frontiers | Frontiers in Earth Science

TYPE Original Research
PUBLISHED 04 October 2022
DOI 10.3389/feart.2022.992442

OPEN ACCESS

EDITED BY
Ataollah Shirzadi,
University of Kurdistan, Iran

REVIEWED BY
Umar Ashraf,
Yunnan University, China
Golale Asghari,
University of Kurdistan, Iran

*CORRESPONDENCE
Ardiansyah Koeshidayatullah,
koeshidayatullah@kfupm.edu.sa
Motaz Alfarraj,
motaz@kfupm.edu.sa

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing,
a section of the journal
Frontiers in Earth Science

RECEIVED 12 July 2022
ACCEPTED 06 September 2022
PUBLISHED 04 October 2022

CITATION
Koeshidayatullah A, Al-Azani S, Baraboshkin EE and Alfarraj M (2022), FaciesViT: Vision transformer for an improved core lithofacies prediction. *Front. Earth Sci.* 10:992442.
doi: 10.3389/feart.2022.992442

FaciesViT: Vision transformer for an improved core lithofacies prediction

Ardiansyah Koeshidayatullah[1*†], Sadam Al-Azani[2†], Evgeny E. Baraboshkin[3,4] and Motaz Alfarraj[2*]

[1]Department of Geosciences, College of Petroleum Engineering and Geosciences, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, [2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI), King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, [3]Skolkovo Institute of Science and Technology, Moscow, Russia, [4]Digital Petroleum LLC, Moscow, Russia

Lithofacies classification is a fundamental step to perform depositional and reservoir characterizations in the subsurface. However, such a classification is often hindered by limited data availability and biased and time-consuming analysis. Recent work has demonstrated the potential of image-based supervised deep learning analysis, specifically convolutional neural networks (CNN), to optimize lithofacies classification and interpretation using core images. While most works have used transfer learning to overcome limited datasets and simultaneously yield a high-accuracy prediction. This method raises some serious concerns regarding how the CNN model learns and makes a prediction as the model was originally trained with entirely different datasets. Here, we proposed an alternative approach by adopting a vision transformer model, known as *FaciesViT*, to mitigate this issue and provide improved lithofacies prediction. We also experimented with various CNN architectures as the baseline models and two different datasets to compare and evaluate the performance of our proposed model. The experimental results show that the proposed models significantly outperform the established CNN architecture models for both datasets and in all cases, achieving an f1 score and weighted average in all tested metrics of 95%. For the first time, this study highlights the application of the Vision Transformer model to a geological dataset. Our findings show that the FaciesViT model has several advantages over conventional CNN models, including (i) no hyperparameter fine-tuning and exhaustive data augmentation required to match the accuracy of CNN models; (ii) it can work with limited datasets; and (iii) it can better generalize the classification to a new, unseen dataset. Our study shows that the application of the Vision transformer could further optimize image recognition and classification in the geosciences and mitigate some of the issues related to the generalizability and the explainability of deep learning models. Furthermore, the implementation of our proposed *FaciesViT* model has been shown to improve the overall performance and reproducibility of image-based core lithofacies classification which is significant for subsurface reservoir characterization in different basins worldwide.

KEYWORDS

lithofacies, geosciences, transformers, AI, deep learning

1 Introduction

Since the seminal work of Krizhevsky et al. (2012) that used deep convolutional neural networks (CNN), known as AlexNet, to win the ImageNet challenge, the application of CNNs has been widely popular and has significantly transformed the landscape of visual and pattern recognition. Availability of big data and easy access to high-performance computing resources allows rapid development of deeper and wider CNN architectures to learn more complex features in an image and produce a prediction accuracy that surpasses human accuracy (LeCun et al., 2015; He et al., 2016). CNNs are particularly powerful for image analysis because CNNs are translational and scale invariance through weight sharing and pooling, respectively. In geosciences, various CNN algorithms have been explored and implemented to perform either supervised or unsupervised learning approach on multi-scale image datasets and analyzes, including basin-scale seismic interpretation (Wrona et al., 2018; Alaudah et al., 2019); Wu et al., 2019 and micro-scale analysis of petrography or computed tomography scan datasets (Alqahtani et al., 2018; Koeshidayatullah et al., 2020; Ferreira et al., 2022).

Previous works have highlighted CNNs as a data-hungry model, and to achieve a high-performance result, a much deeper network is required to increase the receptive fields and capture long-range dependencies (LeCun et al., 2015; He et al., 2016). This issue can be mitigated by the ability of CNNs to be trained in a domain where a large volume of datasets is obtained more easily and transfer the knowledge to a more specific domain where data is difficult to collect and expensive (Weiss et al., 2016). This method is referred to as transfer learning, and it provides a way to optimize machine learning performance when the dataset is limited, such as in geosciences. Since then, most deep learning implementations in geosciences, particularly for image analysis tasks, have relied primarily on transfer learning methods (Li et al., 2017; de Lima et al., 2019; Baraboshkin et al., 2020; Wu et al., 2020). Although this method allows for some breakthroughs in geological image classification and recognition, the model was originally trained on a domain that inherently has different data features and distributions, but can still produce a high-performance result that could raise some concerns in the long run (Pires de Lima and Duarte, 2021; Koeshidayatullah, 2022). Furthermore, this is compounded by the relatively stagnant performance and low explainability of various CNN models, which created the urgency to develop a deeper and wider CNN model. In such a case, while the performance of CNN models may improve, it comes with a significant trade-off, in which the models become computationally expensive and even more challenging to interpret.

Motivated by the success of the self-attention mechanism in natural language processing tasks (Vaswani et al., 2017; Devlin et al., 2018), the recent development of CNN models for vision tasks has also attempted to incorporate such self-attention modules within CNN layers and has been shown to improve the overall image classification performance (Cao et al., 2020). The implementation of attention modules allows the model to focus on specific areas of the input image that has more influence on the prediction. However, one major drawback of using self-attention for the image data set is the quadratic complexity of the sequence length, because the image pixel needs to be unrolled into long 1D sequences, and each pixel needs to attend to all other pixels. These issues make transformer models computationally expensive and inefficient for image analysis. Furthermore, the transformer model requires positional embedding to capture the correct pixel sequence, which requires architectural changes. Recent work proposed a new architecture, Vision Transformer (ViT), to mitigate these issues by splitting the image into 16x16-sized patches and treating these patches as tokens (Dosovitskiy et al., 2020). Furthermore, this model encodes positional embedding and class embedding to capture the spatial relationship between patches and learn the global relationship of the image, respectively. This model achieved state-of-the-art results on the ImageNet dataset, which inspires further works to develop ViT architectures for various computer vision tasks (Chen C.-F. et al., 2021; Liu et al., 2021). Despite these successful applications, the ViT model is well-known for requiring a large number of training datasets. Recent work has focused its efforts on improving the ViT model in order to make the model work with a small dataset.

In geosciences, most image classification and segmentation tasks are dominated by CNNs, and the application of ViT is still significantly limited. To date, only a few studies have experimented with ViT, primarily for remote sensing images (Bazi et al., 2021; Chen H. et al., 2021), and no studies have utilized the Vision Transformer model to perform lithofacies classification on borehole core images. Therefore, our study aims to explore how ViT learns and makes predictions on a geological dataset. Furthermore, we evaluate and validate (i) the performance of ViT to classify the lithofacies of subsurface core images; (ii) the ability of ViT to generalize the learning process and predict new, unseen datasets, and further compare the prediction with conventional CNNs architecture (base and pretrained models); and (iii) the limitation and potential of ViT for classifying limited geological dataset. The successful application of ViT has the potential to improve the generalizability and explainability of deep learning models in geosciences.

2 Core-based lithofacies analysis

The characterization of the subsurface reservoir involves multi-scale and -dimensional analyzes, from one-dimensional

borehole analysis (well log and core samples) to three-dimensional depositional and property modeling (Amel et al., 2015; Al-Ramadan et al., 2020; Anees et al., 2022b; Anees et al., 2022a). In such cases, the classification and interpretation of lithofacies play a key role in providing the basic building block to a more advanced interpretation. Lithofacies (facies)—is defined as a body of rock with certain specified attributes that distinguish it from other rock units (Leeder, 2012). In reservoir characterization, lithofacies classification is particularly important for identifying depositional processes and delineating the depositional environment and quantifying the net-to-gross ratio of reservoir facies (Amel et al., 2015; Koeshidayatull ah et al., 2016; Ashraf et al., 2019). Information related to lithofacies can be obtained from well data (e.g., Gamma-Ray, Neutron-Density), seismic attribute analysis, and cored intervals in a borehole. The latter is the only dataset that provides the ground truth of lithofacies classification and interpretation. However, it is also the most expensive and difficult to obtain and characterize from the subsurface; hence, its availability is relatively limited.

