Explainable Deep Learning for Automatic Rock Classification

Dongyu Zheng[a], Hanting Zhong[a]*, Gustau Camps-Valls[b], Zhisong Cao[c], Xiaogang Ma[d], Benjamin Mills[e], Xiumian Hu[f], Mingcai Hou[a], Chao Ma[a]

[a] *State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation and Key Laboratory of Deep-time Geography and Environment Reconstruction and Applications, MNR, Chengdu University of Technology, Chengdu, 610059, China*

[b] *Image Processing Laboratory, Universitat de València, Paterna, 46980, Spain*

[c] *College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, 610059, China*

[d] *Department of Computer Science, University of Idaho, Moscow, ID 83844, USA*

[e] *School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK*

[f] *School of Earth Sciences and Engineering, Nanjing University, Nanjing, 210023, China*

* *Correspondence:* zhonghanting@cdut.edu.cn

ARTICLE INFO

Keywords:

Explainable deep learning

Knowledge-infused machine learning

Model interpretability

Attention-based modal network

Rock classification

Authorship contribution statement

DZ, Conceptualization, Methodology, Writing & Editing - Original Draft; **HZ**, Formal analysis, Investigation, Writing - Review & Editing; **GC**, Writing - Review & Editing; **ZC**, Methodology, Software, Formal analysis, Writing - Original Draft; **XM**, Writing - Review & Editing; **BM**, Writing - Review & Editing**;** **XH**, Writing - Review & Editing; **MH**, Supervision, Funding acquisition; **CM**, Project administration, Funding acquisition

Abstract

As deep learning (DL) gains popularity for its ability to make accurate predictions in various fields, its applications in geosciences are also on the rise. Many studies focus on achieving high accuracy in DL models by selecting models, developing more complex architectures, and tuning hyperparameters. However, the interpretability of these models, or the ability to understand how they make their predictions, is less frequently discussed. To address the challenge of high accuracy but low interpretability of DL models in geosciences, we study rock classification from thin-section photomicrographs of six types of sedimentary rocks, including quartz arenite, feldspathic arenite, lithic arenite, siltstone, oolitic packstone, and dolomite. These rocks' characteristic framework grains and grain textures are their distinguishing features, such as the rounded or oval ooids in oolitic packstone. We first train regular DL models, such as ResNet-50, on these photomicrographs and achieve an accuracy of over 0.94. However, these models make classifications based on features like cracks, cements, and scale bars, which are irrelevant for distinguishing sedimentary rocks in real-world applications. We then propose an attention-based dual network incorporating both global (overall photomicrograph) and local (distinguishing framework grains) features to address this issue. Our proposed model achieves not only high accuracy (0.99) but also provides interpretable feature extractions. Our study highlights the need to consider interpretability and geological knowledge in developing DL models, in addition to aiming for high accuracy.

Keywords: Explainable deep learning; Knowledge-infused machine learning; Model interpretability; Attention-based modal network; Rock classification

1. Introduction

Deep learning (DL) has been highly effective in a range of tasks in geosciences, including capturing complex relationships in datasets, creating automatic analysis pipelines and building large and efficient models for numerical simulation and inversion (Bergen et al., 2019; Reichstein et al., 2019; Camps-Valls et al., 2021). This is partly due to the large number of tunable parameters and nonlinear model structures available in DL approaches. However, this over-parameterization also leads to reduced interpretability, making it difficult for users to understand the results obtained with these methods (Castelvecchi, 2016; Buhrmester et al., 2021). The lack of interpretability limits the reliability and application of DL models (Mamalakis et al., 2022). Scientists can neither verify whether the predictions of DL models are made based on reasonable references nor can they improve the models' ability of generalization (e.g., Ebert-Uphoff and Hilburn, 2020).

Since the successful application of deep convolutional neural networks (DCNNs) to the classification of photographs from a dataset of 1.2 million images with one thousand classes (Krizhevsky et al., 2012), there have been numerous attempts to use DCNNs for fossil classification (Romero et al., 2020; Liu et al., 2020, 2022) and mineral classification (Maitre et al., 2019; Hao et al., 2019; Wang et al., 2021; Ge et al., 2021; Zheng et al., 2022) from photomicrographs, cathodoluminescence and scanning electron microscope images in the geoscience community. These efforts have largely focused on evaluating the performance of DCNNs using metrics such as accuracy or mean average precision. However, these numerical values can be sensitive to small changes in the input images, and models with "high accuracy" may not necessarily be robust if they rely on irrelevant features for classification (Lei et al., 2018; Yang et al., 2020). Given the inherent heterogeneity of rocks and other geological objects, it is important to understand how DCNNs make their classifications to ensure their applicability in real-world scenarios. To address this issue, recent advances in interpretability algorithms, such as the Class Activation Mapping (CAM) technique, can provide valuable insights into the decision-making process of DCNNs by highlighting the specific regions of an image that are responsible for the classification using gradient information (Zhou et al., 2016; Selvaraju et al., 2017). While these algorithms have been demonstrated to be useful, they have not yet been widely applied to geoscientific tasks. To develop robust and reliable DL models for geosciences, it is necessary to incorporate interpretability algorithms to deduce the decision-making process of DL models.

Rock classification is a fundamental task in geoscience that involves identifying rock types based on observing framework grains, minerals, texture, and structures. The traditional approach for studying features of rocks in detail consists of first slicing and then mounting, which makes rock samples sliced into roughly 30-micrometers-thick thin sections and mounted on glass slides. Then, for affordability and efficiency, thin sections are usually prepared and photographed as three-channel digital images (red, green, and blue, RGB), also known as photomicrographs. Geoscientists can observe thin sections under polarized light microscopes or examine photomicrographs to observe rock compositions, texture, and other characteristics. Sedimentary rocks cover approximately three-quarters of the Earth's surface, and understanding sedimentary rock types is important for characterizing the Earth's landscape and life over time, as well as for assessing reservoir quality in the oil and gas industry (Dickinson and Suczek, 1979; Garzanti et al., 2007; Boggs, 2009). As a result, photomicrograph examination has become a standard workflow in sedimentary geology, and there have been many attempts to use deep learning approaches to classify rocks based on photomicrographs (de Lima et al., 2019; Koeshidayatullah et al., 2020; Tang et al., 2020; Su et al., 2020; Saxena et al., 2021; Li et al., 2022; Liu et al., 2022). These studies have primarily focused on the accuracy of the models but have not yet investigated how deep learning models make their classifications, which may lead to issues with generalization. This high accuracy yet low interpretability of DL models for geosciences restricts the utility of DL in real-world geoscience work.

