrastive learning frameworks in correctly identifying non-geological classes, such as noise, foam and empty space. In contrast, both SimCLR and GeoCLR outperform the LGA in correctly distinguishing geological classes with fewer expert-generated labels ($M/n < 100$). SimCLR correctly identifies foam and empty space in almost all cases, however, it fails to reliably distinguish crayon from the rock on which it was drawn. Regardless of increasing $M/n$, the LGA poorly distinguishes between classes containing single linear features, such as fractures and veins. The gabbroic rock classes share a lot of visual similarity, given they are defined by subdivisions of a property that actually spans a spectrum of values (dark mineral abundance). This is particularly evident at the extreme ends of the color index used to define them in this study. These shared characteristics cause both the LGA and SimCLR to mis-label 5%–8% olivine-bearing gabbro as olivine gabbro, whereas GeoCLR only confuses 4% and 6% of olivine-bearing gabbro for gabbro and olivine gabbro, respectively. For the more mafic-rich (higher proportion of dark mineral) classes, all models mis-label ≥10% of mela-olivine gabbro as olivine gabbro.

### 3.3. Fine Tuning

Figure 7 compares fine-tuned networks pre-trained using the AE, LGA, SimCLR and GeoCLR frameworks to ResNet18. This comparison serves as an indicator of how well the semi-supervised methods outlined in this paper compare to commonly used supervised image classification techniques (K. He et al., 2016; Krizhevsky et al., 2012). Specific $f_1$ scores were generated by averaging the scores of related groups of classes to highlight how well models classify geological ($f_{geo}$), linear ($f_{linear}$), bulk-rock ($f_{geo}$) and noisy ($f_{noise}$) classes (Figure 7). All models except the AE are effective at filtering out noisy classes not of geological interest, whereas linear classes are those most often mis-classified. Yamada et al. (2021) demonstrated that their LGA improved the classification accuracy of linear classes to 53.7%, as they had a characteristic spatial distribution. In this study linear features were the least well classified, even with spatial metadata, as the LGA gave an $f_{linear}$ of 0.43 ± 0.07. Like the LGA, ResNet18, and SimCLR gave a relatively low $f_{linear}$ when compared to $f_{geo}$ and $f_{bulk}$, but are still more accurate than the LGA as they all have $f_{linear} > 0.75$. Unlike all other models, GeoCLR shows almost no variation between its ability to classify linear, bulk and geological features, and all have $f_1$ scores of 0.87 ± 0.01–0.05. This consistent

**Table 4**
*Pearson's Coefficient (r) and p Values Calculated by Comparing the VCD-Based and AI-Based Alteration Log Data*

| Data set | | Window size (cm) 1 | 100 | 400 |
|---|---|---|---|---|
| Alteration % | $r$ | 0.50 | 0.71 | 0.72 |
| | $p$ Value | 0.00 | $6.05 \times 10^{-61}$ | $3.61 \times 10^{-17}$ |
| | $n$ | 39,784 | 392 | 99 |
| Background % | $r$ | 0.48 | 0.68 | 0.65 |
| | $p$ Value | 0.00 | $1.79 \times 10^{-54}$ | $2.74 \times 10^{-13}$ |
| | $n$ | 39,784 | 392 | 99 |

*Note.* Analysis was performed for three different depth resolutions: 1 cm; and for running averages calculated using 1 and 4 m window sizes. $n$ indicates the number of data points compared for each iteration.