Conventional lithofacies classification from core samples relies on visual pattern analysis (e.g., grain size, depositional texture, and structure analysis) (Rothwell and Rack, 2006). Secondary physical and chemical analyzes, such as hardness, acid tests and other nondestructive geochemical techniques, can also be performed to confirm the mineralogy of the lithofacies (Croudace and Rothwell, 2015; McPhee et al., 2015; Amao et al., 2016). Although physical analysis cannot be analyzed in machine learning, machine learning is an excellent tool for replicating visual pattern analysis and performing more robust and unbiased classification (Thomas et al., 2011; Baraboshkin et al., 2018; Martin et al., 2021). Visual-based machine learning, specifically deep neural networks, has been widely applied in the past 5 years to perform lithological classification (de Lima et al., 2019; Baraboshkin et al., 2020; Koeshidayatullah et al., 2020). The primary motivation behind employing machine learning for this task is a time-consuming, expensive, and potentially biased interpretation when performed by a human. To date, it is clear that machine learning will never completely replace humans or geologists in performing this task, but machine learning has the potential to help optimize and standardize the process. Such advancement will not only significantly reduce the cost associated with core description but also improve the reproducibility of core analysis across geologists.

Various models and approaches have been proposed to conduct core-based lithofacies analysis, from support vector machine (SVM) to deep neural networks. A set of works that aims to classify rocks by different types of features. A set of works extracted different color distributions and intensity (e.g., color histogram, hue, saturation) from rock samples and used different algorithms based on statistics (Singh et al., 2004; Harinie et al., 2012), SVM (Lobos et al., 2016; Patel et al., 2016; Patel et al., 2017), combination of statistics and machine learning (Prince et al., 2005; Thomas et al., 2011; Seleznev et al., 2020), to perform lithology classification. In addition, previous works applied LeNet (named CIFAR in the publication) and other convolutional neural networks to classify granite tiles (Ferreira and Giraldi, 2017) and rock images (Zhang et al., 2017). Another work shows the application of deep convolutional neural networks (CNN) to classify different lithologies directly from core images (de Lima et al., 2019). Despite highlighting several limitations of deep learning, this study shows a promising potential of CNN in optimizing core-based lithofacies analysis only from digital core images. Several follow-up studies demonstrated the power of CNN to perform core lithofacies classification by comparing different architectures and found that ResNet architecture (He et al., 2016) outperforms other CNN architectures, achieving up to 95% in analytical precision (Baraboshkin et al., 2018; Ivchenko et al., 2018; Baraboshkin et al., 2020). Recent work by Martin et al., 2021 shows a successful case of coupling core facies and extracted color-channel log(CCL) to predict centimeter-scale lithofacies. This study uses two different CNN, WaveNet (Oord et al., 2016) and Deep TEN (Zhang et al., 2017), to perform sequence-to-sequence learning in the CCL data and texture classification in the core image dataset, respectively. The CNN currently applied for different rock types and analyses: igneous rocks (Fan et al., 2020; Fu et al., 2022), rock quality designation (Alzubaidi et al., 2021), and trace fossils detection (Ayranci et al., 2021; Timmer et al., 2021). Furthermore, a recent work proposed the use of elemental data in addition to images to improve the accuracy of classification (Xu et al., 2021).

A previous study explored how networks learn by extracting different feature maps at different layers, indicating that networks learn rather differently than humans and, most of the time, use unrelated features to predict lithofacies (Baraboshkin et al., 2020). Furthermore, this study shows that the model could not achieve the same level of performance in a new unseen core dataset with similar lithofacies, so the accuracy dropped to close to 50%. This raises a serious concern about the explainability and generalizability of CNN models and how much we can trust the model. While another study highlighted how probability averaging may improve the results of classification (Alzubaidi et al., 2021), the class imbalance problem, which is very typical in the geosciences dataset, could have a detrimental impact on the overall performance of the deep learning model (Koeshidayatullah et al., 2020; Koeshidayatullah, 2022). This is further compounded by the fact that most of these studies rely heavily on the transfer learning method and intensive data augmentation to perform training (Baraboshkin et al., 2020). Therefore, there is an urgent need to explore another deep learning method, such as Vision Transformer, to analyze geological image datasets and improve the explainability of deep learning.

TABLE 1 Datasets description.

Class	Dataset I			Dataset II
	Training	Testing	All samples	Testing
Argillaceous	517	104	621	37
Granite	550	111	661	0
limestone	263	53	316	0
Sandstone-Laminated	534	108	642	217
Sandstone-Massive	561	114	675	60
Siltstone	542	110	652	126
Total	2967	600	3567	440

3 Methodology

This section presents the methodology designed and followed to evaluate ResNet, ViT, and the hybrid structures of ResNet and ViT to classify Core Lithofacies.

3.1 Dataset collection and preparation

Table 1 describes the datasets used in this study collected by Baraboshkin et al. (2020). 10% of the training set are used for validation. The first dataset was collected from different wells placed in Russia, it includes various formations: Bazhenov, Abalak (Vasuganskaya and Georgievskaya), Vikulovskaya, and Domanik. The second dataset is collected from the Achimov formation. The dataset was collected as a photo of core boxes from the RFGF (unified fund of subsurface geological information) website and automatically cropped out from the photo to separate different core box columns. The columns were then sliced into 10 x 10 cm images. Each image is resized to 256 × 256 pixels and normalized to a range of pixel intensities from 0 to 1.

3.2 Augmentation

The augmentation technique is a powerful tool for increasing the generalization performance of models by minimizing overfitting. It can be carried out online or offline. Online augmentation, also known as real-time augmentation, is performed during the training phase, whereas offline augmentation is first implemented on the dataset before training, then the augmented dataset is used for training the model. It is noteworthy that both online and offline augmentations have their benefits and drawbacks. For example, the main disadvantages of offline augmentation include space complexity and implementation complexity (Shorten and Khoshgoftaar, 2019). On the other hand, a main drawbacks of online augmentation is that the original samples of the dataset might not be kept during the training. Another issue with online augmentation compared with offline augmentation is that the size of the dataset is as the same as the size of the original dataset.

In this study, online augmentation is taken into account and implemented on the input dataset using TorchVision Transforms in Pytorch. The considered image augmentation includes cropping, rotation, flipping, color jitter, and Gaussian blur. They are implemented by random selection. To alleviate the aforementioned drawbacks of online data augmentation, we created three data loaders and combined them. The first is for the original data (to ensure the original samples are included in the training) and the other ones for the considered augmentation types with different orders and parameters.

- Random crop: Each input image is first resized to 256 × 256 pixels. Then random subset with 224 × 224 pixels from the original image is created.
- Rotation: Images are rotated at 90 and 180° for the first and the second image augmenters.
- Flipping: horizontal and vertical flips are considered with a probability of 0.5 for the first and second image augmenters.
- Color jitter: this type of augmentation is to randomly change the contrast, brightness, saturation, and hue of an image. The brightness factor of 0.5 is considered to add a random brightness jitter to images for one data loader. However, the brightness for the other data augmenter is not changed. The saturation factor is chosen uniformly from 0 to one to adjust the amount of jitter in saturation. Hue factor of 0.3 is also adopted to add random shade to the images.
- Blur: using Gaussian filter to blur an image. For the first data augmenter the kernel width and height size of 21 is adapted, whereas a kernel size of five is used for the second image augmenter.

