Petroleum Science 22 (2025) 3207–3228

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A large-scale, high-quality dataset for lithology identification: Construction and applications

Jia-Yu Li [a, b], Ji-Zhou Tang [a, b, *], Xian-Zheng Zhao [c, **], Bo Fan [d], Wen-Ya Jiang [e], Shun-Yao Song [e], Jian-Bing Li [e], Kai-Da Chen [f], Zheng-Guang Zhao [g]

[a] School of Ocean and Earth Science, Tongji University, Shanghai, 200092, China
[b] State Key Laboratory of Marine Geology, Tongji University, Shanghai, 200092, China
[c] CNPC Advisory Center, Beijing, 100724, China
[d] Motorola Solutions, Inc, Somerville, 02145, USA
[e] Dagang Oilfield Company, PetroChina, Tianjin, 300280, China
[f] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China
[g] School of Mine Safety, North China Institute of Science and Technology, Sanhe, 065201, Hebei, China

ARTICLE INFO

Article history:
Received 22 September 2024
Received in revised form
14 February 2025
Accepted 16 April 2025
Available online 21 April 2025

Edited by Meng-Jiao Zhou

Keywords:
Geoenergy exploration
Lithology identification
Lithology dataset
Artificial intelligence
Deep learning
Drill core

ABSTRACT

Lithology identification is a critical aspect of geoenergy exploration, including geothermal energy development, gas hydrate extraction, and gas storage. In recent years, artificial intelligence techniques based on drill core images have made significant strides in lithology identification, achieving high accuracy. However, the current demand for advanced lithology identification models remains unmet due to the lack of high-quality drill core image datasets. This study successfully constructs and publicly releases the first open-source Drill Core Image Dataset (DCID), addressing the need for large-scale, high-quality datasets in lithology characterization tasks within geological engineering and establishing a standard dataset for model evaluation. DCID consists of 35 lithology categories and a total of 98,000 high-resolution images (512 × 512 pixels), making it the most comprehensive drill core image dataset in terms of lithology categories, image quantity, and resolution. This study also provides lithology identification accuracy benchmarks for popular convolutional neural networks (CNNs) such as VGG, ResNet, DenseNet, MobileNet, as well as for the Vision Transformer (ViT) and MLP-Mixer, based on DCID. Additionally, the sensitivity of model performance to various parameters and image resolution is evaluated. In response to real-world challenges, we propose a real-world data augmentation (RWDA) method, leveraging slightly defective images from DCID to enhance model robustness. The study also explores the impact of real-world lighting conditions on the performance of lithology identification models. Finally, we demonstrate how to rapidly evaluate model performance across multiple dimensions using low-resolution datasets, advancing the application and development of new lithology identification models for geoenergy exploration.

1. Introduction

Lithology identification is the cornerstone of geoenergy exploration, as the accuracy of lithological characterization directly impacts various geoenergy-related fields, including geothermal energy development efficiency, gas hydrate extraction success, and the safety of gas storage (Xu et al., 2021; Zhao et al., 2022; Liu et al., 2023; Zhang et al., 2024; Tang et al., 2024). Ensuring precise lithology identification is thus a critical challenge in geoenergy exploration. Traditional lithology identification methods primarily rely on the interpretation of well logging data and the visual analysis of drill core samples (Thomas et al., 2011; Ren et al., 2022). Among these, core-based identification—which allows direct observation of the target rock formations—provides the most reliable information (Borsaru et al., 2006; Izadi et al., 2017; Huang et al., 2023). However, this approach often requires manual inspection and involves various physical and chemical techniques, such as X-ray diffraction (XRD), CT scanning, scanning electron microscopy (SEM), isotope analysis, and spectral analysis (Fu et al.,

* Corresponding author.
** Corresponding author
E-mail addresses: jeremytang@tongji.edu.cn (J.-Z. Tang), xzzhao@petrochina.com.cn (X.-Z. Zhao).

https://doi.org/10.1016/j.petsci.2025.04.013

2017). Consequently, it is time-consuming, labor-intensive, and demands highly skilled professionals with extensive field experience (Galdames et al., 2017). The advent of artificial intelligence presents a promising alternative by enabling automated lithology identification (LeCun et al., 2015; Chen and Li, 2022).