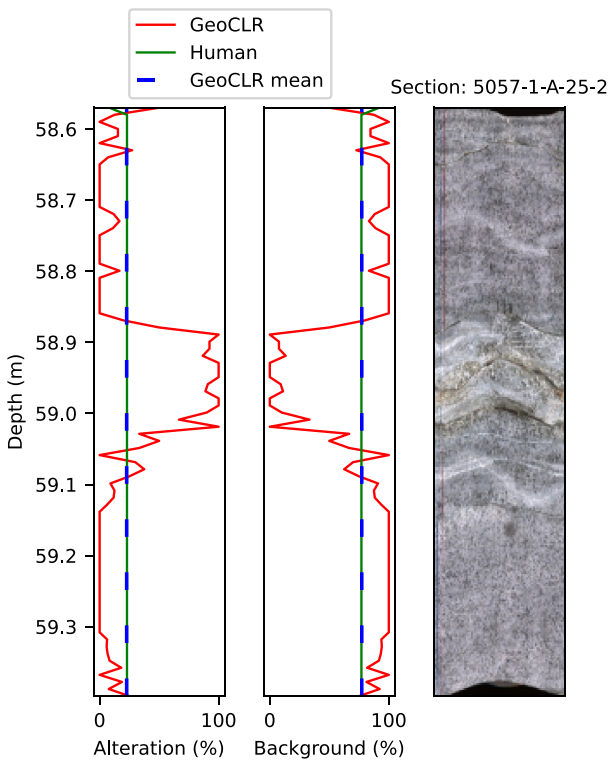


**Figure 9.** Proportions of alteration and relatively fresh background rock through an example Hole GT1A core-section calculated using classifications generated using GeoCLR (red line) are compared to the equivalent data generated by alteration petrologists during the Oman Drilling Project (green line). The mean GeoCLR values through this interval (blue dashed lines) show excellent agreement with the visual core description-based estimates made by human experts. The vertical and horizontal scales of the core section image are equal.

accuracy across class types, combined with its high $f_{all} = 0.90 \pm 0.01$ and low error confirm that GeoCLR outperforms all other models evaluated in this study (Table 3). The classifications made by a fine-tuned model trained using our modified GeoCLR framework can therefore reliably be used to visualize and quantify geological features in a core.

## 4. Automated Alteration Logging

Plotting the relative abundance of fresh and altered rock downhole highlights regions of focused hydrothermal alteration within the ocean crust (Alt et al., 2010; Coggon et al., 2022; Kelemen et al., 2020; Teagle et al., 2023). During scientific drill core description alteration petrologists gather this data using visual estimations of alteration extent. The scale at which estimations are made often varies between expeditions and the quantification has an element of subjectiveness. Here we present a novel and automated approach to evaluating the spatial variations in the alteration extent downhole using the classifications generated by GeoCLR as a demonstration that AI-based approaches can standardize time-intensive geological tasks. Validation of our AI-based method is done by comparing it to an equivalent data set generated by experts during the OmanDP.

The expert-generated alteration data for Hole GT1A includes visual estimations of the average proportion of alteration features (halos, patches, and deformation), as well as relatively fresh background rock within continuous downhole intervals. The depth and length of these intervals were defined by distinct changes in the nature/extent of alteration. To allow comparison with the cm-scale AI-based data through Hole GT1A, we assume that the proportions of alteration features in a given interval are representative of each centimeter of core in that interval. This assumption allowed a continuous downhole visual core description-based (VCD-based) estimate of the extent of alteration and background rock to be calculated by summing the proportions of all alteration types in an interval. A comparable depth-resolution data set was then generated from the AI-based core logging data by calculating the percentage of patches labeled as 'alteration zone' by GeoCLR at each cm downhole (Figure 8). Similarly, the proportion of images labeled as a class of gabbroic rock was used to infer the amount of relatively fresh background rock in each cm downhole. GeoCLR classified images of "alteration zone" with an $f_1 = 0.9$, although 3% and 5% of the validation data set were mis-labeled as foam and vein type A, respectively. Foam was inserted into regions too altered and fractured to be scanned on the DMT core scanner, and veins occur in conjunction with high levels of alteration in the core. Therefore, the presence of these classes are indicative of alteration, so their mis-classification is not expected to significantly bias a plot of alteration extent downhole.