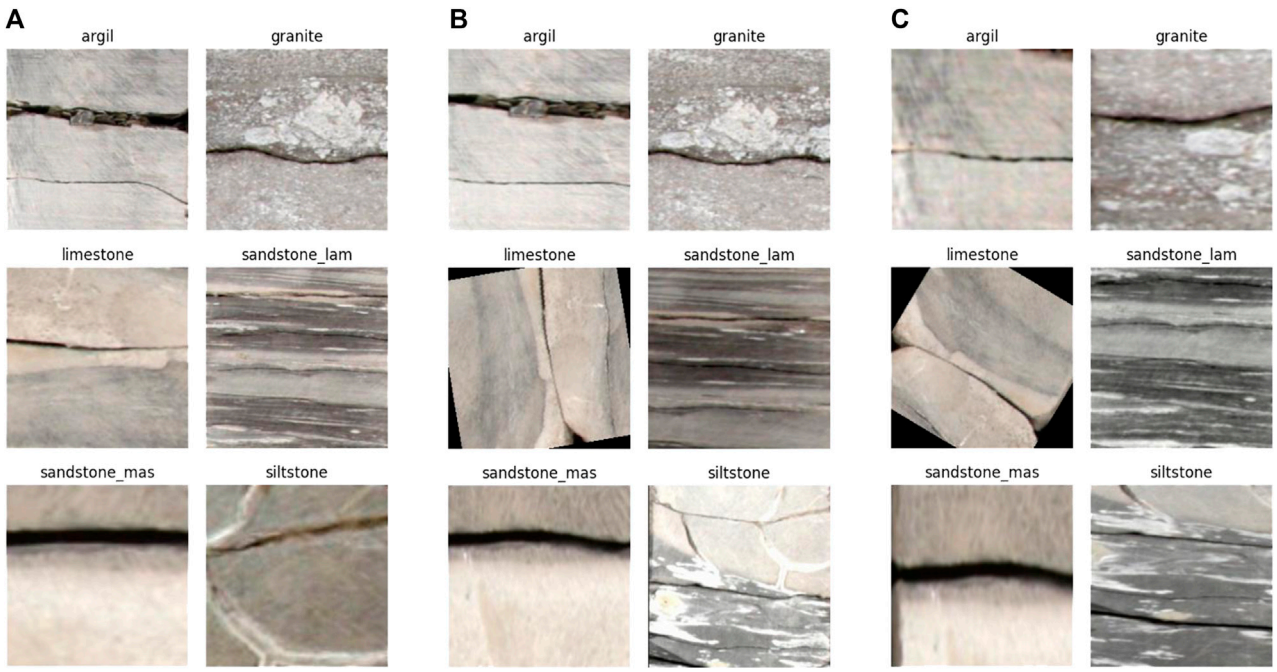

FIGURE 1
Examples of **(A)** original data, **(B)** image augmenter I and **(C)** image augmenter II.

Figure 1 depicts examples of the developed data loaders.

3.3 Training and fine-tuning

Optimization is the process of adjusting model parameters to reduce model error in each training step. Stochastic gradient descent (SGD) optimizer is used with Momentum (Qian, 1999) value of 0.9 in order to help accelerate SGD in the relevant direction and dampens oscillations.

The learning rate warm-up (Goyal et al., 2017) approach is adopted at the beginning of the training steps by which the learning rate increases linearly from zero to the initial learning rate during the defined steps. Then, for the remaining steps the learning rate goes down from the initial learning rate to zero following cosine function/curve.

Dropout as a regularization technique is a significant method to prevent overfitting. It deactivates randomly chosen neurons during training phase by assigning zero values for the selected neurons. Dropout cannot only be applied for the neuron level but also on the path level (Drop-path) in order to prevent co-adapting different depths of sub-networks.

Weight decay is also implemented to help prevent overfitting and the exploding gradient problem issues by adding a penalty term to the cost function. In this study, L2 penalty is applied which leads to shrink the model weights. Cross-entropy loss is applied to adjust the weights during the training phase to minimize the loss value.

$$Loss = -\sum_{i=1}^{n} t_i log\left(p_i\right) \tag{1}$$

where n is the number of classes, in this study $n = 6$, t_i is the truth label and p_i is the Softmax probability for class i.

The transfer learning paradigm is utilized here for both CNN-based and ViT-based architectures due to the small size of the training dataset. Networks were pre-trained using ImageNet Dataset (Deng et al., 2009) and the generated weights are fine-tuned and the networks were re-trained for the classification task of this study.

3.4 Evaluation measures

To evaluate the proposed models, the confusion matrix, precision, precision, recal, and F_1 are considered. A confusion matrix is an $N \times N$ table where N is the number of classes in the dataset. It is a good method to evaluate the performance of classification models especially, for highly imbalanced datasets. It is composed of four main evaluation measures namely: True Positive (TF), True Negative (TN), False Positive (FP) and False Negative (FN). Several different evaluation measures can be formulated from those four measures, including Accuracy, Precision, Recall and F_1.

TABLE 2 Description and variations of ViT (Dosovitskiy et al., 2020) models.

	Layers	Hidden size dim	Heads	MLP dim	Params
ViT-Base (ViT-B)	12	768	12	3027	~86M
ViT-Large (Vit-L)	24	1024	16	4096	~307M
ViT-Huge (ViT-H)	32	1280	16	5120	~632M

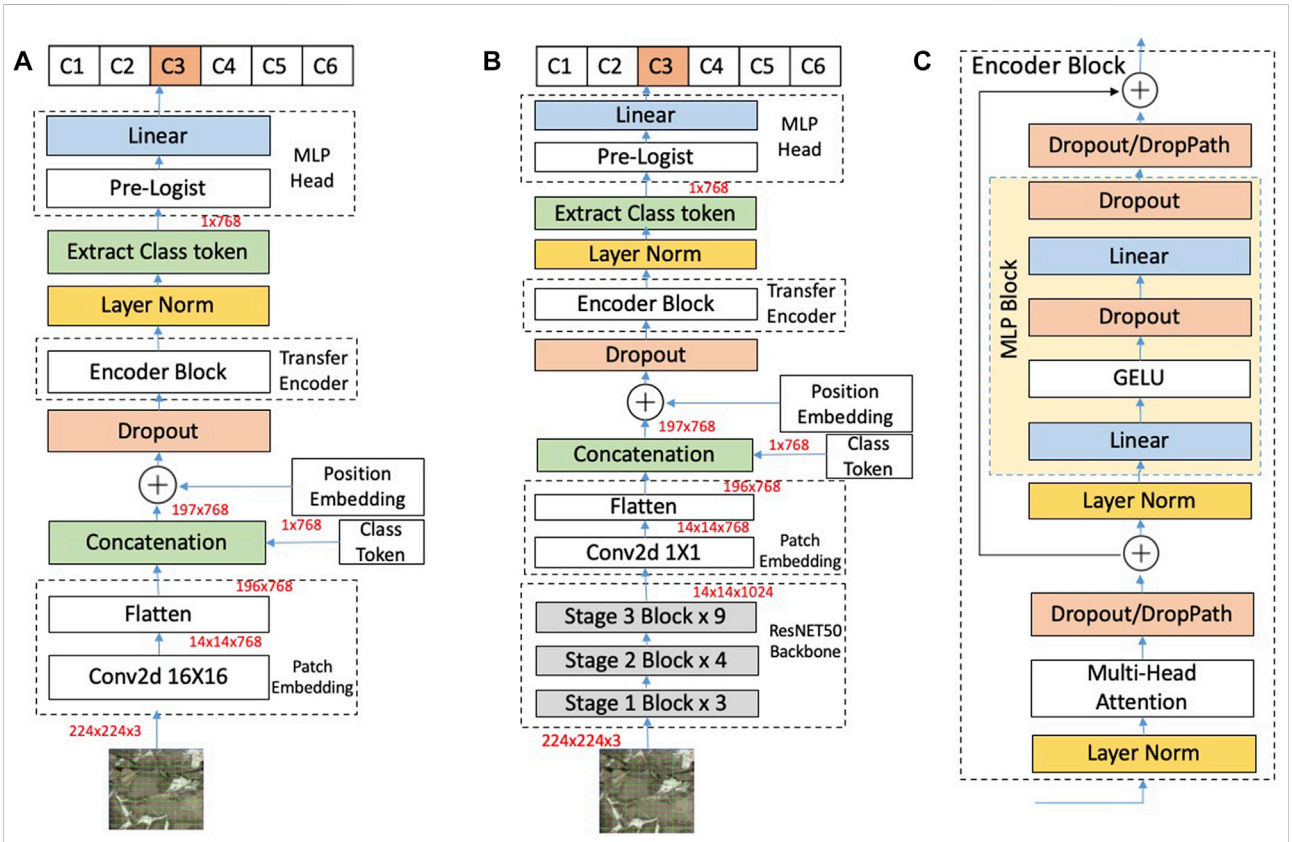

FIGURE 2
An overview of Vision Transformer architecture: **(A)** FaciesViT, **(B)** the hybrid ResNet and ViT architecture, and **(C)** the encoder block. The output layer is represented by Classes from C1 to C6 where C1: Argil, C2: Granit, C3: limestone, C4: Standstone lam, C5: Standstone mas, and C6: Siltstone. Since the input image belongs to limestone, this is highlighted in the output layer.

$$\text{Accuracy} = \frac{\text{number of instances classified correctly}}{\text{total number of instances}} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recal = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = 2 \times \frac{Precision \times Recal}{Precision + Recal} \quad (5)$$

Weighted average (WA) is considered due to the imbalanced nature of the datasets especially Dataset II (El-Alfy and Al-Azani, 2020).

$$WAPrecision = \frac{\sum_i Precision_i * count_i}{\sum_i count_i} \quad (6)$$

$$WARecal = \frac{\sum_i Recal_i * count_i}{\sum_i count_i} \quad (7)$$

$$WAF_1 = 2 \times \frac{WAPrecision \times WARecal}{WAPrecision + WARecal} \quad (8)$$

TABLE 3 Parameters considered for training models.