The development of automated lithology prediction using artificial intelligence (AI) techniques first emerged in the field of lithology identification based on well logging data (Tian et al., 2013; He et al., 2019; Bai et al., 2025). Well logging data, including gamma ray (GR), acoustic (AC), caliper (CAL), and density (DEN) logs, contains rich nonlinear geological features that can serve as indicators for lithology classification (Busch et al., 1987). Two open-source well logging lithology datasets have been widely used by researchers to assess the effectiveness of various methods. One dataset is from the Hugoton and Panoma fields in Kansas, USA (Dubois et al., 2007), and the other is from the Daniudi Gas Field (DGF) and Hangjinqi Gas Field (HGF) in the Huabei Oilfield, China (Xie et al., 2018). Xie et al. (2018) compared the lithology identification performance of five machine learning methods using the DGF and HGF datasets, including naïve Bayes, support vector machine, artificial neural network, random forest, and gradient tree boosting. The results indicate that ensemble methods are effective for the supervised classification of lithology using well log data. Imamverdiyev and Sukhostat (2019) proposed a novel one-dimensional convolutional neural network (1D-CNN) model, trained with various optimization algorithms, and compared its performance with recurrent neural networks, long short-term memory, support vector machine, and *k*-nearest neighbor models using the Hugoton and Panoma Fields dataset, showing more accurate results. Zhao et al. (2023) introduced a classification-enhanced semi-supervised generative adversarial network (CE-SGAN), which can alleviate the impact of imbalanced well logging data. Experimental results on the Hugoton and Panoma Fields dataset and the DGF-HGF dataset, compared with other methods, demonstrate the significant advantages of this approach. Dong et al. (2023) proposed a deep kernel method (DKM) for lithofacies identification using well log curves, which has excellent nonlinear feature fitting capability, superior accuracy, and faster processing speed on the DGF dataset. Overall, the development of lithology identification based on well logging data using artificial intelligence techniques has been greatly facilitated by publicly available well logging lithology datasets. These datasets enable researchers to quickly assess the performance of proposed models and facilitate comparisons with other studies.

In recent years, researchers have demonstrated the potential of lithology identification based on image log analysis (Shi et al., 2023). This approach leverages the fact that images provide two-dimensional data, which contain more information compared to traditional one-dimensional well logging data, thereby enabling more accurate lithology identification (Marmo et al., 2005). Early work by Thomas et al. (2011) proposed an object-based image analysis method for lithology classification using grayscale core images containing four lithologies: carbonate cement, shale, sandstone, and voids. The method employed multi-resolution segmentation to extract features and trained a nearest neighbor classifier, achieving an accuracy of up to 94.29%. Subsequently, Wieling (2013) demonstrated the effectiveness of principal component analysis (PCA) in lithology identification. PCA was used to reduce image features to lower-dimensional representations, followed by classification using methods such as support vector machines (SVM) to map nonlinear features into a linear space for improved classification performance.

With the advancement of deep learning techniques, particularly convolutional neural networks (CNNs), image-based lithology identification has seen significant improvement. Zhang et al. (2017) employed CNNs to perform lithology recognition on a grayscale image dataset comprising 1500 images across three lithology classes: sandstone, shale, and conglomerate, achieving a recognition accuracy of 95%. Baraboshkin et al. (2020) preprocessed core images from various regions in Russia to create a lithology dataset consisting of six classes: blocky sandstone, layered sandstone, limestone, granite, shale, and siltstone, totaling 20,000 images. They tested multiple CNN models on this dataset, with the best model achieving a recognition accuracy of 72%. Alzubaidi et al. (2021) collected a dataset comprising 76,500 images across four lithology classes—sandstone, limestone, shale, and debris (representing non-core material)—from 28 boreholes in Australia. They tested ResNet, Inception-v3, and ResNeXt models, achieving a prediction accuracy of 93.12%. Fu et al. (2022) constructed a lithology dataset with 10 classes and 15,000 images based on publicly available core image databases. They compared the performance of ResNet, ResNeSt, DenseNet, and VGG architectures, achieving a maximum recognition accuracy of 99.60%.

Although existing research has achieved impressive results in lithology identification using image data, each study employs a distinct lithology dataset, making direct comparison of results difficult. Moreover, many of these datasets are not publicly available, and even those built from open sources often require extensive image preprocessing, which limits their accessibility. In contrast to lithology identification based on well logging datasets, the construction of publicly available core image datasets is of significant importance for standardized evaluation and comparison of lithology identification methods.