In this study, we develop an interpretable rock classification DL model to address this issue by incorporating geological knowledge. We focus on sedimentary rock classification from thin-section photomicrographs, a common classification task in computer vision, as the distinguishing features of sedimentary rocks, such as framework grains and textures, are easy to identify visually. We first applied classical DCNNs, such as ResNet-50, to classify six types of sedimentary rocks and evaluated the performance of these models using numerical metrics (accuracy) and interpretable visualizations generated by Gradient-weighted Class Activation Mapping (Grad-CAM). We then develop and test our new attention-based dual-modal network, SedNet, which integrates global (the whole photomicrograph) and local (characteristic framework grains) features. Our results show that classical DCNNs achieve high accuracy but tend to focus on irrelevant parts of the rock photomicrographs, while our proposed model achieves not only high accuracy but also better interpretability, as indicated by the highlighting of distinguishing framework grains in the Grad-CAM visualizations. This study underscores the importance of interpretability and incorporation of geological knowledge in DL geoscience models. It suggests that integrating global and local information may improve the generalization abilities of DL models in this field.

2. Dataset, data preprocessing, and data augmentation

Six types of sedimentary rocks, including quartz arenite, feldspathic arenite, lithic arenite, siltstone, oolitic packstone, and dolomite, were selected in this study for classification (Figure 1; Table 1). Quartz arenite, feldspathic arenite, lithic arenite, and siltstone are four types of siliciclastic rocks, consisting of grains formed by the decomposition of rocks following weathering and deposition. These grains are typically silicates such as quartz and feldspar, but may also include fragments of igneous, metamorphic, and sedimentary rocks.

The differentiation between siliciclastic rocks lies in the composition of the framework grains and grain size. For instance, quartz arenite is a sandstone with more than 95% quartz grains. Therefore, a representative sub-image for quartz arenite showcases a typical quartz grain. Feldspathic arenite and lithic arenite resemble quartz arenite but contain predominantly feldspar and lithic grains, respectively. As such, their corresponding sub-images feature a feldspar grain and a lithic fragment, respectively. Siltstone is a rock type with smaller grains than sandstone, typically less than 0.063 mm, presenting a distinct silty texture. A region displaying this texture was chosen as the siltstone sub-image, rather than a single mineral grain.

In contrast to siliciclastic rocks, carbonate rocks like oolitic packstone and dolomite are formed through chemical or biochemical processes, consisting primarily of calcite or dolostone. Oolitic packstone, characterized by ooids—rounded or oval grains with concentric textures—is represented by a sub-image showcasing an ooid. Dolomite, known for its high interference colors and euhedral crystal forms, is represented by a sub-image highlighting these unique attributes. In this manner, the sub-images for each rock type have been carefully selected to encapsulate their unique mineralogical and textural characteristics, thereby aiding the deep learning network in differentiating between the classes more effectively.

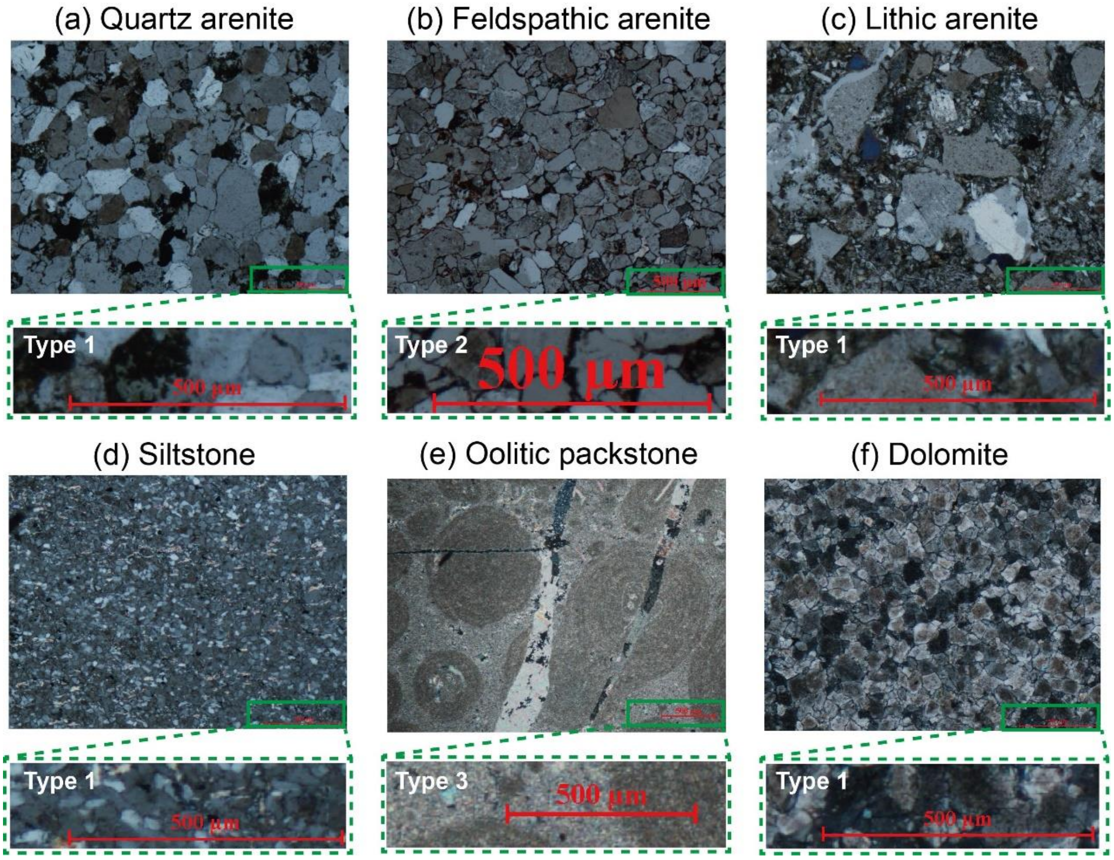

Figure 1. The studied six types of sedimentary rocks and the associated scale bars.

A total of 1356 cross-polarized light photomicrographs were obtained from 15 samples covering the six rock types using high-resolution electronic cameras mounted on Nikon LV100POL microscopes. The images are three-channel RGB with a resolution of 1280 × 860 × 3. The image dataset was split into training, validation, and test sets with a ratio of 6:2:2 (Figure 2). The images were subjected to flipping and rotation to augment the data for model training. The color information in the images was considered important for mineral differentiation and was therefore preserved. Furthermore, the overrepresentation of certain rock types in the dataset (namely quartz arenite, feldspathic arenite, lithic arenite, and dolomite) as opposed to others (siltstone and oolitic packstone) was motivated by the inherent complexity associated with distinguishing these rock types. Quartz arenite, feldspathic arenite, lithic arenite, and dolomite tend to exhibit homogenous grain compositions and textures, thereby posing a significant challenge in their identification. Conversely, rock types such as siltstone and oolitic packstone demonstrate distinct characteristics, such as fine-sized silty textures and oolitic grains, respectively. This deliberate imbalance in the dataset was designed to accommodate these differential complexities inherent in rock identification.