The 1 cm depth resolution of the AI-generated data reveals high frequency shifts in alteration and background extent, whereas the lower-resolution VCD-based data displays sharp step-wise shifts between alteration intervals, which results in only a moderately positive correlation between the data sets (Table 4). Close inspection of a given 1 m section reveals that the AI-based data is capable of picking out small localized spatial variations in alteration that would be impractical for an expert to log (Figure 9). However, the mean AI-based estimates of alteration extent and proportion of background rock through a given section show good agreement with the VCD-based estimates through the same interval (Figure 9), further confirming that the AI-based approach is capable of identifying and quantifying geologically significant features identified by the experts—albeit at higher-

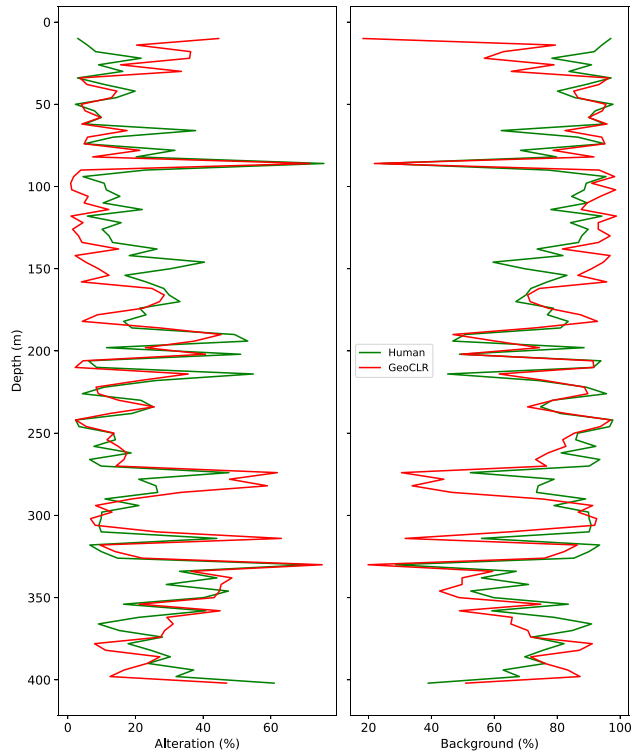


**Figure 10.** Running average (window size = 4 m) for the abundance (%) of altered rock (left) and fresher protolith (right) downhole within Hole GT1A for both VCD-based estimations (green line) and AI-based estimations made using GeoCLR classifications (red line). As the VCD-based data sums to 100% of the core surface, the AI-based data was normalized to also sum to 100%.

resolution. To better visualize the broad variations in alteration extent within Hole GT1A, downhole-running averages for every 1 m (length of a core section) and 4 m (length of a full core) were also calculated to smooth both the AI and VCD-based data (Figure 10). On average the AI-based data at a given depth is 1%–2% higher than the VCD-based data, however, the Pearson's Coefficient for the running averages of 1 and 4 m shows statistically significant positive correlations between the AI and VCD-based data sets (Table 4).

Overall, the large-scale variations in the smoothed VCD-based data are also captured by the AI-based data, and only two major discrepancies are observed at 146 and 365 m (Figure 10). The first of these discrepancies occurs where a highly fractured interval has been visually identified as 40.5% altered, whereas GeoCLR defined most of the interval as 'fracture'. The second occurs where GeoCLR underestimates the amount of background rock by classifying patches of gabbro at this depth as 'crayon', highlighting the importance of minimizing the markings made to the core surface prior to imaging.

The comparable performance of our AI-based approach using images alone to traditionally labor-intensive on-site core description demonstrates that AI methods have the potential to revolutionize current practices in the field. Specifically, rather than dedicating time to visually quantifying features experts could dedicate more time to discrete sample analysis or carrying out more detailed analysis of important intervals. Also, experts could dedicate time to labeling training images on-site while core is on display, as this would further ground classifications to the actual recovered material. One limitation, however, is that cores are imaged one section at a time, so model training could not commence until drilling operations at a site are complete. Also, depending on the amount of recovered material, training time may take too long to be done on-site forcing AI-based approaches to be postponed until post-expedition. Regardless, it is clear that modifying on-site workflow with approaches such as that outlined in this work in mind would save significant amounts of time during a given coring campaign.