In this work, we collected a large number of drill core tray images from publicly available borehole image databases and meticulously curated and cropped them to construct a large-scale, high-quality, open-access Drill Core Image Dataset (DCID). The dataset comprises two versions: DCID-7 and DCID-35. The DCID-7 dataset includes 7 lithological categories, each containing 4000 training samples and 1000 test samples, resulting in a total of 35,000 images. The DCID-35 dataset comprises 35 lithological categories, each with 800 training samples and 200 test samples, also totaling 35,000 images. The dataset features a large volume and a wide range of lithological diversity, with all images having a high resolution of 512 × 512 pixels, making it a solid foundation for lithology identification model training. Additionally, we provide an automated downsampling tool that converts the original high-resolution images into 256 × 256, 128 × 128, 64 × 64, and 32 × 32 pixel formats to accommodate different task requirements. Furthermore, by incorporating 28,000 slightly defective images, the dataset can simulate real-world application scenarios and enhance model robustness through data augmentation. Based on DCID, we systematically investigated the impact of various factors—such as model architectures, model sizes (parameter counts), image resolutions, lighting conditions, and levels of real-world data augmentation—on the performance and robustness of lithology identification in simulated real-world environments. Finally, we demonstrated how low-resolution datasets can be effectively used to rapidly evaluate models across multiple dimensions, including identification accuracy, training time, inference speed, storage efficiency, and robustness.

2. Dataset construction

2.1. Data and label acquisition

The Drill Core Image Dataset (DCID) was developed using core image data from various drill wells in South Australia (Geological Survey of South Australia). The dataset was curated to ensure geological diversity and representativeness across lithological

environments. Core images were sourced from wells with hyLogger spectral scanner data and geologist-provided lithological annotations. Preserved in trays, these samples were scanned, and labels were assigned based on depth-specific lithological descriptions.

Fig. 1 illustrates this process with red sandstone images from the IHAD2 well, collected between 611.4 m and 617.5 m. Geologists classified the 560–636 m interval as "red sandstone", and all images within this range were labeled accordingly. To enhance dataset robustness, images from additional wells within the same lithological category were included when needed. Data were collected from diverse wells (e.g., WJD1, SLT101, IHAD2, etc.) spanning varied geological settings to ensure comprehensive lithofacies coverage.

Raw core tray scans are hosted on the SARIG platform by the Geological Survey of South Australia, providing access to images and lithological descriptions. The DCID includes 35 lithological categories (e.g., sandstone, siltstone, granite) based on geologist interpretations. Sourced from multiple wells in the same region across various depths, the dataset reflects diverse geological environments, with labels aligned to expert annotations, ensuring applicability for lithology identification tasks.

2.2. Data processing

As shown in Fig. 2, constructing the final dataset from core tray images involves three key steps: image cropping, defective sample processing, and downsampling.

2.2.1. Image cropping

After acquiring drill core tray images, the next step is image cropping, which significantly affects lithology identification models. Alzubaidi et al. (2021) investigated cropping sizes (60 × 60, 120 × 120, 180 × 180, and 240 × 240 pixels), finding that smaller images capture fewer lithological features, while larger ones improve performance by including more detail. Due to lateral size constraints in the collected core tray images, this study standardized the cropped size to 512 × 512 pixels to ensure high-resolution capture of lithological features.

2.2.2. Defective sample processing

Images from automated cropping include adverse samples (e.g., fractures, missing sections, artificial marks, impurities), termed "defective samples", which impair lithology identification model training. Two main approaches address these: one boosts recall, the other precision. The first adds a "garbage" class for defective images (Alzubaidi et al., 2021), enhancing recall while preserving data integrity, though it may reduce lithology classification accuracy. The second filters out defective images entirely, training on clean data to improve precision but potentially reducing real-world robustness (Fu et al., 2022).