Parameter	Batch size	Learning rate	Momentum	Weight decay	Decay type	#Epochs	Learning rate warmup (K)
Value	64	1e-02	0.9	1e-04	Cosine	100	1

TABLE 4 Classification reports for CNN-based models for Dataset I and Dataset II.

	Dataset I												
	B-ResNet50			B-ResNet101			P-ResNet50			P-ResNet101			support
	prc	recal	f1	prc	recall	f1	prc	recall	f1	prc	recal	f1	
argil	0.817	0.856	0.836	0.795	0.894	0.842	0.942	0.942	0.942	0.917	0.962	0.939	104
granite	0.973	0.973	0.973	0.973	0.982	0.978	0.982	0.973	0.977	1.000	0.973	0.986	111
limestone	0.839	0.887	0.862	0.906	0.906	0.906	0.906	0.906	0.906	0.925	0.925	0.925	53
sandstone_lam	0.759	0.759	0.759	0.784	0.806	0.795	0.808	0.898	0.851	0.890	0.824	0.856	108
sandstone_mas	0.863	0.772	0.815	0.860	0.860	0.860	0.899	0.860	0.879	0.861	0.868	0.865	114
siltstone	0.728	0.755	0.741	0.817	0.691	0.749	0.885	0.836	0.860	0.809	0.845	0.827	110
Accuracy	0.828			0.852			0.902			0.897			
Weighted Avg	0.830	0.828	0.828	0.852	0.852	0.850	0.903	0.902	0.902	0.898	0.897	0.897	600
Dataset II													
argil	0.442	0.622	0.517	0.299	0.541	0.385	0.647	0.297	0.407	0.364	0.216	0.271	37
granite	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0
limestone	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0
sandstone_lam	0.939	0.355	0.515	0.927	0.410	0.569	0.841	0.512	0.636	0.845	0.502	0.630	217
sandstone_mas	0.154	0.167	0.160	0.067	0.067	0.067	0.241	0.333	0.280	0.200	0.267	0.229	60
siltstone	0.652	0.683	0.667	0.707	0.651	0.678	0.678	0.651	0.664	0.756	0.492	0.596	126
Accuracy	0.445			0.443			0.509			0.443			
Weighted Avg	0.708	0.445	0.510	0.694	0.443	0.516	0.696	0.509	0.576	0.691	0.443	0.535	440

3.5 CNN architecture

Residual-based Convolutional Neural Network (ResNet) (He et al., 2016) is considered as our baseline in this study due to its scalability while achieving satisfying results in image classification and object detection comparing with other CNN-based architectures. ResNet has different structures. In this study we consider ResNet50 and ResNet101 structures. Both ResNet50 and ResNet101 are trained either from scratch or by utilizing the pretraining approach. Therefore, four networks are evaluated as the baseline of this study, namely Basic ResNet50 (B-ResNet50), Basic ResNet101 (B-ResNet101), Pretrained ResNet50 (P-ResNet50) and Pretrained ResNet101 (P-ResNet101). Those models are implemented using Pytorch framework.

3.6 Vision transformers

3.6.1 ViT-base (ViT-B)

Three main models are presented and developed by Dosovitskiy et al. (2020) described in Table 2. Due to the small sizes of the datasets used in this study, the ViT-B structure is selected to be evaluated among others.

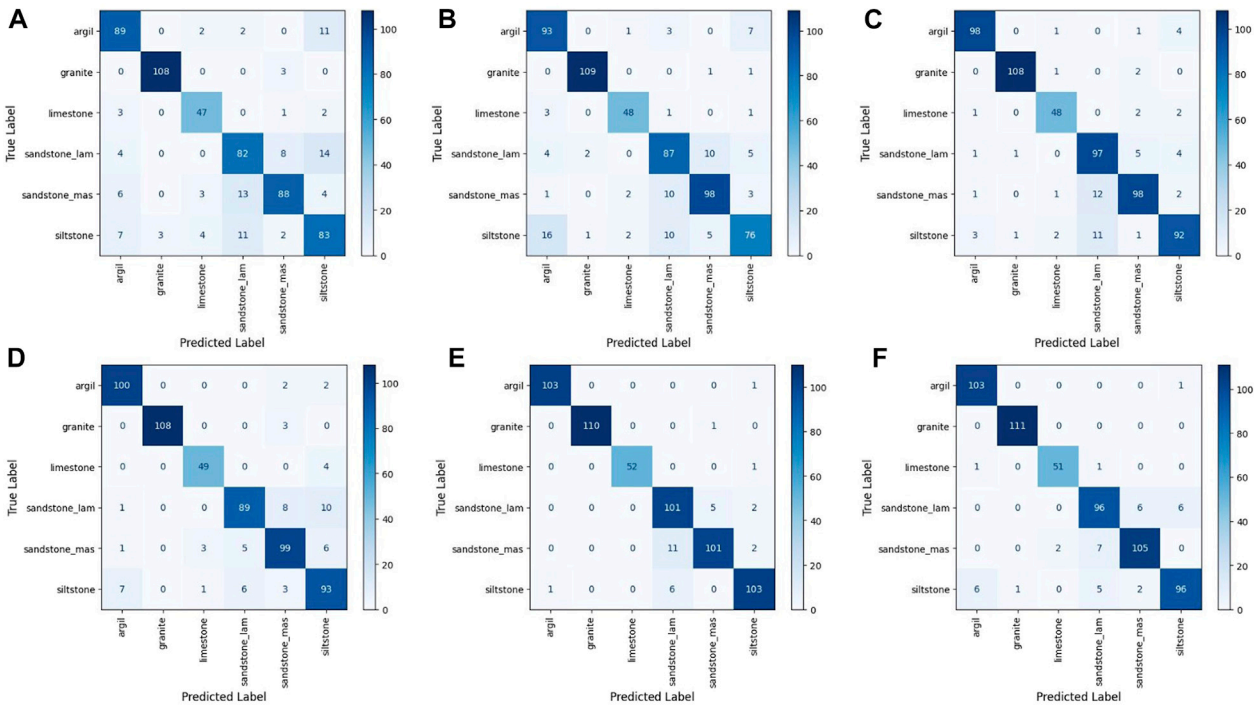

FIGURE 3
Confusion matrices for Dataset I **(A)** B-ResNet50 **(B)**B-ResNet101 **(C)**P-ResNet50 **(D)**P-ResNet101 **(E)** FaciesViT **(F)** P-R50-ViT.

As depicted in Figure 2A, an input image with a resolution of 244 × 224 × 3 first is divided into 16 patches and input into the patch embed block to generate a sequence of 2D patches. The Embedding Block is composed of a convolution layer of 16 × 16 and a flatten layer. The output of the flatten layer (196 × 768) is then contacted with a class token with a size of 1 × 768 and added to the training parameter, Position Embedding, to retain positional information, which is followed by the Dropout Layer. The resulting sequences of embedding vectors are fed into the Encoder Block which is composed of alternating layers of multi-headed self-attention depicted in Figure 2C. The outputs of the Encoder Block is fed into the Layer Norm. The output related to the class token (1 × 768) is then extracted and fed into MLP Head block.

3.6.2 ResNet50-ViT

Instead of feeding image patches into the transformer, feature maps generated using CNN can be fed into the transformers (Dosovitskiy et al., 2020). A hybrid model of CNN, especially ResNet50 and vision transformer is also evaluated in this study (ResNet50-ViT). The ResNet50 network serves as a feature extractor. Therefore, the input to the vision transformer is the features maps generated using ResNet instead of feeding the image patches in the previous vision transformer model. Figures 2B,C depict the structure of the hybrid model.

