CoreScore: a machine learning approach to assess legacy core condition

Mark Fellgett*, Alex Hall, Simon Harris, Magret Damaschke and Andrew Kingdon

British Geological Survey, Nottingham NG12 5GG, UK

MF, 0000-0001-7969-0021; SH, 0000-0002-7408-7255; MD, 0000-0002-3595-4950; AK, 0000-0003-4979-588X

*Correspondence: markf@bgs.ac.uk

Abstract: Today's geoscience challenges often require repurposing of data and samples from legacy boreholes. Collection of new deep core is expensive; maximizing this investment is vital. However, the condition of legacy cores varies due to factors including recovery, sampling, lithology, and storage.

Rock Quality Designation analysis is often undertaken on new core but this only provides a snapshot of core condition and will not be indicative of subsequent condition. Poor core condition can make destructive analytical techniques impossible and also impacts non-destructive techniques including core scanning.

Since 2011, BGS have systematically collected 125 000 core images. This study investigates if core condition of this archive can be assessed using automated analysis by machine learning. A neural network-based approach was used to segment these images. By differentiating imaged core from their background, properties such as number of fragments and total rock area were determined and used to assess core condition. Analysis of outputs demonstrates that with minimal input data, core condition can be rapidly assessed. This allows users to better understand and visualize core. This can be used to qualitatively assess non-destructive data, improve success of destructive sampling through targeted sampling and reduce the time and effort spent interacting with physical material.

Supplementary material: The code for CoreScore is available at https://github.com/BritishGeologicalSurvey/CoreScore/. The photographs analysed are available for download from the British Geological Survey website, https://www.bgs.ac.uk/information-hub/photos-and-images/

Core material from boreholes is critical to the understanding and modelling of subsurface systems. However, acquiring new core is an expensive operation and, in addition to this, UK onshore drilling projects have come under increased public scrutiny due to perceived environmental risks and impacts (Ireland *et al.* 2021). This adds additional complexity to the development of new onshore drilling projects and, as a result, acquisition of new core material.

In the absence of new core material, subsurface research relies heavily on archives of legacy core for the UK landmass and continental shelf. Over 600 km of core material is stored as part of the National Geoscience Data Centre (NGDC) hosted at the British Geological Survey (BGS) Keyworth site. The NGDC archive underpins a huge volume of subsurface research, from large-scale characterization studies (e.g. NIREX 1997; Andrews 2013; Monaghan and The Project Team 2016) to small-scale physical property studies (e.g. Fellgett *et al.* 2019; Payton *et al.* 2021).

Regardless of how carefully core is handled and stored post acquisition, it will degrade over time. The condition of legacy core can be highly variable and is dependent on a number of factors including initial core recovery, sampling (pre- and post-delivery to the NGDC), lithology and physical properties. Some lithologies such as well-cemented sandstones may not degrade much over time but others such as shales may degrade within years of acquisition. Storage techniques such as wrapping and refrigeration can help preserve core condition, but such techniques are costly and due to the volumes of core stored in the NGDC only a small proportion of the core can be preserved in this way. Core which is not specially preserved for the longer term can develop core breaks (biscuiting), post-acquisition salt crust or crumble into small pieces (rubbling) (Fig. 1).

Storage conditions, including humidity, sealing and temperature, alongside manual handling, will also impact core condition. Each of these factors will influence not just the core condition but also the ongoing capacity of that core to be used for research and sampling. For example, if required to sample a 20×50 mm plug for triaxial testing, it may involve examining tens to hundreds of metres of core in order to find core pieces of sufficient

From: Neal, A., Ashton, M., Williams, L. S., Dee, S. J., Dodd, T. J. H. and Marshall, J. D. (eds) 2023. *Core Values: the Role of Core in Twenty-first Century Reservoir Characterization*. Geological Society, London, Special Publications, **527**, 137–151. First published online March 14, 2023, https://doi.org/10.1144/SP527-2021-200

Fig. 1. Image of a 50 cm section from three legacy cores selected to show core in poor condition from three different boreholes stored in the National Geoscience Data Centre. Two of these cores (centre and right) have degraded significantly since acquisition making it difficult to undertake certain types of research. The left-hand core is still suitable for some research, but it would not be possible to obtain a core plug from. Source: contains British Geological Survey materials © UKRI 2022.

size and quality to enable the taking of viable samples. In some highly heterogeneous lithologies there may not be an appropriate core section available to take viable samples (Fig. 1).

Issues with core condition also present a problem for acquisition of non-destructive analyses of core and core scanning techniques. Certain cores or sections of cores may be too broken to allow taking of plugs or thin sections. This reduces the type of analytical work that can be undertaken.

Core scanning allows for consistent and relatively rapid core property measurements, including geophysical, geochemical and structural analysis. The application of core scanning data is manyfold and data can underpin various geological disciplines, including mineral exploration studies (e.g. Tappert *et al.* 2011; Fresia *et al.* 2017), petroleum geology (e.g. Blunt *et al.* 2013; Zhang *et al.* 2019), geotechnics and geohazards (e.g. Kuras *et al.* 2016; Harraden, *et al.* 2019), nuclear waste management (e.g. Smith *et al.* 2020), and environmental studies (e.g. Frisia *et al.* 2012; Ruhl *et al.* 2016). However, some of these core scanning techniques, such as hyperspectral and X-ray fluorescence (XRF), only investigate the top few millimetres of the core. These techniques can be influenced by changes in core surface condition from core breaks to core samples, which leads to a reduction in signal and

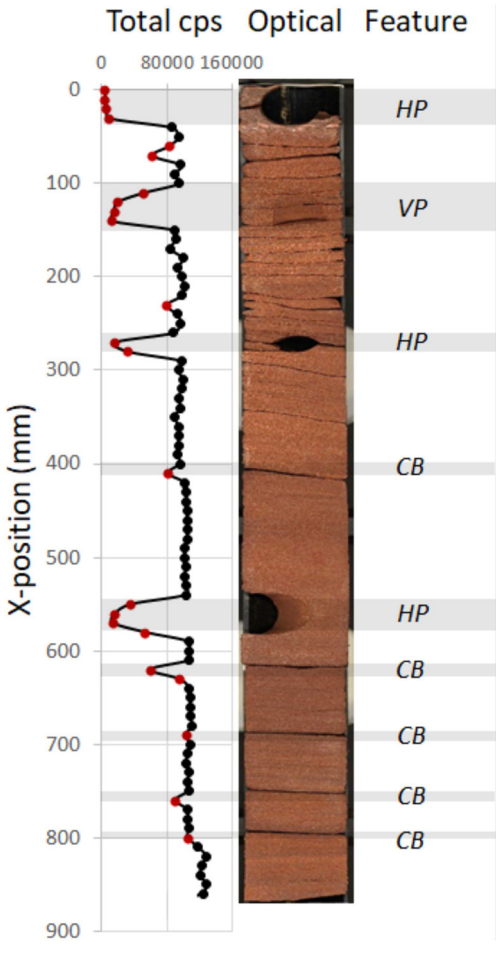

Fig. 2. Image showing how surface profile core scanning (in this example X-ray fluorescence) is affected by core condition. The highlighted core plugs and associated loss of material result in a significant reduction in signal (recorded in total counts per second). Left: depth in core scan section. Centre left: total raw unprocessed counts from XRF surface profile. Centre right: optical image. Right: highlighted features that have an impact on quality of core scan data. HP, horizontal plug; VP, vertical plug; CB, core break.

unreliable results that must be removed from the final analysis (Fig. 2).