This study addresses defective samples by retaining some slightly defective ones for data augmentation. During dataset construction, flawless samples are first screened to form the baseline dataset. Defective samples are then filtered based on their lithological feature proportion, with a threshold of 40%. Samples with at least 40% lithological features—calculated as the proportion of pixels corresponding to the target lithology:

$$\frac{\text{Number of lithological pixels}}{\text{Total number of pixels in the image}} \geq 0.4 \tag{1}$$

2.2.3. Down sampling

Multiscale resolution significantly impacts model training. High resolution enhances pattern and texture recognition by providing detailed information but demands more computational resources and slows training. Lower resolution accelerates training and reduces resource use but may compromise detail, impacting performance on complex tasks. In lithology identification, resolution choice depends on the balance between accuracy and training speed. During dataset construction, we downsampled original 512 × 512 pixel high-resolution images to 256 × 256, 128 × 128, 64 × 64, and 32 × 32 pixels, enabling selection of the optimal resolution for specific needs, balancing accuracy and efficiency.

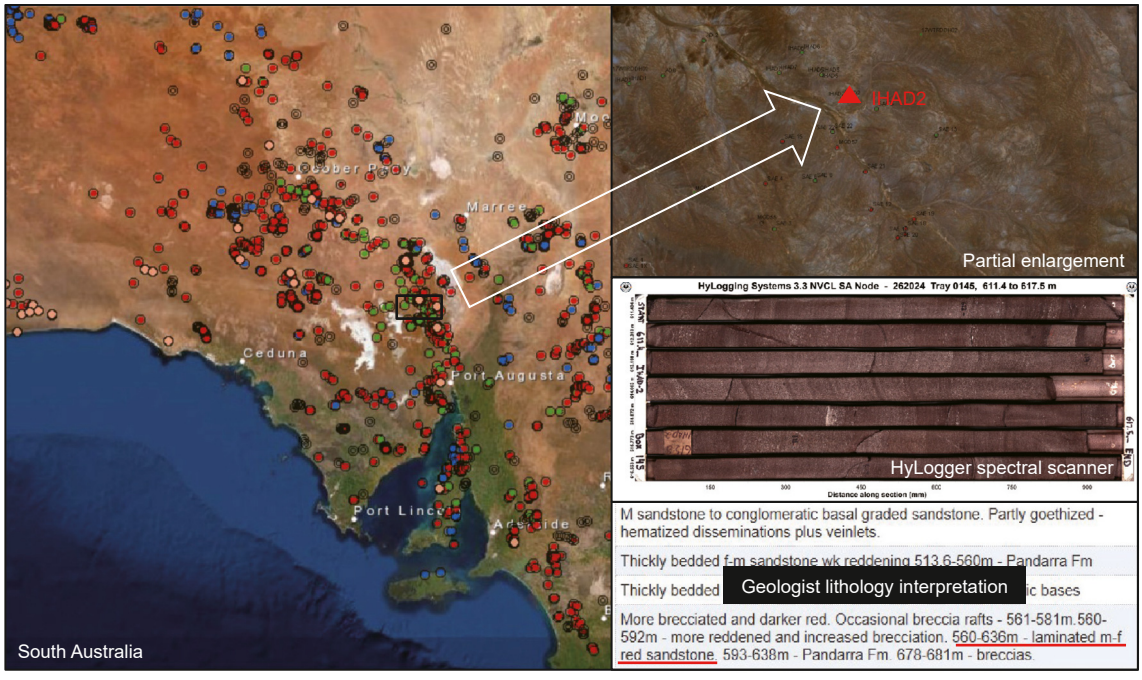


Fig. 1. Schematic diagram of data acquisition (data comes from South Australia, taking the red sandstone of IHAD2 well as an example) (Government of South Australia—Department for Energy and Mining).