## 5. Summary

This study presents a novel semi-supervised machine learning approach for the analysis and classification of geological images that utilizes spatial metadata for improved machine learning accuracy that can be implemented into existing CNN architectures. This method can be applied to any Earth or space image data sets that have accompanying spatial metadata, and implementing this workflow into several state-of-the-art machine learning frameworks has demonstrated that:

1. When only 30 labeled images per class are used for training, incorporating spatial metadata improves the classification accuracy of unsupervised auto-encoder and contrastive learning frameworks by 30% and 11%, respectively. Increasing this to >100 images per class further improves performance over non-spatially guided auto-encoders and contrastive learning by 50.7% and 13.3%, respectively.
2. Of the unsupervised learning models tested, spatially guided contrastive learning (GeoCLR) had the best classification accuracy, regardless of the number of expert-generated annotated images used for training. GeoCLR outperforms both non-spatially guided and supervised methods with an order of magnitude fewer expert-generated annotations and reaches maximum accuracy with ∼100 annotated images per class (1200 images).
3. Fine tuning of unsupervised models improves classification accuracy by an average of 2.25%, and GeoCLR trained with 300 expert-generated annotations per class showed the best performance in this study with a classification accuracy of 90 ± 0.05% after fine tuning. Classes containing linear features, such as veins and fractures, with spatial context extending beyond the frame of a single patch are the least well classified class type for all models except GeoCLR, which labels all types of class with comparable accuracy.
4. Classifications generated using methods described here allow for the automated generation of downhole data sets traditionally created by experts over the course of days to weeks. Comparing downhole estimates of the amount of altered and relatively fresh rock based on both GeoCLR classifications and visual expert estimations indicate a statistically significant positive relationship (Pearson's coefficient = 0.7). Therefore, our automated method provides a reliable and efficient means of analyzing geological images at higher resolutions than would be feasible using current manual approaches.

## Data Availability Statement

All images and geological log data used in this study are available from the Oman Drilling Project website (publications.iodp.org/other/Oman/OmanDP.html).

**Acknowledgments**
Authors would like to thank Dr Jude Coggon (University of Southampton) for her help compiling all OmanDP data used in the study, as well as Dr. Michelle Harris (University of Plymouth) for her input on data presentation and insight into the expert generated data sets created by herself and others during the OmanDP. Finally, we would like to thank Prof. Timothy Henstock (University of Southampton) for his input during the write-up. RMC was funded by a Royal Society University Research Fellowship (URF\R1\180320) and LG by a Royal Society award (RGF\EA\181072) to RMC. DAHT was funded by and NERC-NSF grant (NSFGEO-NERC: NE/W007517/1 "Data mining the deep"). This research used samples and data provided by the Oman Drilling Project. The Oman Drilling Project (OmanDP) has been possible through co-mingled funds from the International Continental Scientific Drilling Project (ICDP; Kelemen, Matter, Teagle Lead PIs), the Sloan Foundation—Deep Carbon Observatory (Grant 2014-3-01, Kelemen PI), the National Science Foundation (NSF-EAR-1516300, Kelemen lead PI), NASA—Astrobiology Institute (NNA15BB02A, Templeton PI), the German Research Foundation (DFG: KO 1723/21-1, Koepke PI), the Japanese Society for the Promotion of Science (JSPS no:16H06347, Michibayashi PI; and KAKENHI 16H02742, Takazawa PI), the European Research Council (Adv: no.669972; Jamveit PI), the Swiss National Science Foundation (SNF: 20FI21163073, Früh-Green PI), JAMSTEC, the TAMU-JR Science Operator, and contributions from the Sultanate of Oman Ministry of Regional Municipalities and Water Resources, the Oman Public Authority of Mining, Sultan Qaboos University, CNRS-Univ. Montpellier, Columbia University of New York, and the University of Southampton.

## References

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., & Cremers, D. (2018). Clustering with deep learning: Taxonomy and new methods. *arXiv Preprint arXiv:1801.07648*.

Al-Mudhafar, W. J. (2017). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *Journal of Petroleum Exploration and Production Technology*, *7*(4), 1023–1033. https://doi.org/10.1007/s13202-017-0360-0

Alt, J. C., Laverne, C., Coggon, R. M., Teagle, D. A., Banerjee, N. R., Morgan, S., et al. (2010). Subsurface structure of a submarine hydrothermal system in ocean crust formed at the east pacific rise, odp/iodp site 1256. *Geochemistry, Geophysics, Geosystems*, *11*(10). https://doi.org/10.1029/2010GC003144