An example of this is the work undertaken by the BGS Core Scanning Facility (Damaschke *et al.* 2023). The core scanning facility is co-located with the NGDC at the BGS Keyworth site, and allows for the collection of large quantities of images, physical and chemical property data from the legacy core archive. A description of the facility and its applications can be found in Damaschke *et al.* (2023).

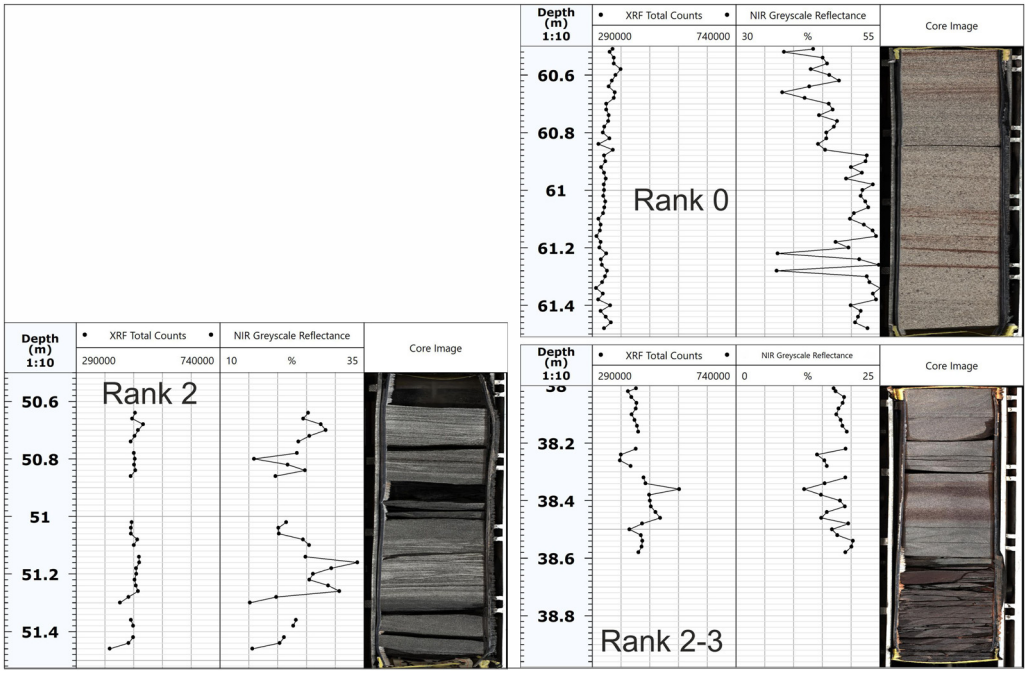

Fig. 3. Example showing how surface profile core scanning intervals are adjusted based on core condition using visual grading system from Damaschke *et al.* (2023). As core quality reduces, surface profile scanning can only be collected over smaller areas of the core. Rank 4 core has not been included as it is not possible to collect surface profile data on such cores (Fig. 1).

When undertaking core scanning BGS currently uses a simple visual assessment of suitability for surface profile scanning. Cores are graded from 0 to 4, with grade 0 considered as representing the best quality surface for profiling data and grade 4 being unsuitable for surface profile scanning (Fig. 3) This methodology is further discussed in Damaschke *et al.* (2023).

Existing methods for assessing core condition are largely based on the Rock Quality Designation (RQD) originally proposed in 1968 (Deere 1968) and reviewed in 1989 (Deere and Deere 1989). RQD specifically provides a mechanism of assessing rock quality at a drill site shortly after its recovery. It is based on the number of natural discontinuities and core loss measurement to calculate an index that expresses core condition from 0% (very poor) to 100% (very good).

The minimum unit of core used by RQD is 10 cm in length and is bounded by natural discontinuities. Induced fractures caused by the drilling and handling of core are not factored into the RQD calculations. Where these features fragment the core, they are ignored and the core length is measured between the two closest natural discontinuities (Deere and Deere 1989).

To be representative of the *in situ* condition of the rock, RQD must also be collected on site shortly after the core is drilled as certain lithologies, such as clays, and shales often break up. Because of this RQD represents a snapshot of core condition; in some clay lithologies RQD may shift from 100% to 0% within hours or days due to post-acquisition core fragmentation (Deere and Deere 1989).

As a result, despite the extensive applications of RQD in engineering geology, it cannot be applied to legacy cores due primarily to fragmentation through drying, core fragment size and minor impacts from disturbance by transport, handling, and storage. A new method of assessing legacy core condition is proposed in order to improve user interaction with the physical material in the NGDC.

Due to the volume of material currently held in the NGDC this method must be automated. One candidate for such automation is the BGS core photography archive, which contains over 125 000 images of core and is detailed below.

British Geological Survey core photography dataset

Procedures for the acquisition of core photography in BGS were set up during the transfer of core from Edinburgh to Keyworth during the closure of the Gilmerton Core Store in 2010 (Howe 2011). In order to

demonstrate that the move did not disrupt the core, a decision was taken to photograph the core before transportation and upon arrival for a subset to assess whether any damage had been caused.

This resulted in the creation of an archive of 125 000 images over an 18-month period. Since 2012 core photography has continued in a less intensive fashion, with newly accessioned core being prioritized.

The core photographs are taken using a Phase One 645DF camera fitted with the Schneider 'Blue Ring' 55 mm f/2.8 lens and the Phase One P45+ digital back. The coreboxes themselves are placed in custom-fabricated plastic trays supported on a custom-built roller table. A ledge at the back of the table supports any required scale or colour calibration bars, as well as the 7″ LCD screen, which displays the core and depth information (Fig. 4). For more detailed information on core photography acquisition, see Howe *et al.* (2012).

The use of the plastic trays means that each photograph can contain 1–6 m of core (*c.* 3–18 ft), depending on core diameter and configuration of the containing corebox. As a result, the original 125 000 images contained 175 000 coreboxes. This gives a maximum core length of 175 000 m, though some coreboxes will hold less than 1 m of core due to losses during drilling and the end of core runs. Following acquisition, the images are processed and converted into JP2 images and JPEG thumbnails. These are then made publicly available through the BGS Photographs and Images webpage (BGS 2021).

The core photographs are a valuable research resource in themselves, but the consistent nature of the acquisition and the large number of images make them an ideal candidate for automated analysis (Martin *et al.* 2021). The number of images also provides information on a wide range of lithologies and types of core, providing opportunities to assess additional factors impacting core condition (Fig. 1).

Image analysis

Any automated system designed to assess core condition from image data alone must be capable of distinguishing between individual core fragments. Traditional approaches that could have been used include segmenting based on pixel values or the use of an edge/line detection algorithm. However, such approaches are unlikely to yield accurate results for the following reasons.

Pixel-based segmentation requires pre-defining a series of 'rules' that classify an individual pixel based on the red–green–blue (RGB) colour values of that pixel. Unfortunately, the wide variety of samples in the core images leads to a wide variety of valid pixel values between individual core fragments. This means any predefined pixel values cannot be used to reliably distinguish core fragments from image to image. For example, the shaded area in an image of a lighter colour rock, may match the unshaded region in a dark rock.

More fundamentally, pixel-based segmentation considers only individual pixels in isolation. Much of the information stored in an image is contained in the spatial context, i.e. the relationship between a given pixel and those around it. By considering

Fig. 4. Example of a core photograph from the BGS core photography collection. Source: contains British Geological Survey materials © UKRI 2022.

individual pixels, but not those in the nearby vicinity, such a rules-based approach excludes much of the available information. While it would be possible to extend the rules-based approach to include nearby pixels, additional pre-processing would still be required to normalize rock shades between images, leading to an ever more complex manually defined set of rules.