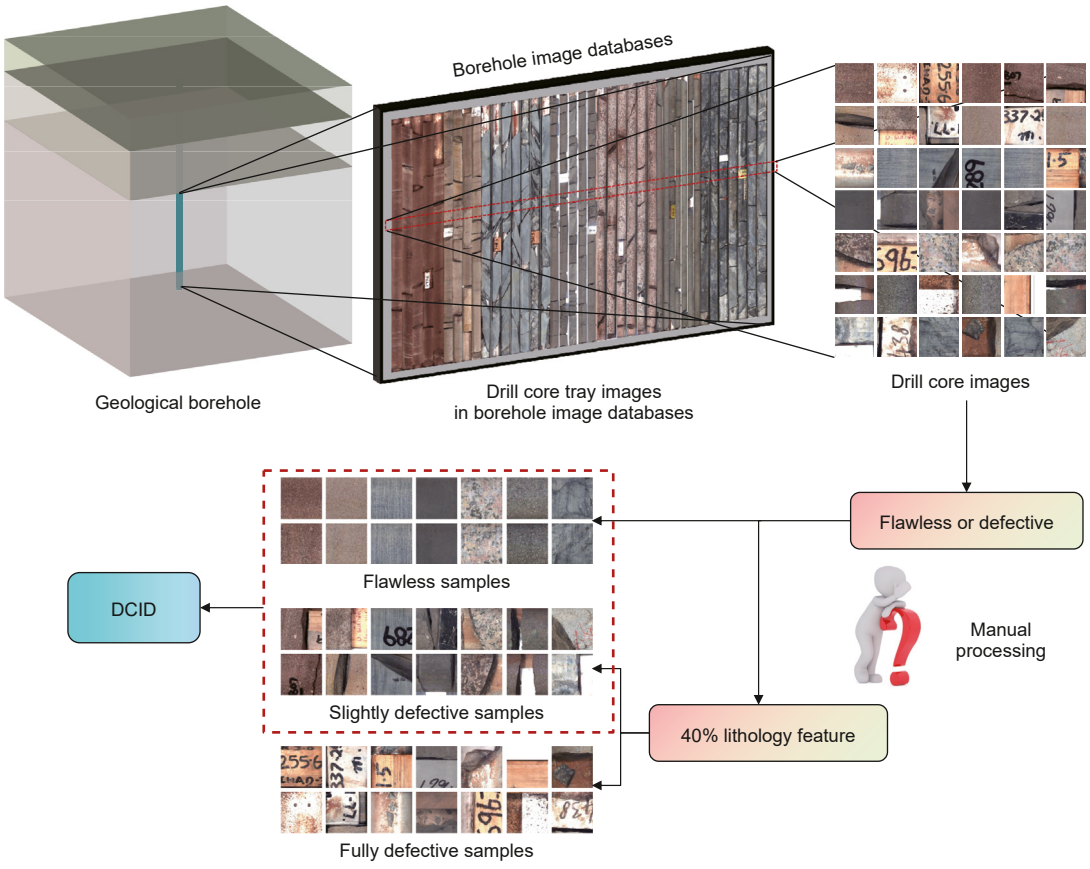


Fig. 2. Dataset construction flow chart: The process begins by acquiring drill core tray images from the geological borehole image database. These images are cropped to extract individual drill core images. Based on the condition of the images (flawless or defective), manual processing is performed. The final DCID dataset consists of two parts: flawless samples and slightly defective samples.

2.3. Drill core image dataset (DCID)

The original Drill Core Image Dataset (DCID) comprises two versions: DCID-7 and DCID-35. This design is inspired by the widely used CIFAR dataset, a benchmark in computer vision research for image classification tasks, which also includes two versions: CIFAR-10 and CIFAR-100. Similarly, the version with fewer categories in our dataset (7 categories) contains more samples per class (5,000), making it well-suited for evaluating the upper bounds of model accuracy in lithology identification. In contrast, the version with more categories (35 categories) has fewer samples per class (1,000), enabling the assessment of model performance in a more complex, fine-grained classification scenario. This dual-version structure offers a balanced framework for evaluating different aspects of model generalization and robustness. As shown in Fig. 3, the DCID-7 dataset comprises 35,000 (512 × 512 pixels) colored core images across 7 categories, with 5000 images per category. Each category is divided into 4000 training images and 1000 testing images in an 8:2 ratio. As shown in Fig. 4, the DCID-35 dataset consists of 35,000 (512 × 512 pixels) colored core images in 35 categories, with 1000 images per category. Each category is divided into 800 training images and 200 testing images in an 8:2 ratio.

The expanded Drill Core Image Dataset (DCID) is named according to the format DCID-R-C-L-I, as illustrated in Fig. 5. In this naming convention, “R” represents the resolution of images in the dataset, with optional parameters of 32, 64, 128, 256, and 512 pixels. “C” indicates the number of categories, with optional parameters of 7 and 35, corresponding to the same meanings as in the original dataset. “L” denotes levels of real-world data augmentation (RWDA), with optional parameters ranging from 0 to 0.4. These RWDA levels are defined as the proportion of slightly defective samples introduced into the dataset relative to the original number of samples:

$$L = \frac{N_{\text{defective}}}{N_{\text{total}}} \tag{2}$$

where $N_{\text{defective}}$ is the number of defective samples, and N_{total} is the total number of samples in the dataset. “I" denotes the index for none (N), train dataset (T), test dataset (E), and all dataset (A), indicating that RWDA is applied to the selected portion of the dataset. Applying RWDA to the test dataset can be understood as simulating the presence of defective data in real-world environments. Customized versions of the dataset can be generated from the original data through a preprocessing program, based on the selected combination of parameters. A detailed list of lithology categories is provided in Table 1. The dataset is publicly available on GitHub and Hugging Face, and can be accessed by searching for “DCID” on either platform.