Alzubaidi, F., Mostaghimi, P., Swietojanski, P., Clark, S. R., & Armstrong, R. T. (2021). Automated lithology classification from drill core images using convolutional neural networks. *Journal of Petroleum Science and Engineering*, *197*, 107933. https://doi.org/10.1016/j.petrol.2020.107933

Camps-Valls, G., Marsheva, T. V. B., & Zhou, D. (2007). Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(10), 3044–3054. https://doi.org/10.1109/TGRS.2007.895416

Chai, H., Li, N., Xiao, C., Liu, X., Li, D., Wang, C., & Wu, D. (2009). Automatic discrimination of sedimentary facies and lithologies in reef-bank reservoirs using borehole image logs. *Applied Geophysics*, *6*(1), 17–29. https://doi.org/10.1007/s11770-009-0011-4

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). https://doi.org/10.48550/arXiv.2002.05709

Coggon, R. M., Sylvan, J. B., Teagle, D. A., Reece, J., Chrsteson, G. L., & Estes, E. R. (2022). *Expedition 390 preliminary report: South Atlantic Transect 1*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.pr.390.2022

Coggon, R. M., Teagle, D. A. H., Sylvan, J. B., Reece, J., Estes, E. R., Williams, T. J., et al. (2024). *Proceedings of the International Ocean Discovery Program, 390/393*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.390393.2024

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Finn, P. G., Udy, N. S., Baltais, S. J., Price, K., & Coles, L. (2010). Assessing the quality of seagrass data collected by community volunteers in Moreton bay Marine Park, Australia. *Environmental Conservation*, *37*(1), 83–89. https://doi.org/10.1017/S0376892910000251

Fu, D., Su, C., Wang, W., & Yuan, R. (2022). Deep learning based lithology classification of drill core images. *PLoS One*, *17*(7), e0270826. https://doi.org/10.1371/journal.pone.0270826

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., & He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv Preprint arXiv:1706.02677*. https://doi.org/10.48550/arXiv.1706.02677

Grant, L. J. C., Evans, A. D., Coggon, R. M., Estep, J. D., McIntyre, A., Slagle, A., et al. (2024). Data report: High-resolution digital imaging of whole-round hard rocks collected during IODP Expeditions 390C, 395E, 390, and 393, South Atlantic Transect, using a DMT CoreScan3. In R. M. Coggon, D. A. H. Teagle, J. B. Sylvan, J. Reece, E. R. Estes, T. J. Williams, et al. (Eds.), *Proceedings of the International Ocean Discovery Program, 390/393*. International Ocean Discovery Program. (In Press). https://doi.org/10.14379/iodp.proc.390393.209.2024

Greenberger, R. N., Harris, M., Ehlmann, B. L., Crotteau, M. A., Kelemen, P. B., Manning, C. E., et al. (2021). Hydrothermal alteration of the ocean crust and patterns in mineralization with depth as measured by micro-imaging infrared spectroscopy. *Journal of Geophysical Research: Solid Earth*, *126*(8), e2021JB021976. https://doi.org/10.1029/2021jb021976

He, J., La Croix, A. D., Wang, J., Ding, W., & Underschultz, J. (2019). Using neural networks and the Markov chain approach for facies analysis and prediction from well logs in the precipice sandstone and evergreen formation, Surat basin, Australia. *Marine and Petroleum Geology*, *101*, 410–427. https://doi.org/10.1016/j.marpetgeo.2018.12.022

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hill, E. J., Pearce, M. A., & Stromberg, J. M. (2021). Improving automated geological logging of drill holes by incorporating multiscale spatial methods. *Mathematical Geosciences*, *53*(1), 21–53. https://doi.org/10.1007/s11004-020-09859-0

Hill, E. J., Robertson, J., & Uvarova, Y. (2015). Multiscale hierarchical domaining and compression of drill hole data. *Computers & Geosciences*, *79*, 47–57. https://doi.org/10.1016/j.cageo.2015.03.005

Jarrard, R. D., Abrams, L. J., Pockalny, R., Larson, R. L., & Hirono, T. (2003). Physical properties of upper oceanic crust: Ocean drilling program hole 801c and the waning of hydrothermal circulation. *Journal of Geophysical Research*, *108*(B4). https://doi.org/10.1029/2001JB001727