3.7 Experimental setup

Our experiments are implemented on a Lambda Workstation of AMD® Ryzen threadripper pro 3975wx 32-cores × 64 and three NVIDIA GeForce RTX 3090 GPUs with a graphic memory of 24 GB for each. Anaconda 4.13.0 (2022-05-19) is configured on Ubuntu 20.04.4 LTS and Pytorch framework is used (Paszke et al., 2019). In addition, Scikit learn package (Pedregosa et al., 2011) is used to report the results. Table 3 presents the parameters considered for training and evaluating the models.

4 Results

4.1 Dataset I

4.1.1 Convolutional neural networks

In this study, we evaluated and compared two different residual-based CNN architectures (ResNet), ResNet50 and

TABLE 5 Classification reports for ViT-based models for Dataset I and Dataset II.

	Dataset I						
	P-R50-ViT			FaciesViT			support
	prc	recal	f1	prc	recal	f1	
Argil	0.936	0.990	0.963	0.990	0.990	0.990	104
Granite	0.991	1.000	0.996	1.000	0.991	0.995	111
limestone	0.962	0.962	0.962	1.000	0.981	0.990	53
sandstone_lam	0.881	0.889	0.885	0.856	0.935	0.894	108
sandstone_mas	0.929	0.921	0.925	0.944	0.886	0.914	114
Siltstone	0.932	0.873	0.901	0.945	0.936	0.941	110
Accuracy	0.937			0.950			
Weighted Avg	0.937	0.937	0.936	0.952	0.950	0.950	600
Dataset II							
Argil	0.700	0.378	0.491	0.429	0.162	0.235	37
granite	0.000	0.000	0.000	0.000	0.000	0.000	0
limestone	0.000	0.000	0.000	0.000	0.000	0.000	0
sandstone_lam	0.870	0.585	0.700	0.797	0.705	0.748	217
sandstone_mas	0.267	0.333	0.296	0.447	0.567	0.500	60
siltstone	0.769	0.556	0.645	0.697	0.659	0.678	126
Accuracy	0.525			0.627			
Weighted Avg	0.745	0.525	0.612	0.690	0.627	0.651	440

ResNet101 to train and perform image classification on the first core image dataset. In the first experiment, we train the CNNs from scratch following the learning parameters established in Table 3. Overall, ResNet101, which has more layers, skip connections, and deeper networks, perform better than ResNet50 achieving up to 86% in all classification metrics, particularly the weighted average of f1 score, precision, and accuracy (Table 4). In contrast, RestNet50 only achieved 83% in both the f1 score, precision, and weighted average across the metrics (Table 4). In the second experiment, we pre-trained the ResNet models with the ImageNet dataset and with learning parameters similar to those in the previous experiment. All classification metrics in both models have improved significantly yielding up to 90% in the weighted average of all metrics and accuracy, an increase of 8.5% from the first experiment (Table 4). Unlike the first experiment, the ResNet50 performs slightly better than the ResNet101 during the second experiment (Table 4).

The result table and the confusion matrix further show that the baseline ResNet50 model particularly struggled to classify laminated sandstone, massive sandstone, and siltstone (Figure 3 and Table 4). The model is often confused between laminated sandstone and siltstone. In contrast, the baseline ResNet101 model faces some difficulties to differentiate between argillite and siltstone (Figure 3). However, this mistake is rather consistent with a typical human error as argillite and siltstone can have very similar appearances. Although pre-trained models achieved much improved results, the models still exhibit issues similar to baseline models, particularly in distinguishing between sandstones and siltstone (Figure 3).

4.1.2 Vision transformer

For Dataset I, two Vision Transformer-based architectures, mentioned above, were evaluated. This was conducted to have a better insight into the actual performance of ViT in the geosciences dataset. Both models were pre-trained with the

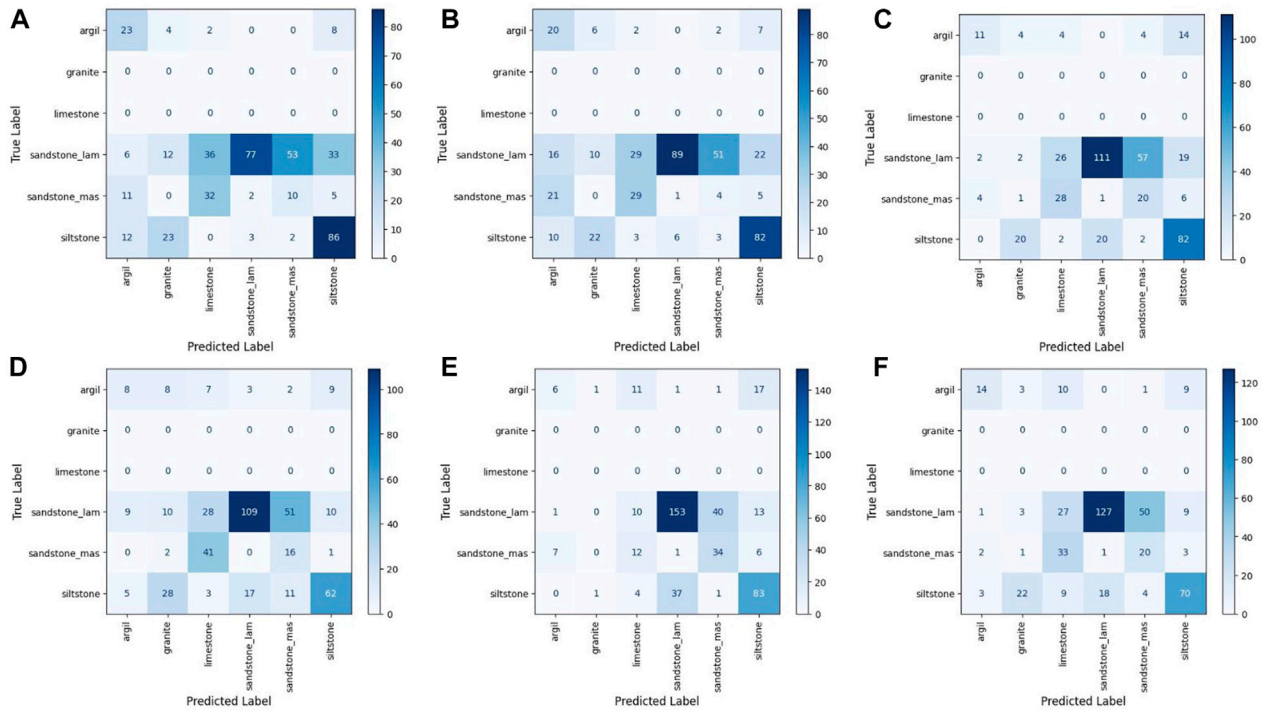

FIGURE 4
Confusion matrices for Dataset II **(A)** B-ResNet50, **(B)**B-ResNet101 **(C)**P-ResNet50 **(D)**P-ResNet101 **(E)** FaciesViT **(F)** P-R50-ViT.

TABLE 6 Summary of results. The best results are presented in bold.

Dataset	Model	Precision	Recall	f1	Acc
Dataset I	B-ResNet50	0.8297	0.8283	0.8284	0.8283
	B-ResNet101	0.8521	0.8517	0.8504	0.8517
	P-ResNet50	0.9035	0.9017	0.9019	0.9017
	P-ResNet101	0.8977	0.8967	0.8968	0.8967
	FaciesViT	**0.9517**	**0.9500**	**0.9503**	**0.9500**
	P-R50-ViT	0.9366	0.9367	0.9363	0.9367
Dataset II	B-ResNet50	0.7079	0.4455	0.5102	0.4455
	B-ResNet101	0.6938	0.4432	0.5160	0.4432
	P-ResNet50	0.6961	0.5091	0.5763	0.5091
	P-ResNet101	0.6911	0.4432	0.5354	0.4432
	FaciesViT	0.6898	**0.6273**	**0.6510**	**0.6273**
	P-R50+ViT	**0.7445**	0.525	0.6116	0.5250

ImageNet dataset and with hyperparameters similar to CNN models (Table 3). In general, both ViT-based models significantly outperform all CNN architectures, with an increase of up to 15% in all classification metrics (Table 5). Our proposed ViT achieved 95% in both the weighted average of the f1 score and other metrics (Table 5), while the hybrid model shows a slightly lower performance, achieving around 93% in both the weighted average of the f1 score and other metrics in the test data set (Table 5). Furthermore, the FaciesViT model shows the most stable performance across different classes and classification metrics (Figure 4 and Table 5).