Table 1. Rock type descriptions

Rock type	Framework grains	Grain size (mm)	Distinguishing features	Scalar bar type
Quartz arenite	>90% quartz grains, trace feldspar or lithic fragments	~ 0.25-0.5	Over 90% quartz grains that are gray and clean under XPL	Type 1
Feldspathic arenite	~40-60% quartz grains, 20-40% feldspar grains, the rest are lithic grains	~ 0.25-0.5	Over 1/5 feldspar grains that are grey and mostly dirty. Tabular in shape, with cleavages or twinning under XPL	Type 2
Lithic arenite	30-50% lithic grains, the rest are quartz and feldspar grains	~ 0.063-0.5	Over 1/3 lithic grains that are volcanic or sedimentary rock fragments	Type 1
Siltstone	Mostly are quartz and feldspar grains	<0.063	Silty-sized grains that are grey under XPL	Type 1
Oolitic packstone	Ooids	0.25-2	Oval or rounded ooids grains with concentric fabrics	Type 2
Dolomite	Dolostone	<0.063	Euhedural-subhedural dolostone that are colorful under XPL	Type 1

See scalar bar type in Figure 1.

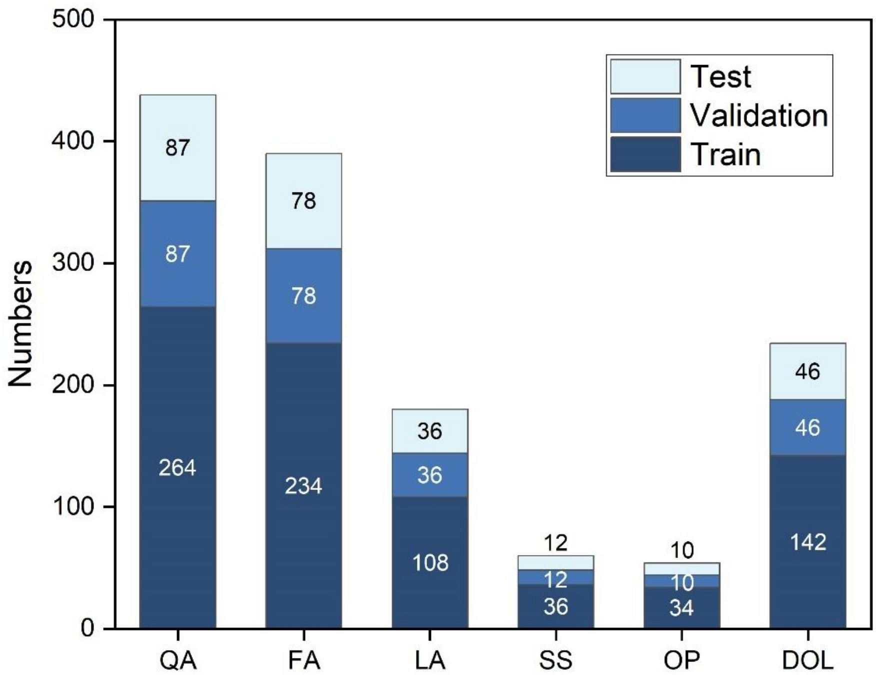

Figure 2. Numbers of training, validation, and test datasets. QA, quartz arenite; FA, feldspathic arenite; LA, lithic arenite; SS, siltstone; OP, oolitic packstone; DOL, dolomite.

3. Model implementation and interpretability

3.1 Model architecture

We introduce a dual-modal network called SedNet to classify sedimentary rocks using thin-section photomicrographs. Traditional DCNNs can only focus on a small area in the images due to the limited size of the convolution kernels. Therefore, SedNet incorporates an attention mechanism and dual-modal input to obtain both kernel-sized and grain-sized information, resulting in improved performance. The architecture of SedNet is depicted in Figure 3a. It consists of four modules: a dual feature extraction module, a channel attention module, a fused feature extraction module, and an output module. The dual feature extraction module comprises two parallel CNNs, each with two Residual Convolution blocks and a global pooling layer (Conv blocks in Figure 3a) activated by Rectified Linear Unit (ReLU) which returns the input values if the input is positive, and 0 if the input is negative. To capture both global and local features in the rock classification, the dual feature extraction module takes as input both the rock photomicrographs and cropped images of distinctive framework grains within the original photomicrographs. For instance, for feldspathic arenite images, one input would be the original thin-section images, while the other would be a representative euhedral-subhedral feldspar grain with twining (see Section 2 for details). The channel attention module includes Squeeze-and-Excitation (SE) blocks, adapted from SE-Net (Hu et al., 2018). These SE blocks enable the neural network to emphasize important features and suppress less important features of the input data. This assignment procedure of feature importance is achieved through the Squeeze and Excitation operations. The Squeeze operation is a global pooling operation that converts a N × N × C matrix ($\mathrm{U_c}$) into a 1 × 1 × C matrix ($\mathrm{Z_c}$), as shown in Eq. (1):

$$\mathrm{Z_c} = \mathrm{F_{sq}}(\mathrm{U_c}(i,j)) = \frac{1}{\mathrm{N}^2}\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{N}} \mathrm{U_c}\,(i,j), \quad (1)$$

Where $\mathrm{F_{sq}}$ represents the Squeeze operation, c represents c^{th} channel, i and j represent the element of the i^{th} row and j^{th} column, respectively. The output of the Squeeze operation is a vector that contains information of the input feature maps (Squeeze vector in Figure 3a).

The Excitation operation takes the output of the Squeeze operation as the input, and then produces a set of channel-wise weights for each feature. This operation processes the 1 × 1 × C matrix with two fully connected layers and applies the sigmoid activation function to limit the output values to the range between 0 and 1. These values are then multiplied by the original N × N × C matrix for each channel, as shown in the following equations:

$$\alpha = F_{ex}(Z, W) = \sigma(g(Z, W)) = \sigma(W_2\delta(W_1 Z)), \quad (2)$$

$$X_c = F_{scale}(U_c, \alpha_c) = \alpha_c U_c, \quad (3)$$

α is a vector representing the weight of each channel, W_1 and W_2 are learnable weight matrices. g represents the dimensionality-reduction procedure, δ denotes the ReLU activation function, σ denotes the sigmoid activation function, X_c is the output of the channel attention module, α_c is the c^{th} element of α. The output of the Excitation operation contains information from the input feature maps with weighted feature importance (Excitation vector in Figure. 3a).

The fused feature extraction module operates the Hadamard product, also known as the element-wise product, and this operation enables the neural network to contain both global and local information from the input rock images. The Hadamard product takes two matrices of the same size and produces a new matrix where each element is the product of the corresponding elements of the original matrices (Fused Conv blocks in Figure 3a; as shown in Eq. 4). For example, given matrices A and B of dimensions $m \times n$ from the previous channel attention blocks, the fused feature extraction is the Hadamard product of A and B:

$$(A \odot B)_{ij} = (A)_{ij} * (B)_{ij}, \quad (4)$$

$\odot$ and $*$ denote the Hadamard operation, which is an element-wise multiplication, i and j represent the element of the i^{th} row and j^{th} column, respectively. The output module consists of global pooling (Fused feature vector in Figure 3a), fully connected and softmax layers (softmax in Figure 3a) that calculates the predicted probability for each category and outputs the category with the highest probability as the final prediction.