The challenges with edge/line detection algorithms associated with the core images is that it is difficult to distinguish edges between core fragments, background and shadows, as they are visually dissimilar. In addition, fragments themselves tend to contain edges between light and dark areas, which may lead to an edge detection algorithm identifying one core fragment as two (or more) fragments.

Traditional line detection algorithms also require the user to specify a number of parameters that constrain the sensitivity of the detection. In the case of the core images, it is difficult to determine the correct level of sensitivity due to the range of interfaces between surfaces. Finally, the most prominent edges within the photographs are those of the boxes the core is stored in (AlZayer 2019). All of these factors mean that it is not possible to predefine an algorithm with a single set of parameters that define an edge.

Rather than predefining parameters and using them to segment images into core fragments, a more pragmatic approach is to use an algorithm capable of automatically learning the parameters necessary to segment images. Such algorithms are broadly defined as 'machine learning' (ML). Most ML approaches trade off the necessity to predefine parameters for a relatively large, representative training set. For this reason, the problem presented here is an ideal candidate for an ML-based solution.

Workflow summary

ML algorithms can be divided into 'supervised' and 'unsupervised'. Supervised algorithms require a labelled dataset, where a label refers to a desired output for every data point. A 'data point' in this case refers to a single pixel from a core image. An unsupervised algorithm would take the data with no labels and attempt to extract underlying patterns, typically by clustering similar data points together. Although this could be conceptually useful, the aim of this study was to automatically label pixels so a supervised approach was necessary and this choice largely dictated the designed workflow.

Supervised algorithms require, at minimum, a training set and a distinct testing set. Every data point (X) in both the training and testing sets requires a corresponding label (Y). For image segmentation problems, the algorithm is initiated with randomly assigned parameters and makes predictions on a batch of training set images. These predictions are compared to the labels for those images, and the internal parameters are adjusted to minimize the difference between the prediction and actual values. This process is repeated across the entire training set. To assess model performance, the now trained model is used to predict outputs for the testing set and these are compared to the test set labels.

For this project, the data consist of 29 core images, a subset of the original 125 000 core images. This subset was split into 25 training images and 4 test images. For most ML algorithms, such a small number of distinct images would be too few for meaningful training. However, in this case, the individual images were of relatively high resolution (4784×7107px). This meant the dataset contained a sufficient amount of information to train up a classifier when given the appropriate choice of algorithm. Five additional unlabelled images were also available for qualitative model evaluation. These were manually selected from the core library as images that would be difficult to segment manually due to poor core condition.

To provide labels for each image, a mask was produced. This consisted of an array identical in size to the original image. The array consisted of values from 0 to 5, where each value represented the class of the corresponding pixel on the original image. The classes were: void, rock fragment, paper, core plug, text and box. The open source tool, label-tool (Kim and Veulemans 2021), was used to perform the labelling, allowing users to draw polygons around appropriate regions of each image and label them accordingly (Fig. 5). The labels were saved as .json files and converted into the image masks in .png format.

Following the labelling exercise, a decision was made to combine the labels for Rock Fragment 1 and Rock Fragment 2 (Fig. 5). The initial decision for the use of two categories was to distinguish if a photograph had multiple cores in. Every photograph will have at least one core present, though its position in the image may vary. Splitting the rock into two categories that were only valid for specific sections of the image reduced the impact of the training data. To mitigate this after the labelling process, the workflow converts Rock Fragment 1 and Rock Fragment 2 into a single category.

Algorithm choice and architecture

The chosen algorithm for image segmentation was a U-Net. U-Nets are a sub-class of convolutional neural networks (CNNs), which are a family of neural networks designed for image analysis. A U-Net was specifically chosen as this architecture is designed for image segmentation and performs well on training datasets with a limited number of

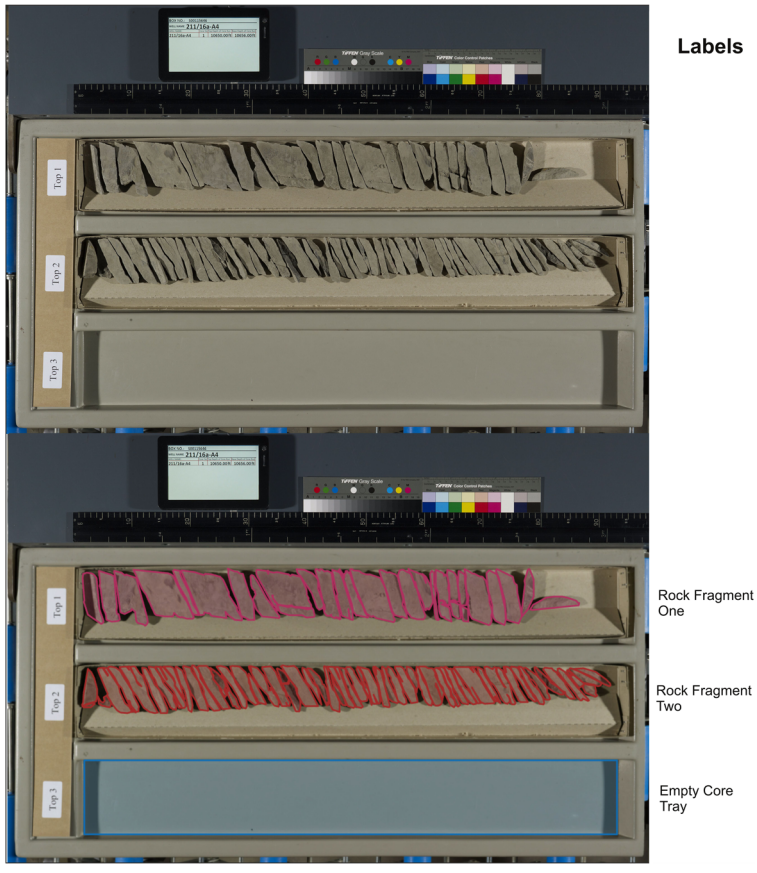

Fig. 5. Example of a labelled image used in the machine learning (ML) workflow. Human operators traced round each core fragment to create a series of labelled polygons known as a mask. This was then input to the processing workflow to train the model. Source: contains British Geological Survey materials © UKRI 2022.

relatively high-resolution images (Ronneberger *et al.* 2015). This is in contrast to more general CNN architectures that tend to require a large number of relatively low-resolution training images (Krizhevsky, *et al.* 2017).

U-Nets consist of an encoder and a decoder. The encoder repeatedly increases the image dimensionality by applying 2D kernels. This results in a growing 'stack' of 2D 'filters', which are trained to identify distinctive features in the image as the kernel parameters are adjusted. The decoder reverses this process; progressively reducing the dimensionality of the stack until a 2D image is output. Using an appropriate loss function, the network seeks to minimize the error between this output and the corresponding mask for the image.

The U-Net utilized in this study used the Resnet34 architecture for the encoder–decoder layers, as the resnet family have a long history of good performance for image classification problems and are easy to implement (He *et al.* 2015). In this case, Resnet34 was available as a pre-built architecture in the FastAI library (Howard *et al.* 2018) which was used to build and train the U-Net. The code used is available at the project github repository (Walsh *et al.* 2021).

Model training and testing

The U-Net was initialized with weights and biases from a Resnet34 model pre-trained on the ImageNet dataset (Deng *et al.* 2009). Using a pre-trained model reduced overall training time as, although ImageNet contains 1000 classes that have no relation to the core segmentation problem, many of the resulting filters were expected to correspond to important features within our images. Training was carried out on a NVidia Quadro RTX 4000 over 100 epochs, requiring 45 min in total. Binary cross-entropy between the prediction and the mask was used as a

loss function, where the model sought to maximize the number of pixels in the prediction and the mask with identical values. Pixel accuracy was also recorded as a training metric.