The confusion matrix shows that the hybrid CNN-ViT model has almost correctly predicted all classes, except between laminated and massive sandstones (Figure 3). Although this issue may not occur for geologists, it can be understood when looking at the training and test datasets of these two classes as they can sometimes appear similar (Figure 1). The FaciesViT model seems to fix this problem and can achieve 95% accuracy (Figure 3). However, the model still struggled slightly to differentiate between argillite and siltstone (Figure 3).

4.2 Dataset II

In this study, we examined the generalizability of all models by performing a blind test on a new, unseen core dataset (Dataset II). This core dataset is collected from a different geological basin but has almost all the rock types as in Dataset I, except granite and limestone. Across the different CNN models, baseline and pre-trained, both the accuracy and the weighted average of the f1 score have decreased significantly, only 44 and 57%,

respectively (Table 4). While the recall values show similar scores across CNN architectures, precision could produce around 70% in Dataset II (Table 4) with the baseline ResNet achieving the highest value among all CNN-based models. The model most correctly predicted siltstone and laminated sandstone but at the same time struggles to classify the rest of the rock types and even made a classification error between laminated and massive sandstones (Figure 4).

Although the ViT models also experienced a similar performance decrease in this dataset. Both ViT-based models still outperform CNN models, achieving up to 65% in both f1-score and accuracy (Table 5). Furthermore, the base ViT model performs slightly better than the hybrid ViT-CNN model (Tables 5 and 6). Compared to all models, the weighted average precision score of the hybrid model yields the highest value, reaching 75% (Table 5). However, its performance is not stable and only achieved 52% in the weighted average recall score. The results of our experiment with Dataset II further show that our proposed FaciesViT model exhibits the most consistent and stable performance across all the evaluation metrics. Similarly, both transformer-based models show high accuracy when classifying laminated sandstone and siltstone (Figure 4 and Table 5). The confusion matrix shows that the proposed FaciesViT model can generalize better and predict all types of rock more equally than the other CNN models (Figure 4).

5 Discussion

5.1 Vision transformer for core lithofacies classification

Applications of deep learning to automate core lithofacies classification have reached state-of-the-art results, matching geologist-level classification for core interpretation (de Lima et al., 2019; Baraboshkin et al., 2020; Falivene et al., 2022). Now more than ever, the need to automate subsurface geological interpretation has peaked due to rapid digital transformation and streamlined data transfer. However, the true potential of deep learning for core image classification remains under-explored due to limited data availability and a high-quality labeled dataset. This is further compounded by the narrow focus of applying Convolutional Neural Networks and supervised learning to perform this task. Recent developments of a deep learning algorithm for computer vision have introduced the implementation of a transformer-based algorithm, known as Vision Transformer (ViT; Dosovitskiy et al., 2020). This algorithm differs from CNN because it focuses on the sequence of images and patches rather than individual pixels. Furthermore, this model benefits from a multi-head self-attention mechanism to learn the importance of features for the classification and optimize the classification with a limited dataset. Recent work indicates that Vision Transformer has outperformed CNN in many computer vision tasks, including image classification, image superresolution, and segmentation (Liu et al., 2021; Xie et al., 2021). However, the applications of ViT are very limited in geosciences and further exploration is required to fully uncover the power and applicability of the transformer-based algorithm to conduct visual recognition in geosciences.

In this study, for the first time, we proposed and developed a novel transformer-based framework to perform fully automated core lithofacies classification using datasets that have been studied by Baraboshkin et al. (2020). To evaluate and validate our proposed model, we conducted several experiments using recent CNN architectures (ResNet50 and ResNet101) and a hybrid CNN-ViT model. Our results show that the proposed FaciesViT model is much superior to the CNN and hybrid CNN-ViT models, achieving a weighted average f1 score 15% higher than all CNN architectures (Table 6). In addition, we further tested the trained ViT model (train to Dataset I) to perform classification on the entirely new and unseen dataset (Dataset II). A similar experiment was previously performed and showed that their best CNN model that achieved $>90\%$ could not generalize the information to the new dataset and only achieved a score of less than 50% f1-score even after extensive data augmentation (Baraboshkin et al., 2020). In our study, we showed that the proposed ViT model could generalize better when tested on the unseen data set compared to other CNN algorithms. This phenomenon suggests that ViT learns better than CNN and could transfer knowledge from the dataset where it is trained to the out-of-distribution dataset. This is evident through the visualization of the layers using the attention rollout to show what features are important for the classification (Figure 5) and most are similar to the characteristics used by geologists to identify the different types of rocks. Among the different methods, the mean attention rollout provides the most information on how the FaciesViT model informed its decision (Figure 5). This can be justified as the FaciesViT works better with the original images than with the feature maps extracted using CNN which is a major advantage of using ViT because the ViT model can outperform CNN models without requiring preprocessing steps and feature extraction processes. This provides a major advance towards developing a general deep learning model for image classification using transformed-based architecture. As reported in previous work, ViT has the potential to replace CNN in performing various computer vision tasks because it provides (i) a more efficient and robust model; (ii) a model with higher generalizability, and (iii) a more explainable deep learning model.

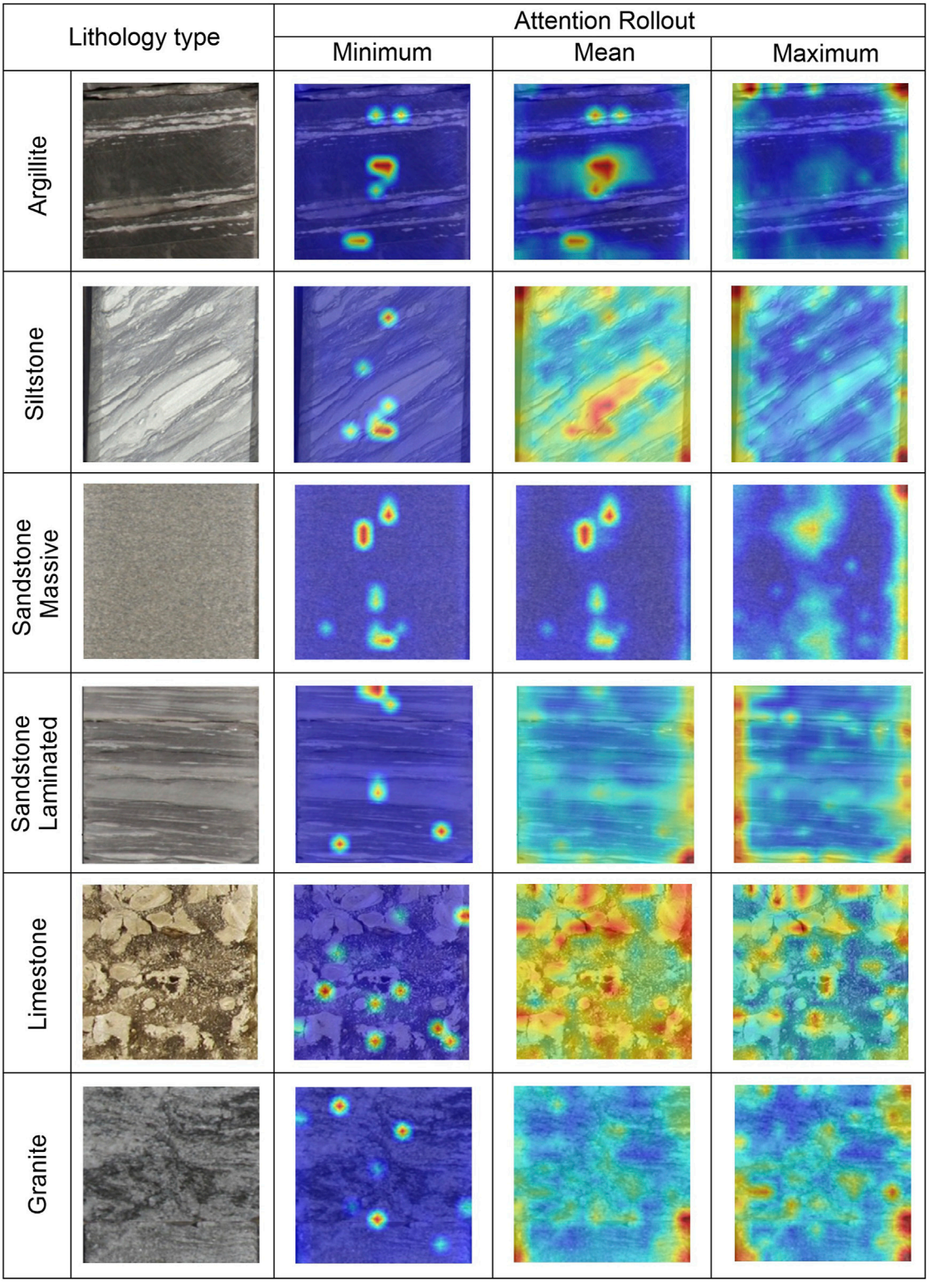

FIGURE 5
Visualizations of learned features within the transformer layers using attention rollout in different rock types.