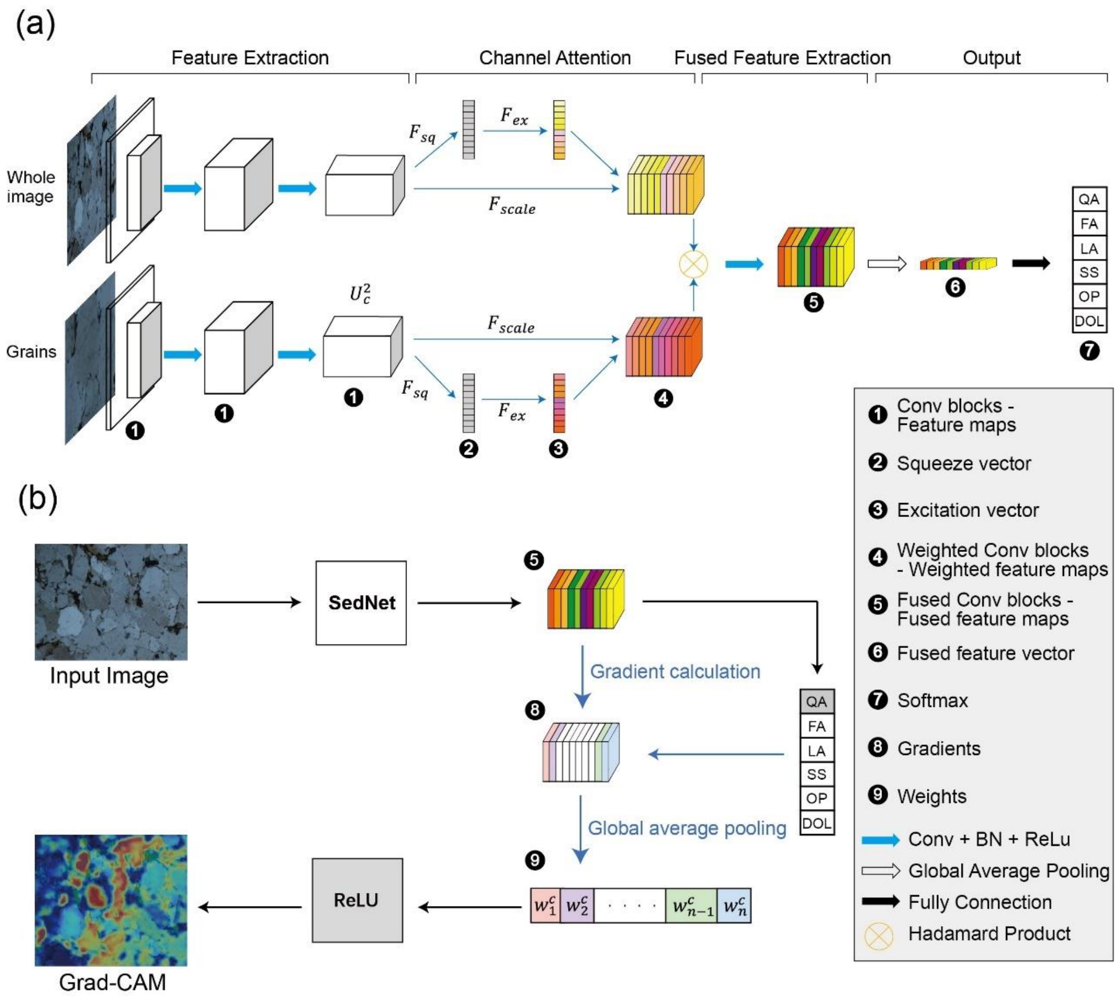

Figure 3. (a) Overall structure of the SedNet. (b) Workflow of the calculation of Grad-CAM, adapted from Selvaraju et al. (2017). The marked numbers indicate key components of SedNet; see text for explanations.

3.2 Grad-CAM visualization

Grad-CAM is a technique that visualizes the regions in an image that are most influential in the decision-making process of DCNNs (Selvaraju et al., 2017). In contrast to the traditional CAM algorithm, Grad-CAM offers enhanced performance due to its distinct approach to calculate class activation maps. Rather than employing the global average pooling over the feature map of the last convolution layer, as done in the CAM, Grad-CAM computes the gradient of the output class with respect to the final convolutional layer of the DCNN. This gradient information is then globally average-pooled to yield the neuron importance weights. These weights are crucial in highlighting which features in the map are most important for predicting the class, thereby resulting in a more effective and comprehensive visualization of the class activations. In this study, we calculated the gradient of the output of the DCNNs with respect to the feature maps of the last convolution layer (Gradient in Figure 3b). The target convolutional layer is M^k, and k denotes the k^{th} convolutional layer (Figure 3b). The weight of M^k can be calculated by Eq. (5):

$$w_k^c = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial y^c}{\partial M_{ij}^k}, \quad (5)$$

Where c represents a category, w_k^c is a vector representing the weight of each channel (Weights in Figure 3b) in M^k, y^c represents the score belonging to a certain category c. Then the Grad-CAM can be obtained through the weighted combination of forward activation and follow it with ReLU:

$$I_k^c = ReLU(\sum_{k=1}^{C} w_k^c \cdot M^k), \quad (5)$$

Where I_k^c represents the Grad-CAM of the target convolution layer.

4. Results and discussion

To thoroughly evaluate both the classical DCNNs and SedNet, we compared numerical metrics and Grad-CAM visualizations. To ensure a fair comparison, we used the same datasets and hyperparameters, such as training and test data, data augmentation, batch size, and learning rate, for all models. Additionally, to test the interpretability of DCNNs, we used scale bars for different rock types with various lengths and sizes. This allowed us to better understand how the models were making their classifications.

4.1 The classical models have high accuracies but tend to use irrelevant feature extractions

The three classical classification models, EfficientNet-B2 (Tan and Le, 2019), MobileNet-V3 (Howard et al., 2017), and ResNet-50 (He et al., 2016), achieved high accuracy and low loss values. ResNet-50 and MobileNet-V3 rapidly converged at approximately 16 epochs, while EfficientNet-B2 reached its plateau after approximately 30 epochs of training (Figure 4). The training accuracies for EfficientNet-B2, MobileNet-V3, and ResNet-50 are 0.9425, 0.9886, and 0.9871, while the test accuracies are 0.9187, 0.9458, and 0.9474 (Figure 4; Table 2).