Of the 25 images in the training set, 3 were used for validation to calculate model training metrics. Therefore, 22 images were used to train the model, 3 to validate and 4 to test. The trained model was then used to predict masks for the test set. These predictions were compared to the actual masks for the test set images to obtain an overall pixel accuracy for the model.

Results

The progression of model loss and accuracy over the training epochs are plotted in Figure 6. As expected, training loss fell asymptotically with the number of epochs as the model was able to distinguish between increasingly fine features. Unusually, the validation loss was generally lower than the training loss. This is probably due to the small number of underlying images forming the validation set – just 3 raw images. As the training set contained more variety, it is reasonable to assume that prediction on the validation set in this case was relatively easy.

Performance of the model on the test set images is shown qualitatively in Figure 7. In addition, predictions for the unlabelled images, which were characterized as 'difficult' to manually label by the project, are shown in Figure 8. In both cases the predictions

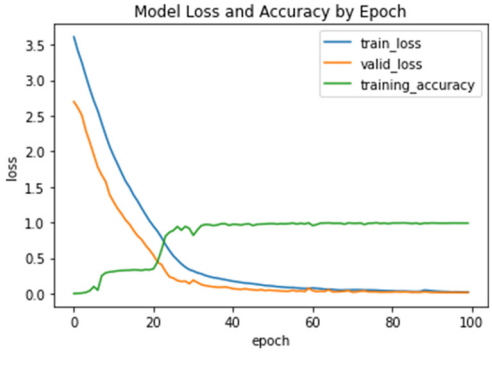

Fig. 6. Progression of model loss and accuracy values over the course of the training run. The accuracy score is the number of correct predictions made on the training dataset. The loss is the difference between the desired target state (training data) and the current model output. As the model begins to correctly identify core fragments, the training accuracy increases, i.e. more core fragments are correctly identified. There is no sign of an accuracy loss discrepancy, i.e. the loss continues to fall, while the accuracy remains the same. If the model was over trained, a single wrong prediction would lead to a significant difference in loss, with no change in accuracy.

clearly exhibit an ability to distinguish between core fragments as well as other image artefacts.

Overall predictive accuracy over the test set was 97.3%. This meant the model correctly predicted the class of 97.3% of pixels in the test set images. Model accuracy on predicting the rock classification was considered separately as identifying rock fragments is the most important application for the model. For the test set, rock prediction accuracy is summarized as follows:

- The precision was 87.7%, meaning 87.7% of pixels in the test set images that were labelled as 'rock' were correctly classified as such.
- The false positive rate was 5.7%, meaning 5.7% of pixels in the test set images that were predicted to be 'rock' were incorrectly classified as such.
- The false negative rate was 2.3%, meaning 2.3% of pixels in the test set images that were predicted not to be rock were actually rock.

The false positive and false negative predictions are shown in Figure 9. In general, the relatively high false negative rate appears to be due to the model misidentifying areas with few surface features as card. Many of these rock areas have a similar colour to the card in the background, so this is unsurprising.

Another contributor to the false negative rate is the model correctly identifying rock plugs as 'non-rock' areas. These predictions are actually correct but were mislabelled by the human operators in the masks. So, these areas actually artificially increase the false negative rate in this case.

The false positive rate is largely derived from the model being unable to distinguish the core breaks between individual fragments. Although the rate is relatively low, correct identification of these areas is crucial for the proposed applications as inferring core breaks is the only reliable means to automatically count the number of distinct core fragments. This led to a model which does appear to predict core condition metrics proportional to the actual values, but is not able to reliably state the number of fragments in a highly fragmented core. Future improvements may be possible through more extensive training on a training dataset that contains more of these highly fragmented core images. However, a balance may have to be struck between high model sensitivity for core breaks and higher rate of false positives for other artefacts that may appear superficially similar to core breaks.

The edges of core samples (i.e. the interface between the core and the box) also contribute to both the false positive and false negative rate. This is possibly due to the presence of shadows in these areas. Errors in these areas are of less concern for automating core condition assessment.

The pixel accuracy is only an indicator of the model's ability to distinguish correctly between

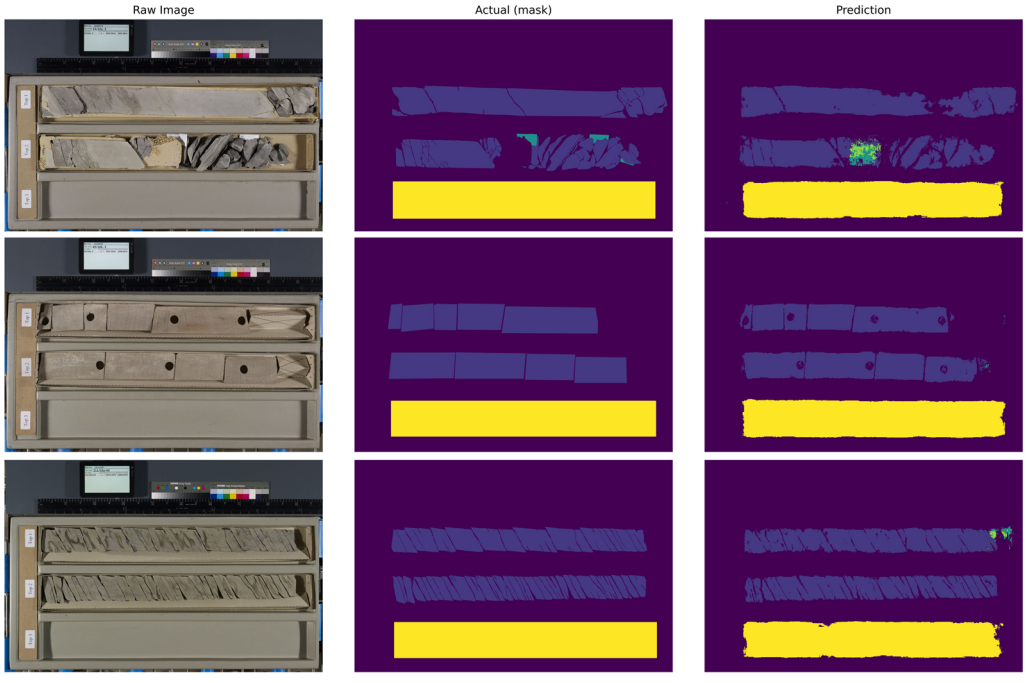

Fig. 7. Original test set images, shown with ground-truth masks and predictions from the trained model. Even with a small training dataset the model can recognize the majority of core within the box. Left-hand panel: original core image; centre panel: mask generated by human operator (Fig. 5); right-hand panel: prediction of core from trained ML model when applied to the original image. Source: contains British Geological Survey materials © UKRI 2022.

different types of material in the image. It does not directly produce an indication for degree of fragmentation. In order to do this, a number of metrics were derived from the pixel values in the prediction masks. The prediction masks were used to derive the following metrics for each image:

- Relative Rock Area: the proportion of the image taken up by rock.
- Total Rock Perimeter: the total perimeter of all rock regions in pixels. More fragmented cores are expected to have a higher perimeter.
- Average Fragment Perimeter: the average perimeter of every individually identified rock fragment, where an individual rock fragment is a single contiguous zone of 'rock' pixels (automatically enumerated by the regionprops function from the scikit-image package).
- Number of Fragments: the total number of distinct rock fragments distinguished by the model.
- Total Rock Area: the number of pixels identified as rock.
- Perimeter Complexity: the total perimeter area divided by the total rock area. Higher quality cores would be expected to have a less complex perimeter.