Kelemen, P. B., Matter, J. M., Teagle, D. A., Coggon, J. A., & the Oman Drilling Project Science Team (2020). Oman drilling project: Scientific drilling in the Samail ophiolite, sultanate of Oman. In *Proceedings of the Oman drilling Project*. International Ocean Discovery Program. https://doi.org/10.14379/OmanDP.proc.2020

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*(6), 84–90. https://doi.org/10.1145/3065386

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, *26*(2), 195–197. https://doi.org/10.1038/nbt1386

Kumar, P., & Chauhan, S. (2022). Study on temperature ($\tau$) variation for simclr-based activity recognition. *Signal, Image and Video Processing*, *16*(6), 1667–1672. https://doi.org/10.1007/s11760-021-02122-x

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks* (Vol. 10, p. 3361).

Ma, Y. Z. (2011). Lithofacies clustering using principal component analysis and neural network: Applications to wireline logs. *Mathematical Geosciences*, *43*(4), 401–419. https://doi.org/10.1007/s11004-011-9335-8

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, *6*, 39501–39514. https://doi.org/10.1109/ACCESS.2018.2855437

Olmstead, M. A., Wample, R., Greene, S., & Tarara, J. (2004). Nondestructive measurement of vegetative cover using digital image analysis. *HortScience*, *39*(1), 55–59. https://doi.org/10.21273/HORTSCI.39.1.55

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms* (Vol. 55). Spartan Books.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Teagle, D., Reece, J., Coggon, R., Sylvan, J. B., Christeson, G. L., Williams, T. J., & Estes, E. R. (2023). International ocean discovery program expedition 393 preliminary report: South Atlantic Transect 2. In *International Ocean Discovery Program Expedition Preliminary Report* (Vol. 393). https://doi.org/10.14379/iodp.pr.393.2023

Thomas, A., Rider, M., Curtis, A., & MacArthur, A. (2011). Automated lithology extraction from core photographs. *First Break*, *29*(6). https://doi.org/10.3997/1365-2397.29.6.51281

Tominaga, M., Teagle, D. A., Alt, J. C., & Umino, S. (2009). Determination of the volcanostratigraphy of oceanic crust formed at superfast spreading ridge: Electrofacies analyses of odp/iodp hole 1256d. *Geochemistry, Geophysics, Geosystems*, *10*(1). https://doi.org/10.1029/2008GC002143

Tominaga, M., & Umino, S. (2010). Lava deposition history in odp hole 1256d: Insights from log-based volcanostratigraphy. *Geochemistry, Geophysics, Geosystems*, *11*(5). https://doi.org/10.1029/2009GC002933

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Van Etten, A., Lindenbaum, D., & Bacastow, T. M. (2018). Spacenet: A remote sensing dataset and challenge series. *arXiv Preprint arXiv:1807.01232*. https://doi.org/10.48550/arXiv.1807.01232

Wang, F., & Liu, H. (2021). Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2495–2504). https://doi.org/10.48550/arXiv.2012.09740

Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning* (pp. 478–487). https://doi.org/10.48550/arXiv.1511.06335

Yamada, T., Massot-Campos, M., Prügel-Bennett, A., Pizarro, O., Williams, S. B., & Thornton, B. (2022). Guiding labelling effort for efficient learning with georeferenced images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 593–607. https://doi.org/10.1109/TPAMI.2021.3140060

Yamada, T., Prügel-Bennett, A., & Thornton, B. (2021). Learning features from georeferenced seafloor imagery with location guided autoencoders. *Journal of Field Robotics*, *38*(1), 52–67. https://doi.org/10.1002/rob.21961

Yamada, T., Prügel-Bennett, A., Williams, S. B., Pizarro, O., & Thornton, B. (2022). Geoclr: Georeference contrastive learning for efficient seafloor image interpretation. *arXiv Preprint arXiv:2108.06421*. https://doi.org/10.55417/fr.2022037

Zhang, P., Sun, J., Jiang, Y., & Gao, J. (2017). Deep learning method for lithology identification from borehole images. In *79th EAGE Conference and Exhibition* (Vol. 2017, pp. 1–5). European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.201700945