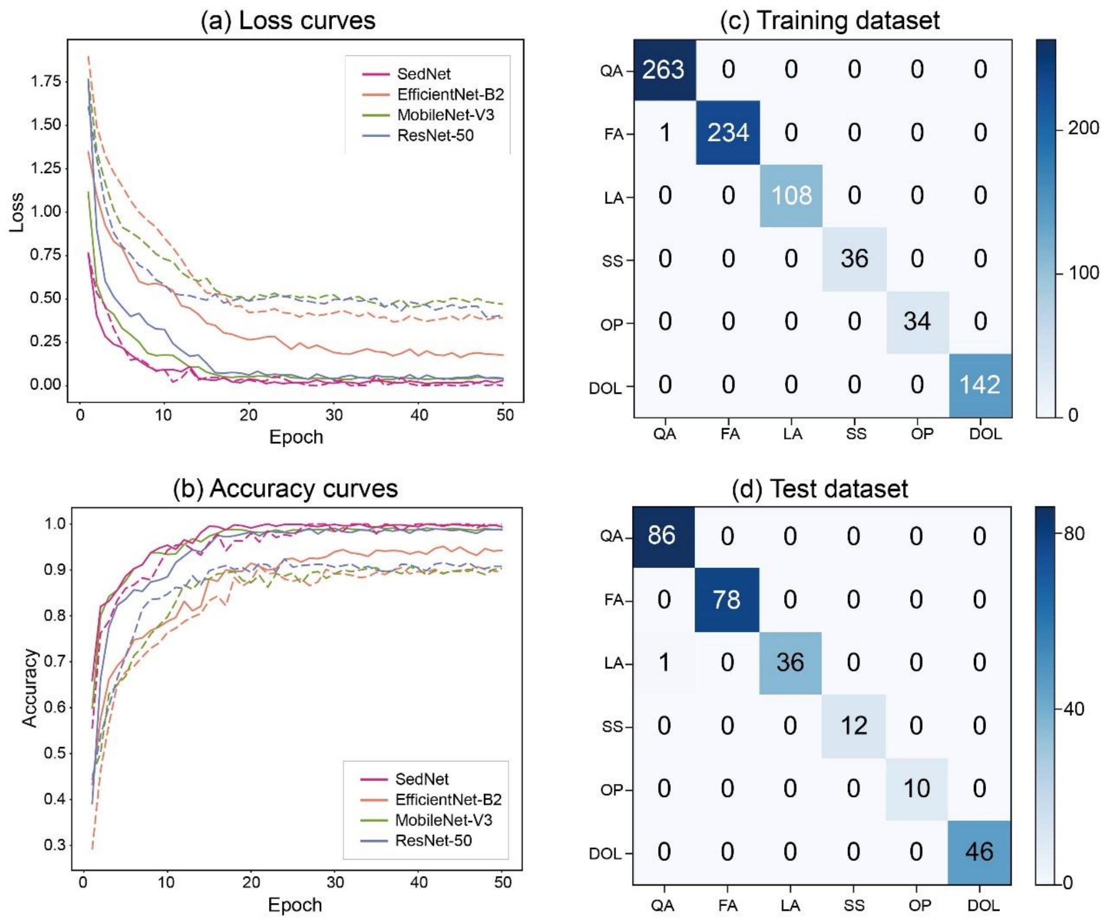

Figure 4. (a) and (b) show the loss and accuracy of the EfficientNet-B2, MobileNet-V3, ResNet-50, and SedNet. The results of training dataset are represented in solid lines, and results of validation dataset are represented in dashed lines. (c) and (d) are confusion matrices of SedNet using the training and test dataset, respectively. QA, quartz arenite; FA, feldspathic arenite; LA, lithic arenite; SS, siltstone; OP, oolitic packstone; DOL, dolomite.

Table 2. Evaluation metrics and loss of trained deep learning models

		SedNet	EfficientNet-B2	MobileNet-V3	ResNet-50
Accuracy	Train	***0.9988***	0.9360	0.9878	0.9866
	Test	***0.9963***	0.8993	0.9030	0.9104
	Mean	***0.9975***	0.9176	0.9454	0.9485
Precision	Train	***1.0000***	0.9651	0.9963	0.9926
	Test	***1.0000***	0.9414	0.9453	0.9457
	Mean	***1.0000***	0.9533	0.9708	0.9692
Recall	Train	***0.9988***	0.9688	0.9914	0.9938
	Test	***0.9963***	0.9526	0.9528	0.9606
	Mean	***0.9975***	0.9607	0.9721	0.9772
F1-score	Train	***0.9994***	0.9669	0.9938	0.9932
	Test	***0.9981***	0.9470	0.9490	0.9531
	Mean	***0.9988***	0.9569	0.9714	0.9732
Loss	Train	***0.0315***	0.1773	0.0453	0.0425
	Validation	***0.0011***	0.3918	0.4704	0.4058
	Mean	***0.0163***	0.2846	0.2579	0.2242

The Grad-CAM visualizations of the three models show uninterpretable features, suggesting that these models base their predictions on irrelevant features rather than the distinctive characteristics of sedimentary rocks (Figure 5). Although the characteristic features of the six rock types are distinct and easily recognizable by geologists, the classical DCNNs failed to focus on these features. For example, the characteristic features of quartz arenite, feldspathic arenite, lithic arenite and siltstone include the framework grains and their sandy or silty size, but the classical three models focus on other features such as the matrix and even scale bars. While the MobileNet-V3 model focuses on parts of the framework grains in feldspar arenite, the visualization map indicates that it focuses on the entire area rather than the distinctive feldspar grains.

Furthermore, the visualization map shows that the EfficientNet-B2 model identifies siltstone based on a few grains. But, since the investigated siltstones are heterogeneous in composition and grain size, the whole image or at least most of the siltstone photomicrograph should be considered. The classical models also perform poorly with carbonate rocks, where ooids, colorful interference colors, and euhedral dolostone crystal forms are important features for geologists. However, the DCNNs again tend to focus on the matrix, scale bar, and only a small number of grains (Figure 5). As the evaluation rules of the classical DCNNs are not interpretable, these models may not be able to make accurate predictions in real-world classification applications.