All predicted parameters are shown in Table 1.

Discussion

Neural network approaches have been applied to a wide variety of Earth science projects in recent years, from facies prediction (Martin *et al.* 2021) through to forecasting of sea ice (Andersson *et al.* 2021). These types of approaches are favoured due to the generation of reliable results from a small amount of training data. Existing python libraries such as fastai (Howard *et al.* 2018) and TensorFlow (Abadi *et al.* 2015) also reduce the time taken to produce workflows to address this problem.

The machine learning-based approach demonstrated here has the immediate benefit of near-instant assessment of new images. The training set required an average of 30 min per image to label; however, complex images (e.g. Fig. 5) took several hours to label precisely. Thus, the ML approach provides clear time-saving benefits while producing consistent predictions, which is not guaranteed with human operators.

Direct time-saving comparisons between the ML model and human operators is less relevant than the performance of the model itself. However, in this

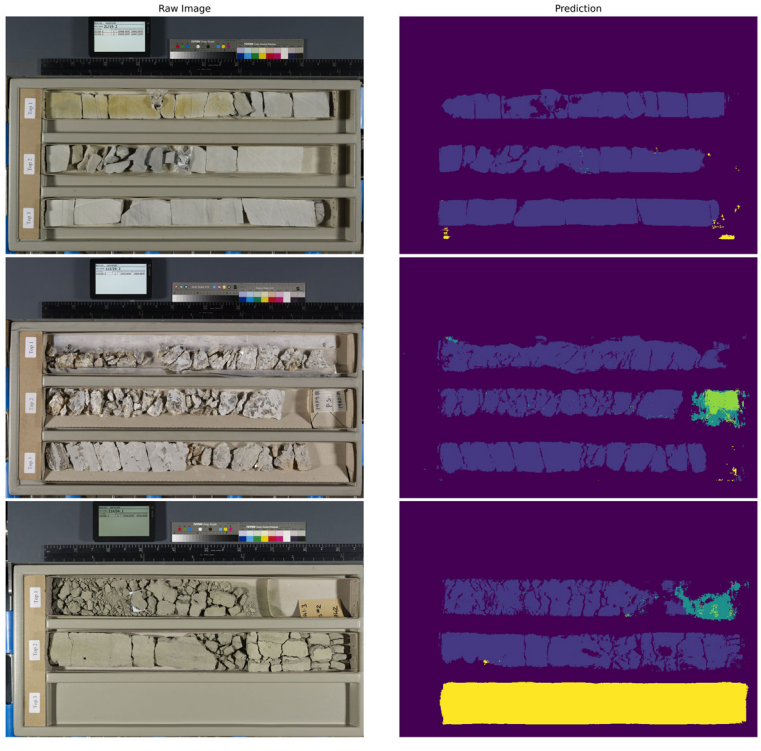

Fig. 8. Original images (left) and model predictions (right) for core identified as being in particularly poor condition. Given the condition of the core, these images were not labelled by human operators due to the time required to produce a mask. The ML model prediction captures a significant amount of the core variability even with a small training dataset. Source: contains British Geological Survey materials © UKRI 2022.

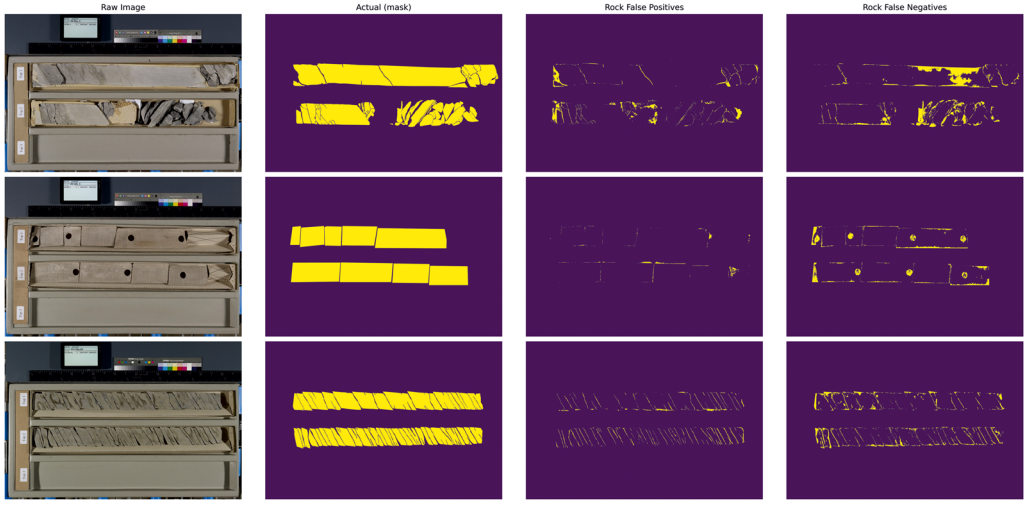

Fig. 9. Visualization of the areas of test set images that were mislabelled for the 'rock' class. Left: original image; centre left: mask produced by human operator; centre right: areas of the core where the model predicted core but the human operator did not; right: areas where the human operator labelled the image as rock but the model did not. The false negatives in the second image show the model recognizing core plugs even though these were not specifically labelled in the training dataset. Source: contains British Geological Survey materials © UKRI 2022.

Table 1. *Predicted metrics for the test set images shown in Figure 7*

Image	Relative rock area (pixels)	Total rock perimeter (pixels)	Average fragment perimeter	Number of fragments	Total rock area (pixels)	Perimeter complexity
test set 1	0.16	8984.79	598.99	15	85 269	0.11
test set 2	0.15	5414.20	541.42	10	81 267	0.07
test set 3	0.15	10 695.59	1782.60	6	80 044	0.13

These parameters may be used as a summary for overall core quality.

case the resources required to undertake this work using human operators is beyond what is practically possible. This is a consequence of the number of images held by BGS, currently 125 000. Manually labelling these images would take, on average 62 500 h or approximately 21 years of a person's time working 8 h a day, 365 days a year. As a result, this task could never be undertaken economically without the use of ML methods, which in this case required on the order of hours to train and provide near real-time labelling. Non-ML automated methods such as those discussed in the 'image analysis' section would still require significant time from a specialist to pre-determine a set of rules for segmentation. By effectively shifting this learning onto the training phase of the algorithm itself, the supervised learning approach allowed a non-domain specialist to produce viable predictions for rock presence.

Areas for model development

The chosen architecture for CoreScore required a relatively small dataset for training compared to a more generic CNN. This is unusual for ML models, but was offset by the advanced image labelling required to construct a dataset. Simpler image classification problems tend to require single labels for an entire image. In this case, producing polygons for every relevant region in the image was the most time-consuming aspect of the process. The polygons themselves provided contiguous regions in the images that the U-Net learned to recognize in new images. Although the predicted outputs of the network were simple pixel masks, it was relatively simple to extract contiguous zones of predicted rock in order to compute the total number of fragments in a given image. It is important to note that the algorithm was not explicitly trained to compute the number of fragments – rather it was scored on its ability to match individual pixels correctly. A further development in the future could utilize a custom loss function to reward the learner for correctly predicting the number of fragments; or a similar derived metric.