5.2 Future recommendation and limitation

This study is designed as a proof-of-concept on the application of ViT for core lithofacies classification and explores how it improved the overall performance of a deep learning algorithm for such a task when compared with traditional CNN algorithms. Hence, we also acknowledged several limitations of this work, including: (i) our study uses a fairly limited dataset ($< 10k$) and an imbalance dataset which may have a negative impact on the learning and prediction processes. Therefore, an additional dataset is required to unravel the actual potential of ViT; and ((ii) we only examined the six most common lithofacies in the subsurface reservoir and uses a general category to group them. A typical reservoir characterization would require a more detailed analysis and grouping. For example, limestone can be further divided into at least five lithofacies, such as mudstone, wackestone, packstone, grainstone, and boundstone which will hold more information than just using limestone as a category. Although we have tested our model to predict an unseen dataset from different geological settings to test the transferability of our model, future works should consider using more detailed lithofacies types to better represent actual subsurface reservoir conditions and complexities. In addition, ViT-based models require larger training dataset and more computational recourses compared to CNN-based models due to the complexity of the used attention mechanism. We recommend using more advanced ViTs overcoming such limitations. Furthermore, a more advanced data augmentation (e.g., label smoothing, CutMix; Koeshidayatullah, 2022) may further improve the performance of the model and optimize the training processes.

6 Conclusion

- This study, for the first time, utilized a transformer-based architecture *FaciesViT*, to perform lithofacies classification directly from core images.
- Our proposed model can match the performance of other CNN models without heavy data augmentation. Furthermore, the model can generalize better for the unseen dataset, which provides a significant step forward in the application of deep learning to lithofacies interpretation.
- The attention rollout technique shows that the algorithm bases its classification on features used by geologists to differentiate and classify lithofacies. This improves the overall explainability and transferability of the model.
- Although the model can predict an unseen and out-of-distribution dataset with an accuracy of up to 65%, a more diverse and larger volume of dataset would help the prediction of our model to other datasets.

Data availability statement

The raw data is available upon a reasonable request to the corresponding authors.

Author contributions

AK: Conceptualization, Writing—original draft, Data curation, Writing—review and editing, equal contribution. SA-A: Conceptualization, Methodology, Writing—original draft, Writing—review and editing, equal contribution. EB: Data Curation, Writing—review and editing. MA: Conceptualization, Methodology, Writing—review and editing.

Acknowledgments

The authors would like to thank King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia for all its support. AK would like to thank the support from the CPG startup fund (SF-21011) for the resources and facilities to conduct this study. The authors would also like to thank SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI) for the support and computing resources. We acknowledged the constructive discussions with Dmitry Koroteev and Denis Orlov from Skoltech, Russia in the preparation of the dataset and manuscript.

Conflict of interest

EB is employed by Digital Petroleum LLC, Moscow, Russia.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Al-Ramadan, K., Koeshidayatullah, A., Cantrell, D., and Swart, P. K. (2020). Impact of basin architecture on diagenesis and dolomitization in a fault-bounded carbonate platform: Outcrop analogue of a pre-salt carbonate reservoir, red sea rift, nw Saudi Arabia. *Pet. Geosci.* 26, 448–461. doi:10.1144/petgeo2018-125

Alaudah, Y., Michałowicz, P., Alfarraj, M., and AlRegib, G. (2019). A machine-learning benchmark for facies classification. *Interpretation* 7, SE175–SE187. doi:10.1190/int-2018-0249.1

Alqahtani, N., Armstrong, R. T., and Mostaghimi, P. (2018). "Deep learning convolutional neural networks to predict porous media properties," in *SPE Asia Pacific oil and gas conference and exhibition*. (OnePetro). Brisbane, Australia: SPE.

Alzubaidi, F., Mostaghimi, P., Swietojanski, P., Clark, S. R., and Armstrong, R. T. (2021). Automated lithology classification from drill core images using convolutional neural networks. *J. Petroleum Sci. Eng.* 197, 107933. doi:10.1016/j.petrol.2020.107933

Amao, A. O., Al-Ramadan, K., and Koeshidayatullah, A. (2016). Automated mineralogical methodology to study carbonate grain microstructure: An example from oncoids. *Environ. Earth Sci.* 75, 666. doi:10.1007/s12665-016-5492-x

Amel, H., Jafarian, A., Husinec, A., Koeshidayatullah, A., and Swennen, R. (2015). Microfacies, depositional environment and diagenetic evolution controls on the reservoir quality of the permian upper dalan formation, kish gas field, zagros basin. *Mar. Petroleum Geol.* 67, 57–71. doi:10.1016/j.marpetgeo.2015.04.012

Anees, A., Zhang, H., Ashraf, U., Wang, R., Liu, K., Mangi, H., et al. (2022b). Identification of favorable zones of gas accumulation via fault distribution and sedimentary facies: Insights from hangjinqi area, northern ordos basin. *Front. Earth Sci. (Lausanne).* 9, 822670. doi:10.3389/feart.2021.822670

Anees, A., Zhang, H., Ashraf, U., Wang, R., Liu, K., Abbas, A., et al. (2022a). Sedimentary facies controls for reservoir quality prediction of lower shihezi member-1 of the hangjinqi area, ordos basin. *Minerals* 12, 126. doi:10.3390/min12020126

Ashraf, U., Zhu, P., Yasin, Q., Anees, A., Imraz, M., Mangi, H. N., et al. (2019). Classification of reservoir facies using well log and 3d seismic attributes for prospect evaluation and field development: A case study of sawan gas field, Pakistan. *J. Petroleum Sci. Eng.* 175, 338–351. doi:10.1016/j.petrol.2018.12.060

Ayranci, K., Yildirim, I. E., Waheed, U. b., and MacEachern, J. A. (2021). Deep learning applications in geosciences: Insights into ichnological analysis. *Appl. Sci.* 11, 7736. doi:10.3390/app11167736

Baraboshkin, E. E., Ismailova, L. S., Orlov, D. M., Zhukovskaya, E. A., Kalmykov, G. A., Khotylev, O. V., et al. (2020). Deep convolutions for in-depth automated rock typing. *Comput. Geosciences* 135, 104330. doi:10.1016/j.cageo.2019.104330

Baraboshkin, E., Ivchenko, A., Ismailova, L., Orlov, D., Baraboshkin, E. Y., and Koroteev, D. (2018). "Core photos lithological interpretation using neural networks," in *20th international sedimentological congress*. Quebec, Canada: IAS.

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., and Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Remote Sens.* 13, 516. doi:10.3390/rs13030516

Cao, R., Fang, L., Lu, T., and He, N. (2020). Self-attention-based deep feature fusion for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* 18, 43–47. doi:10.1109/lgrs.2020.2968550

Chen, C.-F. R., Fan, Q., and Panda, R. (2021). "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision* (Montreal, Canada: IEEE), 357–366.