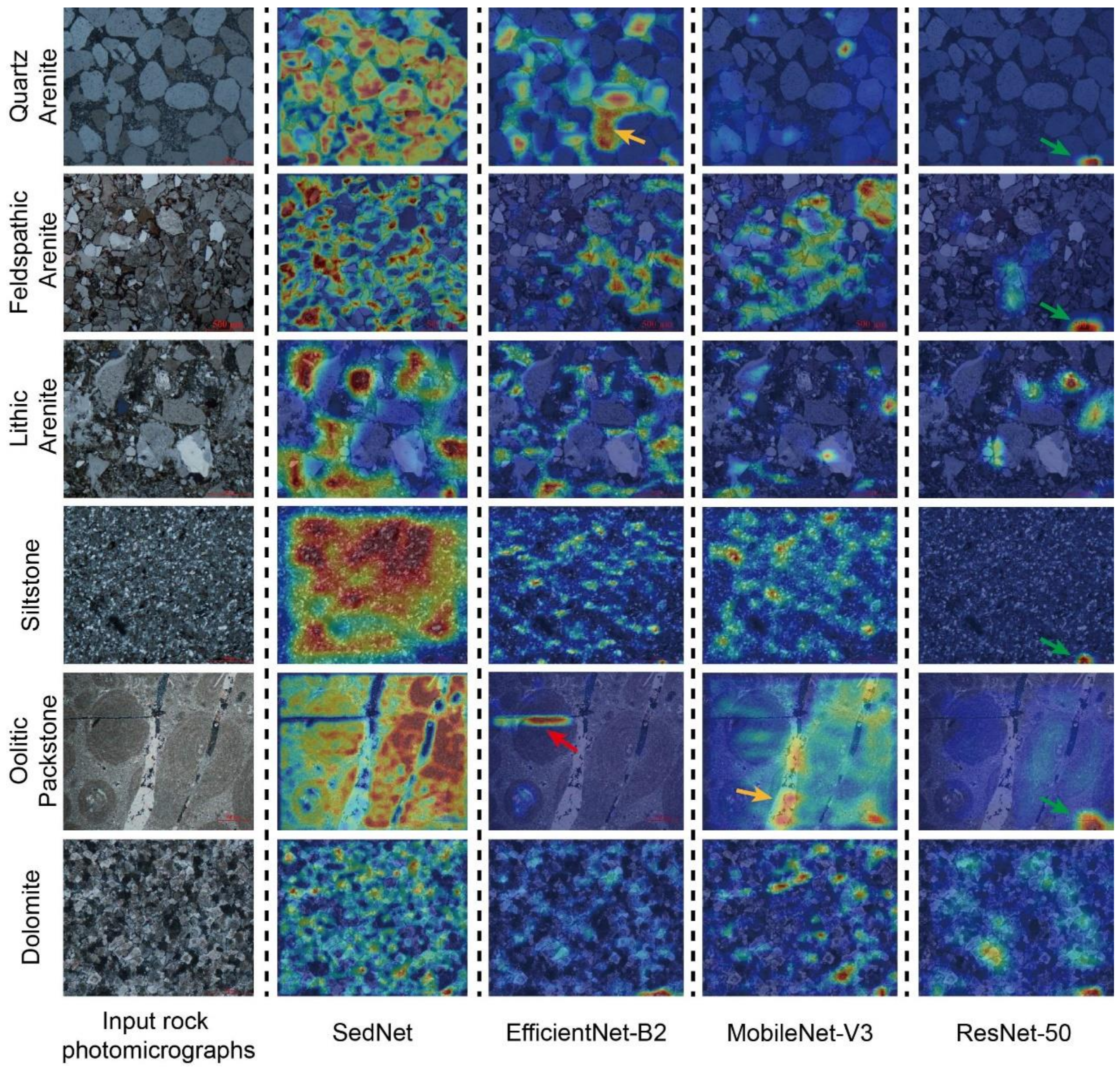

Figure 5. Grad-CAM of SedNet, EfficientNet-B2, MobileNet-V3, and ResNet-50. The red highlighted regions are the parts where models give more weight. Yellow arrows indicate the highlighted cements, red arrows indicate highlighted cracks, the green arrows indicate highlighted scale bars.

4.2 SedNet has not only high accuracy but also interpretable feature extractions

In comparison to the classical DCNNs, SedNet achieved not only high accuracy but also interpretable visualization maps. SedNet quickly converged at around 16 epochs, with training and test accuracies of 0.9942 and 1 (Figure 4; Table 2). The confusion matrix for SedNet shows that the model made very few mistakes, with only one feldspathic arenite being misclassified as a quartz arenite (Figure 4). Given the close similarity between quartz and feldspathic arenite with high quartz content, SedNet demonstrated excellent classification performance.

In addition to its accuracy, the Grad-CAM visualizations of SedNet are interpretable and align with geologists' knowledge (Figure 5). The visualization maps for quartz, feldspathic, and lithic arenite show that SedNet focuses on the most prominent quartz, feldspar, and lithic grains, indicating that the classification was based on these distinctive framework grains. For siltstone, the majority of the area in the photomicrographs is highlighted, consistent with the distinguishing feature of silty texture. Similarly, the visualization maps for the two carbonate rocks provide interpretable results with the ooids and dolostone in the photomicrographs being highlighted. SedNet used the characteristic framework grains and overall rock textures, rather than the matrix, scale bars, and cracks, to classify the rocks, which mimics the approach of geologists and therefore has a high potential for real rock classification projects.

4.3 Maximizing accuracy and decision-making power: the importance of interpretability and geological knowledge

In geosciences, accuracy is a crucial factor in developing and evaluating DL models. However, we argue that geological knowledge plays a key role in providing context and understanding for accurate predictions, and therefore its incorporation into DL models is necessary. Without this knowledge, even highly accurate models may not be reliable or meaningful. Most contemporary CNNs (e.g., ResNet, He et al., 2016) are characterized by millions of parameters, necessitating extensive image datasets for optimal performance.

Although these models are effective in numerous tasks, their suitability for specific domains necessitates additional validation. For example, in tasks like rock image identification, acquiring a large dataset is often unfeasible. Traditional CNNs tend to overfit by memorizing distinct features (e.g., cracks, as mentioned in this study) in smaller datasets, rather than identifying the features geologists typically use for rock classification. However, this overfitting is often undetected when evaluated solely through the numerical metrics. Therefore, the integration of geological knowledge into the development and evaluation of DL models for geoscience applications is essential to ensure accuracy, relevance, and interpretability (Barnes et al., 2020; McGovern et al., 2019; Ebert-Uphoff and Hilburn, 2020).

There is a trend in the geoscience community towards using machine learning models with a better understanding of their inner workings. For example, recent research (Zhao et al.,2019; Doucet et al., 2022; Zou et al., 2022) introduced feature importance and Shapley Additive Explanations (SHAP; Lundberg and Lee, 2017), an interpretable algorithm based on cooperative game theory, into geochemistry studies. Using SHAP values, trace elements in basalts can be used to classify tectonic settings and identify new geochemical differences between basalts from convergent and divergent boundaries. Toms et al. (2020) introduced layerwise relevance propagation for identifying meaningful patterns in ENSO (El Nino-Southern Oscillation) phase identification and seasonal prediction. This algorithm provides transparency to machine learning by propagating relevance from the output layer back through the network to the input layer based on the relative contribution of each neuron (Bach et al., 2015). These interpretable algorithms have shown great value in meteorology, being used to make subseasonal forecasts (Mayer and Barnes, 2021) and to reveal slowdowns in decadal climate warming (Labe and Barnes, 2022).