The nature of the solution presented here allowed for an easy qualitative assessment of model performance, in addition to the accuracy metric provided by the model. Every prediction on the test set consisted of a prediction mask which can be visualized as an image. Viewing these predicted masks alongside the original images not only allows an operator to visually confirm the model is capable of making sensible predictions but it may also allow for iterative model improvement. For example, in the first test set image (Fig. 7), we see rock regions that are mislabelled as card. If similar misclassifications are seen in many test set images, it may be possible to address this in a future training run.

Correcting repeated misclassifications may involve designing a pre-processing filter to accentuate or remove such regions. Alternatively, a solution may be applied at the model-level by applying a custom loss function, tailored to disproportionately increase the loss value for the specific misclassification (Ebert-Uphoff *et al.* 2021); this would make the model more sensitive to these regions.

An important aspect of CoreScore that has not yet been considered in detail is the balancing of the training dataset. The training images were selected at random by a human operator from the full BGS core image dataset. In initial model tests it was found that performance was worse when applied to lighter-coloured core material than with darker-coloured material (Fig. 10). This is likely a result of lighter cores being under-represented in the initial training dataset. As the full dataset consists of a wide variety of lithologies and rock types, representative samples of many of these lithologies would probably be necessary to train up a classifier capable of distinguishing fragments in any core. It is difficult to predict exactly how many training samples, and the variety thereof that would be required to build a truly general classifier that could predict on any unseen new sample; future studies could feasibly repeat the work of this study with a test set of explicitly different samples to measure sensitivity of the model against training set variability.

However, dataset variability cannot be considered in isolation. An ideal dataset would contain different lithologies with the same relative frequency we expect to encounter them in any new images. The process of representing different classes of data in the correct proportion is referred to as

'stratification' and is not possible to achieve perfectly since we can never know how new unseen data will be stratified. In addition, stratification should be carried out based on other metrics such as core condition. For example, a classifier trained only on core in excellent condition would not be able to segment poor quality core.

Ensuring that the training dataset for future runs is well stratified may prove challenging as automatically assessing these attributes in the existing data is difficult without an existing model. Although the model does not directly suffer from the nondeterministic bias seen in human operators, it is susceptible from bias derived from incorrectly stratified data. To mitigate this, it is anticipated that future use of this tool will require a new training set that proportionally represents all rock types in the NGDC.

Another aspect which may artificially raise the model accuracy is positional bias. All training photographs were of broadly the same format: split into four horizontal sections. The top section housed the colour reference card and background, the second and third sections usually housed core, and the bottom section was usually empty. A learner that simply predicted 'background' for the top 25% of the image, 'rock' for the middle 50% and 'card' for the bottom 25% would score a reasonably high accuracy. Indeed, we see excellent predictive capability for 'card' in the bottom section of all test set images.

To address the issue of positional bias in future iterations, data augmentation through geometric transformation will be necessary (Shorten and Khoshgoftaar 2019). The simplest solution would be to rotate training images in 90° increments and duplicate to remove this positional context. Another approach to this issue would be to slice the images so that each section of core is separated and 'background'/'card' areas are minimized. This may be a more comprehensive approach, but the additional investment required means that the simpler translational technique will be attempted and evaluated first.

The labelling of the training data may also have introduced errors and artefacts in the model. In the initial training phase five types of labels were used, despite the principal interest being confined to the presence or absence of core. Instructions were provided to interpreters but the authors now believe that fewer labels should have been used in the initial development of the tool. This would have improved model performance through the reduction of false negatives. However, some false negatives would have persisted in cases where core plugs have been taken from intact rock fragments (Figs 8 & 9).

One limiting factor on the use of Core Score is the available hardware, specifically GPU memory. The training set was small enough for the computational time to train the model to not be a major concern. However, the high resolution of the input images initially caused the model to exceed the available GPU memory of 8Gb. The solution was to train with a batch size of 1, which was appropriate due to the small number of images. If higher-resolution images are to be used in the future, or a batch size increase is necessary, future implementations would require either a higher performance GPU, or reduction of input image sizes by slicing to only include the relevant sections.

One challenge when applying machine learning techniques to the BGS photography collection is that the photographs were not collected with ML techniques in mind. This results in some artefacts in the processing in particular caused by shadows at the edges of each core box (AlZayer 2019; Fig. 10). In similar projects, this has not posed major issues, especially when the photography is carried out under controlled conditions (Hall *et al.* 2021). A sufficiently well-trained classifier would be able to identify these artefacts and account for

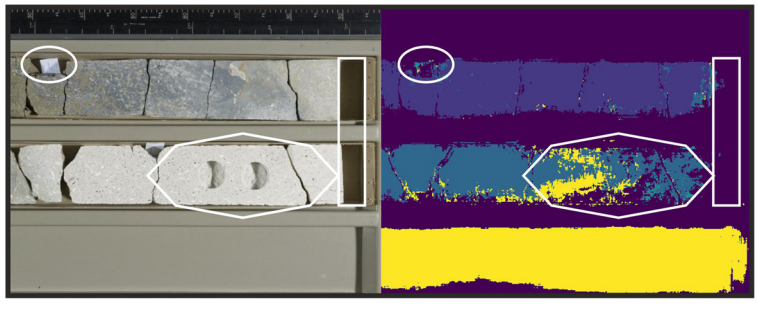

Fig. 10. CoreScore model performance and artefacts on an unseen image. Ovals highlight the effect of paper on the model. Shadows on the core caused by the edge of the corebox are highlighted by the white rectangle. The white polygons show an area of lighter core that is not well identified by the model and is likely the result of lighter material being under-represented in the training dataset. Source: contains British Geological Survey materials © UKRI 2022.

them. However, artefact removal via pre-processing is preferred where applicable. A pre-processing step would be particularly beneficial to the U-Net approach since the training set consists of a small number of high-resolution images, so artefacts will be seen relatively few times during training.

In addition to this there are also many other artefacts in the images introduced from items within the Coreboxes themselves. This includes: Paper Labels (Fig. 10); Sub Sampling (Figs 2 & 10); Plastic/wrappers (Fig. 1), resin and spacers (an object inserted where a section of core is removed. All of these features will impact the performance of the tool, reducing the accuracy of the predictions. The overall result may make CoreScore more suitable for reconnaissance level characterization.

The capability of the model to automatically identify areas of rock allowed for direct computation of parameters that may be used for core assessment. Ideally, the total number of distinct fragments identified in each image would be used as a direct proxy for core condition. This method would likely automatically downgrade finely laminated cores with planes of weakness which may open post-drilling. These cores may have an increased value from a geological perspective, so user requirements need to be considered as part of the process. In addition, discriminating between areas of rock and core breaks was the main contributor to false positive values in the predictions. This led to a model that struggled to accurately identify the pixels in 'transitional zones' between fragments. When deriving core quality metrics from these pixel predictions, there was a tendency for smaller fragments to become single larger contiguous zones which were labelled as a single fragment.

However, the total perimeter of rock in each image was also computed and appears to be proportional to fragment count. This is visually demonstrated in Figure 7 where we see a core with many fragments that are falsely identified as a few large sections of rock. The prediction is still able to trace out a relatively large perimeter around these collections of individual fragments. Future iterations of this methodology may seek to better differentiate between individual fragments, or alternatively seek to characterize core using a single parameter derived solely from the core area and total perimeter.

All of these factors have an impact on attempts to calculate a single core quality index from CoreScore outputs. If the tool was run on a closed dataset where the training images were representative of the core as a whole then a normal distribution could be fitted to calculate a relative core index (AlZayer 2019). Such an index can be valuable for specific use cases but the development of CoreScore is currently geared towards a single tool which can be applied irrespective of core condition. As such, a universal index is not currently the target for immediate development. BGS also continues to acquire new core photographs as part of its Digital Collections programme. These new images will be incorporated into the tool, initially on a borehole-by-borehole basis. It remains a target of the project to introduce core-condition categories based on the outputs of the tool. It is intuitive that a core with fewer core breaks and higher total rock area will be in a better condition than one with more breaks and a lower rock area. Though at this stage the model has only been run on a small number of images so a proposed classification would not be meaningful beyond this specific dataset.