Chen, H., Qi, Z., and Shi, Z. (2021). Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi:10.1109/tgrs.2021.3095166

Croudace, I. W., and Rothwell, R. G. (2015). *Micro-XRF studies of sediment cores: Applications of a non-destructive tool for the environmental sciences*, 17. Berlin, Germany: Springer.

de Lima, R. P., Suriamin, F., Marfurt, K. J., and Pranter, M. J. (2019). Convolutional neural networks as aid in core lithofacies classification. *Interpretation* 7, SF27–SF40. doi:10.1190/int-2018-0245.1

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition* (FL, United States: IEEE), 248–255.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. New York, USA. *arXiv* [Preprint]. doi:10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. New York, USA. *arXiv* [Preprint]. doi:10.48550/arXiv.2010.11929

El-Alfy, E.-S. M., and Al-Azani, S. (2020). Empirical study on imbalanced learning of Arabic sentiment polarity with neural word embedding. *J. Intelligent Fuzzy Syst.* 38, 6211–6222. doi:10.3233/jifs-179703

Falivene, O., Auchter, N. C., Pires de Lima, R., Kleipool, L., Solum, J. G., Zarian, P., et al. (2022). Lithofacies identification in cores using deep learning segmentation and the role of geoscientists: Turbidite deposits (gulf of Mexico and north sea). *Am. Assoc. Pet. Geol. Bull.* 106, 1357–1372. doi:10.1306/03112221015

Fan, G., Chen, F., Chen, D., and Dong, Y. (2020). Recognizing multiple types of rocks quickly and accurately based on lightweight cnns model. *IEEE Access* 8, 55269–55278. doi:10.1109/access.2020.2982017

Ferreira, A., and Giraldi, G. (2017). Convolutional neural network approaches to granite tiles classification. *Expert Syst. Appl.* 84, 1–11. doi:10.1016/j.eswa.2017.04.053

Ferreira, I., Ochoa, L., and Koeshidayatullah, A. (2022). On the generation of realistic synthetic petrographic datasets using a style-based gan. *Sci. Rep.* 12, 12845. doi:10.1038/s41598-022-16034-4

Fu, D., Su, C., Wang, W., and Yuan, R. (2022). Deep learning based lithology classification of drill core images. *Plos one* 17, e0270826. doi:10.1371/journal.pone.0270826

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., et al. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. New York, USA. *arXiv* [Preprint]. doi:10.48550/arXiv.1706.02677

Harinie, T., Janani Chellam, I., Sathya Bama, S., Raju, S., and Abhaikumar, V. (2012). "Classification of rock textures," in *Proceedings of the international conference on information systems design and intelligent applications 2012 (India 2012) held in visakhapatnam, India, january 2012* (Berlin, Germany: Springer), 887–895.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Las Vegas, United States: IEEE), 770–778.

Ivchenko, A. V., Baraboshkin, E. E., Ismailova, L. S., Orlov, D. M., Koroteev, D. A., and Baraboshkin, E. Y. (2018). "Core photo lithological interpretation based on computer analyses," in *Proceedings of the IEEE northwest Russia conference on mathematical methods in engineering and technology* (Saint-Petersburg, Russia: IEEE), 10–14.

Koeshidayatullah, A., Al-Ramadan, K., and Hughes, G. W. (2016). Facies mosaic and diagenetic patterns of the early devonian (late pragian–early emsian) microbialite-dominated carbonate sequences, qasr member, jauf formation, Saudi Arabia. *Geol. J.* 51, 704–721. doi:10.1002/gj.2678

Koeshidayatullah, A., Morsilli, M., Lehrmann, D. J., Al-Ramadan, K., and Payne, J. L. (2020). Fully automated carbonate petrography using deep convolutional neural networks. *Mar. Petroleum Geol.* 122, 104687. doi:10.1016/j.marpetgeo.2020.104687

Koeshidayatullah, A. (2022). Optimizing image-based deep learning for energy geoscience via an effortless end-to-end approach. *J. Petroleum Sci. Eng.* 215, 110681. doi:10.1016/j.petrol.2022.110681

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. neural Inf. Process. Syst.* 25.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature* 521, 436–444. doi:10.1038/nature14539

Leeder, M. R. (2012). *Sedimentology: Process and product*. Berlin, Germany: Springer Science & Business Media.

Li, N., Hao, H., Gu, Q., Wang, D., and Hu, X. (2017). A transfer learning method for automatic identification of sandstone microscopic images. *Comput. Geosciences* 103, 111–121. doi:10.1016/j.cageo.2017.03.007

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision* (Montreal, Canada: IEEE), 10012–10022.

Lobos, R., Silva, J. F., Ortiz, J. M., Díaz, G., and Egaña, A. (2016). Analysis and classification of natural rock textures based on new transform-based features. *Math. Geosci.* 48, 835–870. doi:10.1007/s11004-016-9648-8

Martin, T., Meyer, R., and Jobe, Z. (2021). Centimeter-scale lithology and facies prediction in cored wells using machine learning. *Front. Earth Sci. (Lausanne).* 491. doi:10.3389/feart.2021.659611

McPhee, C., Reed, J., and Zubizarreta, I. (2015). *Core analysis: A best practice guide*. Amsterdam, Netherlands: Elsevier.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). Wavenet: A generative model for raw audio. New York, USA. *arXiv* [Preprint]. doi:10.48550/arXiv.1609.03499

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. neural Inf. Process. Syst.* 32.

Patel, A. K., Chatterjee, S., and Gorai, A. K. (2017). "Development of online machine vision system using support vector regression (svr) algorithm for grade prediction of iron ores," in 2017 fifteenth IAPR international conference on machine vision applications (MVA), Nagoya, Japan, 08-12 May 2017 (Nagoya, Japan: IEEE), 149–152.

Patel, A. K., Gorai, A. K., and Chatterjee, S. (2016). *Development of machine vision-based system for iron ore grade prediction using Gaussian process regression (gpr).* Minsk, Belarus.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.48550/arXiv.1201.0490

Pires de Lima, R., and Duarte, D. (2021). Pretraining convolutional neural networks for mudstone petrographic thin-section image classification. *Geosciences* 11, 336. doi:10.3390/geosciences11080336

Prince, C., Dixon, M., and Haynes, L. (2005). "The use of high-resolution core imagery in reservoir characterization: An example from unlithified miocene turbidites," in *Paper SCA2005-02, society of core analysts annual international symposium*. Toronto, Canada: SPWLA.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Netw.* 12, 145–151. doi:10.1016/s0893-6080(98)00116-6

Rothwell, R. G., and Rack, F. R. (2006). New techniques in sediment core analysis: An introduction. *Geol. Soc. Lond. Spec. Publ.* 267, 1–29. doi:10.1144/gsl.sp.2006.267.01.01

Seleznev, I., Abashkin, V., Chertova, A., Makienko, D., Istomin, S., Romanov, D., et al. (2020). "Joint usage of whole core images obtained in different frequency ranges for the tasks of automatic lithotype description and modeling of rocks' petrophysics properties," in *Geomodel 2020* (Gelendzhik, Russia: European Association of Geoscientists & Engineers), 2020, 1–5.

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60–48. doi:10.1186/s40537-019-0197-0

Singh, M., Javadi, A., and Singh, S. (2004). "A comparison of texture teatures for the classification of rock images," in *International conference on intelligent data engineering and automated learning* (Berlin, Germany: Springer), 179–184.

Thomas, A., Rider, M., Curtis, A., and MacArthur, A. (2011). *Automated lithology extraction from core photographs. first break* 29. Netherlands: EAGE.

Timmer, E., Knudson, C., and Gingras, M. (2021). Applying deep learning for identifying bioturbation from core photographs. *Am. Assoc. Pet. Geol. Bull.* 105, 631–638. doi:10.1306/08192019051

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in neural information processing systems* (CA, United States: MIT Press), 30.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *J. Big Data* 3, 9–40. doi:10.1186/s40537-016-0043-6

Wrona, T., Pan, I., Gawthorpe, R. L., and Fossen, H. (2018). Seismic facies analysis using machine learning. *Geophysics* 83, O83–O95. doi:10.1190/geo2017-0595.1

Wu, B., Meng, D., Wang, L., Liu, N., and Wang, Y. (2020). Seismic impedance inversion using fully convolutional residual network and transfer learning. *IEEE Geosci. Remote Sens. Lett.* 17, 2140–2144. doi:10.1109/lgrs.2019.2963106

Wu, X., Liang, L., Shi, Y., and Fomel, S. (2019). Faultseg3d: Using synthetic data sets to train an end-to-end convolutional neural network for 3d seismic fault segmentation. *Geophysics* 84, IM35–IM45. doi:10.1190/geo2018-0646.1

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* 34, 12077–12090. doi:10.48550/arXiv.2105.15203

Xu, Z., Shi, H., Lin, P., and Liu, T. (2021). Integrated lithology identification based on images and elemental data from rocks. *J. Petroleum Sci. Eng.* 205, 108853. doi:10.1016/j.petrol.2021.108853

Zhang, H., Xue, J., and Dana, K. (2017). "Deep ten: Texture encoding network," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Hawaii, United States: IEEE), 708–717.