In our study, geologists can easily distinguish sedimentary rocks based on their unique framework grains and textures. However, classical convolutional neural networks (DCNNs) often placed more weight on scale bars, cements, and cracks, as evident in the Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations (Figure 5). The rock photomicrographs used in the study were intentionally presented with various styles of scale bars (Figure 1) and the studied quartz arenite was the only rock type with cracks. As a result, the scale bars and cracks became the most distinguishing features of the DCNNs. Such "noise" can be introduced easily if the rock samples are photographed by different institutes or in different facilities, and can significantly impact the final outputs. It has been noted that even slight image transformations can alter the predictions of DCNNs (Azulay and Weiss, 2018). One potential solution to this generalization issue is a collection of big datasets. However, it may not always be feasible for geologists to collect the same volume of images as datasets such as ImageNet. In these cases, the assessment of DCNN outputs cannot rely solely on numerical metrics like accuracy, as these only indicate the algorithm’s performance on a known dataset. Without a full understanding of how the algorithm works, trained models may still face generalization issues. Our proposed dual-modal network is a potential solution for image classification tasks. It emphasizes distinguishing features in the DL models and achieves high accuracy and interpretable Grad-CAM visualizations by integrating global and local features (Figure 5).

The proposed dual-modal network represents one way in enhancing the performance and interpretability of DCNNs, achieved by integrating domain-specific modules. In addition to the integration of bespoke modules into the network, alternative strategies also show potential in this respect. A major challenge in applying deep learning to geosciences is the lack of labeled training data, often due to subjective or labor-intensive labeling processes. One solution is to generate synthetic datasets based on fundamental geological principles. This not only helps in tuning model parameters but also ensures model robustness when applied to real-world problems. This technique has been used successfully in various applications, such as detecting permeability from rock images by generating samples of porous media and assigning permeability labels using the Boltzmann method (Wu et al., 2018), and seismic interpretation where large datasets can be generated through forward modelling (Wu et al., 2023).

Another approach is to integrate prior knowledge into the loss function. This requires DCNNs to conform to the training dataset while also adhering to the prior knowledge, such as physical laws defined by partial differential equations, and this type of neural network is also known as the physics-informed neural network. The powerful approximation and high expressivity capacities of DCNNs enable them to infer solutions within the complex space defined by the governing physical laws, thereby enhancing model performance and understanding (Cuomo et al., 2022; Zhang et al., 2023). In the realm of rock imaging, the prior knowledge can be the experience of geologists that some rocks or grains are more identifiable due to their distinctive features. By applying different weights to the regularization terms in the loss function, the model's ability to analyze difficult images can be improved. Given the desire for a physical understanding of DL models in geosciences, it is expected that interpretability and geological knowledge will be further incorporated in future applications of DCNNs. Combining interpretability and geological knowledge with Deep Learning can create more effective and transparent models for geoscience applications.

5. Conclusions

To examine the significance of interpretability in DL models for geoscientific tasks, we conducted automatic sedimentary rock classification using thin-section photomicrographs and DL models. While classical deep convolutional neural networks (DCNNs) such as ResNet, EfficientNet, and MobileNet achieved high accuracy (up to 0.96), their Grad-CAM visualizations were often not geological-reasonable. Framework grains and textures are key features for distinguishing the studied sedimentary rocks. However, classical DCNNs tended to classify rocks based on irrelevant features, as indicated by the highlighted regions of scale bars, cements, and cracks. To address this issue, we proposed an attention-based dual network that inputs both the original thin-section photomicrographs and framework grains. By combining information from the whole images and framework grains, our proposed model achieves not only higher accuracy (0.99) but also produces interpretable visualization heatmaps in which framework grains were given more weight in the classification process. Our study highlights the importance of considering interpretability and geological knowledge in developing DL models, in addition to aiming for high accuracy.

Acknowledgement

The authors would like to thank Tingting Gong for providing thin sections and Jiaqi Wu for helping in capturing photomicrographs. Thanks to three anonymous reviewers for constructive reviews that substantially improved the manuscript. This work was financially supported by National Natural Science Foundation of China (Grant Nos. 41888101, 42050104, 42050102, 42202125, 42172137). Additional support was provided by the IUGS Deep-time Digital Earth (DDE) Big Science Program.

Code availability statement

The code is made open on Github repository at:

https://github.com/MudRocw1/SedNet_explainable-deep-learning-network

References

Azulay, A., Weiss, Y., 2018. Why do deep convolutional networks generalize so poorly to small image transformations? arXiv Prepr. arXiv1805.12177.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10, e0130140.

Barnes, E.A., Toms, B., Hurrell, J.W., Ebert-Uphoff, I., Anderson, C., Anderson, D., 2020. Indicator Patterns of Forced Change Learned by an Artificial Neural Network. J. Adv. Model. Earth Syst. 12. https://doi.org/10.1029/2020MS002195

Bergen, Karianne J., Paul A. Johnson, V. Maarten, Beroza., G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science (80-.). 363. https://doi.org/10.1126/science.aau0323

Boggs Jr, S., Boggs, S., 2009. Petrology of sedimentary rocks. Cambridge university press.

Buhrmester, V., Münch, D., Arens, M., 2021. Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. Mach. Learn. Knowl. Extr. 3, 966–989. https://doi.org/10.3390/make3040048

Camps-Valls, G., Tuia, D., Zhu, X.X., Reichstein, M., 2021. Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences. John Wiley & Sons.

Castelvecchi, D., 2016. The black box of AI. Nature 538, 20–23.

Cuomo, S., Di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F., 2022. Scientific machine learning through physics–informed neural networks: Where we are and what's