Another method to assess core condition would be to examine core fragmentation and fragment size for different core diameters and lithologies to create an index by rock type. It is, for example, expected that clay- or mud-rich lithologies will have degraded while in storage (Deere and Deere 1989) so may have larger numbers of core breaks. Furthermore it is anticipated that non-geological parameters such as core diameter or whether the core was archived in whole round or slabbed state will also impact core condition. These parameters are not currently stored in machine-readable formats and the data extraction will be a time-consuming process. However, it remains the authors' ambition to capture this information and incorporate it into the tool.

CoreScore applications

The initial concept for developing CoreScore was to improve efficiency in the BGS Core Scanning Facility by not scanning core that was too broken to generate reliable data. However, there are a number of legacy cores that are of sufficient quality to be scanned but their condition may still impact the scanning data (e.g. Fig. 2). It is difficult to quantify the magnitude of this impact due to the lack of publicly available hard rock datasets from core scanning alongside there being no method of quantifying legacy core condition in a consistent manner.

The impact of core condition on core scanning is also dependent on the purpose of collecting core scan data. If there is a large contrast in rock properties, such as an interbedded sequence of evaporites and muds, volumetric measurements are likely to identify this variation unless the core has rubbled. However, if the user would like to look at chemical changes that can indicate variations in cement type down core then even small fractures can have an impact on the dataset.

Information on core condition is needed to make sense of core scanning data. For example, the dataset from UKGEO's Glasgow shows that within sections of poor core condition, scan data cannot be collected (Fig. 3). However providing a single grade for a box does not distinguish sections of intact core, sections

of missing core and sections of core too fragmented to scan. Thus, outputs from CoreScore could be used to qualitatively assess core scanning outputs and identify if data gaps are associated with missing core or broken core.

CoreScore utilization has implications for methodologies for core preservation and sampling in a Core Store environment. Currently users taking destructive samples from core held at BGS Keyworth must either access core photographs or attend site in person to identify sample locations in conjunction with the chief curator or conservator. Where there is a significant section of core (e.g. a hundred metres) this process can involve a lot of physical effort and time to retrieve the cores and lay them out in viewing bays. CoreScore will streamline this process by prescreening core to identify sections where there is sufficient intact rock to allow for sampling. This will minimize unnecessary core handling, decreasing both risks of damage to core material and manual handling injuries to the repository staff. This objective remains an aspiration that is yet to be realized, but CoreScore represents progress toward this aim.

CoreScore can also promote core preservation. Take, for example, the case where two boreholes are drilled near to each other and sample the same stratigraphy. One borehole is mentioned in a publication and has been heavily sampled but the other has not. CoreScore could identify this and be used to direct users to the second core, allowing preservation of the first borehole while improving the type and quantity of samples a user can take. It could also flag where samples had been taken, allowing individuals to investigate what historical data may be available.

The ability to automatically generate simple metrics to assess core can be used to improve the visualization of long stretches of core. Due to hardware and performance considerations it can be impractical to load hundreds of metres of core images into visualization software. The text outputs of CoreScore allow users to present core condition over hundreds of metres but then load high-resolution photography for sections of interest. In addition to this, if a section of fresh core was repeatedly photographed over time then CoreScore could also be used to assess the speed of core degradation in different lithologies and to inform future curation best practice.

Conclusions

Project CoreScore has demonstrated that ML methods and workflows can be used to rapidly assess core condition from images alone. This methodology has been demonstrated to perform with a high degree of accuracy (over 95%) and precision (87.7%) when compared to a manually labelled image.

CoreScore represents an opportunity to change how users interact with physical core material stored in the NGDC. This has implications for all users of the NGDC, from a PhD student taking a single sample from a core to large basin-scale multi-well characterization studies by industry or consortia.

The utilization of existing U-Net algorithms has allowed this to be achieved with a minimal training dataset. This has not only reduced the time spent compiling training datasets but also means that the model can be easily adapted to focus on lithologies not currently represented in the dataset, initially on a project-by-project basis.

However, the content and collection of the training dataset has implications for the deployment of the tool. Results suggest that model performance would be replicated in lighter lithologies, such as limestones or darker lithologies, such as shales. As a result, ongoing tool development will require the collection of training data from areas of interest, either spatially or stratigraphically.

The model has been shown to segment images that humans would find difficult or time-consuming to manually assess. It also provides a consistent methodology which would be difficult to achieve with human operators. It is worth noting that economically it would be difficult for humans to interpret the existing core photography dataset within BGS based on size alone. As a result, ML tools represent one of the few options of unlocking information from this photography dataset.

Scaling up the existing prototype to the full dataset (125 000+ images) would require rapid assessment of new images and require enhanced computing power. We cannot yet conclude that such a model would be able to achieve high segmentation accuracy on new data from the full range of lithologies housed in the NGDC. However, by scaling up our current implementation of an image processing pipeline and U-Net, we hope to extend the predictive capability of this model across a comprehensive range of core samples.

The workflows utilized in this project have been applied to a number of geoscience problems. This demonstrates the versatility of these types of approaches, particularly to consistent datasets of images, and even legacy datasets where factors impacting machine learning were not considered at the time. An example of this is the performance of CoreScore on core images where items such as plastic and paper obscure sections of the core.

Full utilization of the CoreScore outputs will depend on integration with other processes that involve legacy core, such as core scanning. There are, however, many other potential use cases and discussions with future stakeholders will be critical to the development of CoreScore.

Acknowledgements This manuscript was published with the permission of the Executive Director of the British Geological Survey. Contains public sector information licensed under the Open Government Licence v3.0. The authors would like to thank Zayad AlZayer, Jo Walsh and Vyron Christodoulou for their contributions to this project, and Benjamin Wood, Tanya Richmond, Thomas Fletcher and Jonathan Atherley-Mercantas for their contributions to the training dataset.

The authors would also like to thank Francisco Brito, Romaine Graham, Matthew Haines and an anonymous reviewer for constructive comments to improve this manuscript.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions **MF**: conceptualization (lead), methodology (supporting), project administration (lead), writing – original draft (lead), writing – review & editing (lead); **AH**: data curation (supporting), formal analysis (lead), investigation (lead), methodology (lead), software (lead), validation (lead), writing – original draft (supporting), writing – review & editing (supporting); **SH**: data curation (lead), methodology (supporting), writing – original draft (supporting), writing – review & editing (supporting); **MD**: data curation (supporting), investigation (supporting), methodology (supporting), writing – original draft (supporting), writing – review & editing (supporting); **AK**: conceptualization (supporting), project administration (supporting), resources (lead), writing – original draft (supporting), writing – review & editing (supporting).

Funding Funding from the British Geological Survey (ID0EKTBG3974) to MWF supported this research.

Data availability The code for CoreScore is available at https://github.com/BritishGeologicalSurvey/CoreScore/. The photographs analysed are available for download from the British Geological Survey website, https://www.bgs.ac.uk/information-hub/photos-and-images/.

References

Abadi, M., Agarwal, A. *et al*. 2015. TensorFlow, Large-scale machine learning on heterogeneous systems [Computer software], https://doi.org/10.5281/zenodo.4724125

AlZayer, Z. 2019. Digital Rock forensics using machine learning and computer vision, combining approaches for identifying core integrity and basic classification. Independent project report, MSc, Royal Holloway University.

Andersson, T.R., Hosking, J.S. *et al*. 2021. Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, **12**, 5124, https://doi.org/10.1038/s41467-021-25257-4

Andrews, I.J. 2013. *The Carboniferous Bowland Shale gas study: geology and resource estimation*. British Geological Survey for Department of Energy and Climate Change, London, UK.

Blunt, M.J., Bijeljic, B. *et al*. 2013. Pore-scale imaging and modelling. *Advances in Water Resources*, **51**, 197–216, https://doi.org/10.1016/j.advwatres.2012.03.003

British Geological Survey 2021. Photographs and images: BGS Information Hub. British Geological Survey, https://www.bgs.ac.uk/information-hub/photos-and-images/ (last accessed November 2021)

Damaschke, M., Fellgett, M.W., Howe, M.P.A., Watson, C.J. and Condon, D. 2023. Unlocking national treasures: The core scanning approach.

Deere, D.U. 1968. Geological considerations. *In*: Stagg, K.G. and Zienkiewicz, O.C. (eds) *Rock mechanics in engineering practice, Chapter 1*. Wiley, New York, 1–20.

Deere, D.U. and Deere, D.W. 1989. *Rock quality designation (RQD) after twenty years*. Contract Report GL-89-1. US Army Corps of Engineers.

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

Ebert-Uphoff, I., Lagerquist, R. *et al*. 2021. CIRA Guide to Custom Loss Functions for Neural Networks in Environmental Sciences – Version 1, https://doi.org/10.48550/arXiv.2106.09757

Fellgett, M.W., Kingdon, A., Waters, C.N., Field, L., Shreeve, J., Dobbs, M. and Ougier-Simonin, A. 2019. Lithological constraints on borehole wall failure; a study on the Pennine Coal Measures of the United Kingdom. *Frontiers in Earth Science*, **7**, 163, https://doi.org/10.3389/feart.2019.00163

Fresia, B., Ross, P.S., Gloaguen, E. and Bourke, A. 2017. Lithological discrimination based on statistical analysis of multi-sensor drill core logging data in the Matagami VMS district, Quebec, Canada. *Ore Geology Reviews*, **80**, 552–563, https://doi.org/10.1016/j.oregeorev.2016.07.019

Frisia, S., Borsato, A., Drysdale, R.N., Paul, B., Greig, A. and Cotte, M. 2012. A re-evaluation of the palaeoclimatic significance of phosphorus variability in speleothems revealed by high-resolution synchrotron micro XRF mapping. *Climate of the Past*, **8**, 2039–2051, https://doi.org/10.5194/cp-8-2039-2012

Hall, A., Gillespie, M., Everett, P., Christodoulou, V. and Walsh, J. 2021. Machine learning applied to pore-space geometry in sandstones: a tool for evaluating grain-scale similarity? *Quarterly Journal of Engineering Geology and Hydrogeology*, **55**, https://doi.org/10.1144/qjegh2020-183

Harraden, C.L., Cracknell, M.J., Lett, J., Berry, R.F., Carey, R. and Harris, A.C. 2019. Automated core logging technology for geotechnical assessment: A study on core from the Cadia East porphyry deposit. *Economic Geology*, **114**, 1495–1511, https://doi.org/10.5382/econgeo.4649

He, K., Zhang, X., Ren, S. and Sun, J. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

Howard, J., Gugger, S., Bekman, S., Ingham, F., Monroe, F., Shaw, A. and Thomas, R. 2018. Fastai, https://github.com/fastai/fastai

Howe, M. 2011. Gilmerton Core Sample Collection: Keyworth transfer methodology. British Geological Survey, OR/11/049 (Unpublished), http://nora.nerc.ac.uk/id/eprint/531280/

Howe, M., Tulloch, G. and Giles, J. 2012. High quality core images from UK Continental Shelf. Paper presented at PETEX 2012, London, UK, 20–22 November. Petroleum Exploration Society of Great Britain, https://nora.nerc.ac.uk/id/eprint/19842/

Ireland, M.T., Brown, R., Wilson, M.P., Stretesky, P.B., Kingdon, A. and Davies, R.J. 2021. Suitability of legacy subsurface data for nascent geoenergy activities onshore United Kingdom. *Frontiers in Earth Science*, **9**, 376, https://doi.org/10.3389/feart.2021.629960

Kim, S. and Veulemans, J. 2021. label-tool, https://github.com/Slava/label-tool

Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, **60**, 84–90, https://doi.org/10.1145/3065386

Kuras, O., Shreeve, J., Smith, N., Graham, J. and Atherton, N. 2016. Enhanced characterisation of radiologically contaminated sediments at Sellafield by MSCL core logging and x-ray imaging. Paper presented at the Near Surface Geoscience 2016 – 22nd European Meeting of Environmental and Engineering Geophysics, 2016, cp-495. European Association of Geoscientists & Engineers, https://doi.org/10.3997/2214-4609.201601918

Martin, T., Meyer, R. and Jobe, Z. 2021. Centimeter-scale lithology and facies prediction in cored wells using machine learning. *Frontiers in Earth Science*, **9**, 491, https://doi.org/10.3389/feart.2021.659611

Monaghan, A.A. and The Project Team 2016. *Overview of the 21CXRM Palaeozoic Project – a regional petroleum systems analysis of the offshore Carboniferous and Devonian of the UKCS*. British Geological Survey Commissioned Report, CR/16/047.

Nirex 1997. *Sellafield Geological and Hydrogeological Investigations: Assessment of Insitu Stress Field at Sellafield.* Nirex Report S/97/003.

Payton, R.L., Fellgett, M.W., Clark, B.L., Chiarella, D., Kingdon, A. and Hier-Majumder, S. 2021. Pore-scale assessment of subsurface carbon storage potential: implications for the UK Geoenergy Observatories project. *Petroleum Geoscience*, **27**, https://doi.org/10.1144/petgeo2020-092

Ronneberger, O., Fischer, P. and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/arXiv.1505.04597

Ruhl, M., Hesselbo, S.P. *et al*. 2016. Astronomical constraints on the duration of the Early Jurassic Pliensbachian Stage and global climatic fluctuations. *Earth and Planetary Science Letters*, **455**, 149–165, https://doi.org/10.1016/j.epsl.2016.08.038

Shorten, C. and Khoshgoftaar, T.M. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, **6**, 60, https://doi.org/10.1186/s40537-019-0197-0

Smith, N.T., Shreeve, J. and Kuras, O. 2020. Multi-sensor core logging (MSCL) and X-ray computed tomography imaging of borehole core to aid 3D geological modelling of poorly exposed unconsolidated superficial sediments underlying complex industrial sites: An example from Sellafield nuclear site, UK. *Journal of Applied Geophysics*, **178**, 104084, https://doi.org/10.1016/j.jappgeo.2020.104084

Tappert, M., Rivard, B., Giles, D., Tappert, R. and Mauger, A. 2011. Automated drill core logging using visible and near-infrared reflectance spectroscopy: A case study from the Olympic Dam IOCG deposit, South Australia. *Economic Geology*, **106**, 289–296, https://doi.org/10.2113/econgeo.106.2.289

Walsh, J., Christodoulou, V. and Hall, A. 2021. CoreScore, https://github.com/BritishGeologicalSurvey/CoreScore/

Zhang, P., Lee, Y.I. and Zhang, J. 2019. A review of high-resolution X-ray computed tomography applied to petroleum geology and a case study. *Micron*, **124**, 102702, https://doi.org/10.1016/j.micron.2019.102702