next. J. Sci. Comput. 92, 88.
de Lima, R.P., Bonar, A., Duarte Coronado, D., Marfurt, K., Nicholson, C., 2019. Deep convolutional neural networks as a geological image classification tool. Sediment. Rec. 17, 4–9. https://doi.org/10.2110/sedred.2019.2.4
Dickinson, W.R., Suczek, C.A., 1979. Plate tectonics and sandstone compositions. Am. Assoc. Pet. Geol. Bull. 63, 2164–2182.
Doucet, L.S., Tetley, M.G., Li, Z.-X., Liu, Y., Gamaleldien, H., 2022. Geochemical fingerprinting of continental and oceanic basalts: A machine learning approach. Earth-Science Rev. 104192.
Ebert-Uphoff, I., Hilburn, K., 2020. Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. Bull. Am. Meteorol. Soc. 101, E2149–E2170. https://doi.org/10.1175/BAMS-D-20-0097.1
Garzanti, E., Doglioni, C., Vezzoli, G., Ando, S., 2007. Orogenic belts and orogenic sediment provenance. J. Geol. 115, 315–334.
Ge, S., Wang, C., Jiang, Z., Hao, H., Gu, Q., 2021. Dual-input attention network for automatic identification of detritus from river sands. Comput. Geosci. 151, 104735. https://doi.org/10.1016/j.cageo.2021.104735
Hao, H., Guo, R., Gu, Q., Hu, X., 2019. Machine learning application to automatically classify heavy minerals in river sand by using SEM/EDS data. Miner. Eng. 143, 105899. https://doi.org/10.1016/j.mineng.2019.105899
He, Kaiming, Xiangyu Zhang, Shaoqing Ren, Sun, J., 2016. Deep residual learning for image recognition., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv Prepr. arXiv1704.04861.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.
Koeshidayatullah, A., Morsilli, M., Lehrmann, D.J., Al-Ramadan, K., Payne, J.L., 2020. Fully automated carbonate petrography using deep convolutional neural networks. Mar. Pet. Geol. 122, 104687. https://doi.org/10.1016/j.marpetgeo.2020.104687
Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.
Labe, Z.M., Barnes, E.A., 2022. Predicting Slowdowns in Decadal Climate Warming Trends With Explainable Neural Networks. Geophys. Res. Lett. 49. https://doi.org/10.1029/2022GL098173
Lei, S., Zhang, H., Wang, K., Su, Z., 2018. How training data affect the accuracy and robustness of neural networks for image classification.
Li, D., Zhao, J., Ma, J., 2022. Experimental Studies on Rock Thin-Section Image Classification by Deep Learning-Based Approaches.
Liu, T., Li, C., Liu, Zongbao, Zhang, K., Liu, F., Li, D., Zhang, Y., Liu, Zhigang, Liu, L., Huang, J., 2022. Research on Image Identification Method of Rock Thin Slices in Tight Oil Reservoirs Based on Mask R-CNN. Energies 15. https://doi.org/10.3390/en15165818
Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y., Song, H., 2022.

Automatic taxonomic identification based on the Fossil Image Dataset (> 415,000 images) and deep convolutional neural networks. Paleobiology 1–22.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Maitre, J., Bouchard, K., Bédard, L.P., 2019. Mineral grains recognition using computer vision and machine learning. Comput. Geosci. 130, 84–93. https://doi.org/10.1016/j.cageo.2019.05.009

Mamalakis, A., Ebert-Uphoff, I., Barnes, E.A., 2022. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. Environ. Data Sci. 1, 1–17. https://doi.org/10.1017/eds.2022.7

Mayer, K.J., Barnes, E.A., 2021. Subseasonal Forecasts of Opportunity Identified by an Explainable Neural Network. Geophys. Res. Lett. 48, 1–9. https://doi.org/10.1029/2020GL092092

McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., Smith, T., 2019. Making the black box more transparent: Understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100, 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1

Romero, I.C., Kong, S., Fowlkes, C.C., Jaramillo, C., Urban, M.A., Oboh-Ikuenobe, F., D'Apolito, C., Punyasena, S.W., 2020. Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. Proc. Natl. Acad. Sci. U. S. A. 117, 28496–28505. https://doi.org/10.1073/pnas.2007324117

Saxena, N., Day-Stirrat, R.J., Hows, A., Hofmann, R., 2021. Application of deep learning for semantic segmentation of sandstone thin sections. Comput. Geosci. 152, 104778. https://doi.org/10.1016/j.cageo.2021.104778

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Su, C., Xu, S., Zhu, K., Zhang, X., 2020. Rock classification in petrographic thin section images based on concatenated convolutional neural networks. Earth Sci. Informatics 13, 1477–1484.

Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., Gao, Y., 2018. Is Robustness the Cost of Accuracy?--A Comprehensive Study on the Robustness of 18 Deep Image Classification Models, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tang, D.G., Milliken, K.L., Spikes, K.T., 2020. Machine learning for point counting and segmentation of arenite in thin section. Mar. Pet. Geol. 120. https://doi.org/10.1016/j.marpetgeo.2020.104518

Toms, B.A., Barnes, E.A., Ebert-Uphoff, I., 2020. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. J. Adv. Model. Earth Syst. 12,

1–20. https://doi.org/10.1029/2019MS002002
Wang, C., Ge, S., Jiang, Z., Hao, H., Gu, Q., 2021. Computers and Geosciences SiamFuseNet : A pseudo-siamese network for detritus detection from polarized microscopic images of river sands. Comput. Geosci. 156, 104912. https://doi.org/10.1016/j.cageo.2021.104912
Wu, J., Yin, X., Xiao, H., 2018. Seeing permeability from images: fast prediction with convolutional neural networks. Sci. Bull. 63, 1215–1222.
Wu, X., Ma, J., Si, X., Bi, Z., Yang, J., Gao, H., Xie, D., Guo, Z., Zhang, J., 2023. Sensing prior constraints in deep neural networks for solving exploration geophysical problems. Proc. Natl. Acad. Sci. U. S. A. 120, 1–12. https://doi.org/10.1073/pnas.2219573120
Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R.R., Chaudhuri, K., 2020. A closer look at accuracy vs. robustness. Adv. Neural Inf. Process. Syst. 33, 8588–8601.
Zhang, Y., Zhu, X., Gao, J., 2023. Seismic inversion based on acoustic wave equations using physics-informed neural network. IEEE Trans. Geosci. Remote Sens. 61, 1–11.
Zhao, Y., Zhang, Y., Geng, M., Jiang, J., Zou, X., 2019. Involvement of slab-derived fluid in the generation of Cenozoic basalts in Northeast China inferred from machine learning. Geophys. Res. Lett. 46, 5234–5242.
Zheng, D., Wu, S., Ma, C., Xiang, L., Hou, L., Chen, A., Hou, M., 2022. Zircon classification from cathodoluminescence images using deep learning. Geosci. Front. 101436. https://doi.org/10.1016/j.gsf.2022.101436
Zou, S., Chen, X., Brzozowski, M.J., Leng, C., Xu, D., 2022. Application of machine learning to characterizing magma fertility in porphyry Cu deposits. J. Geophys. Res. Solid Earth 127, e2022JB024584.

List of Figures

Figure 1. The studied six types of sedimentary rocks and the associated scale bars.

Figure 2. Numbers of training, validation, and test datasets. QA, quartz arenite; FA, feldspathic arenite; LA, lithic arenite; SS, siltstone; OP, oolitic packstone; DOL, dolomite.

Figure 3. (a) Overall structure of the SedNet. (b) Workflow of the calculation of Grad-CAM, adapted from Selvaraju et al. (2017). The marked numbers indicate key components of SedNet, see text for explanations.

Figure 4. (a) and (b) show the loss and accuracy of the EfficientNet-B2, MobileNet-V3, ResNet-50, and SedNet. (c) and (d) are confusion matrices of SedNet using the training and test dataset, respectively. QA, quartz arenite; FA, feldspathic arenite; LA, lithic arenite; SS, siltstone; OP, oolitic packstone; DOL, dolomite.

Figure 5. Grad-CAM of the SedNet, EfficientNet-B2, MobileNet-V3, and ResNet-50. The

red highlighted regions are the parts where models give more weight. Yellow arrows indicate the highlighted cements, red arrows indicate highlighted cracks, the green arrows indicate highlighted scale bars.

List of Tables