3. Methodology

3.1. Convolutional neural networks

Convolutional neural networks (CNNs), inspired by the hierarchical processing of the visual cortex, are deep learning models designed to automatically learn spatial hierarchies of features from input data (LeCun et al., 2015). A CNN typically consists of

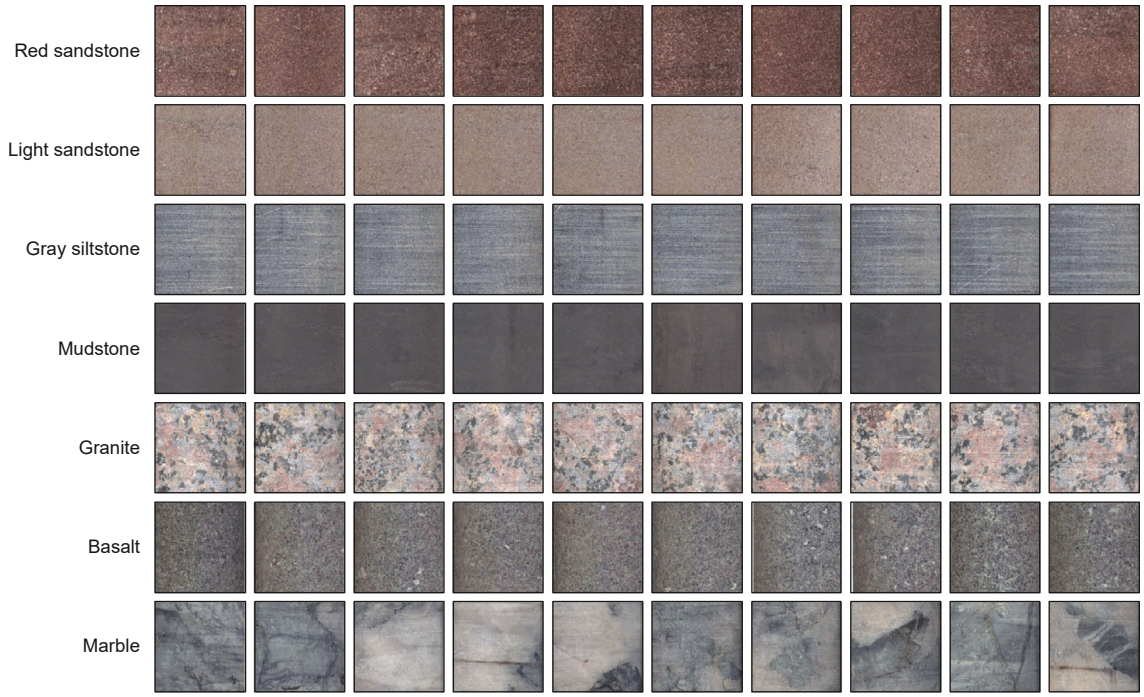


Fig. 3. Representative image samples from the DCID-7 dataset. From top to bottom: red sandstone, light sandstone, gray siltstone, mudstone, granite, basalt, and marble.

convolutional layers, nonlinear activation functions, and pooling layers. The 2D convolution operation can be expressed as

$$(f*g)(i,j) = \sum_m \sum_n f(m,n) \cdot g(i-m, j-n) \tag{3}$$

where f is the input image or feature map, g is the convolution kernel, (i, j) are the pixel coordinates of the output feature map. m and n are the indices of the convolution kernel. CNNs have demonstrated exceptional performance in image processing, computer vision, and pattern recognition.

3.1.1. VGG models

VGG is a family of convolutional neural network models developed by the Visual Geometry Group at the University of Oxford, introduced in 2014 (Simonyan and Zisserman, 2014). Designed to evaluate the impact of network depth on image classification performance, VGG employs a simple yet effective architecture based on sequential 3 × 3 convolutional kernels and max-pooling layers. As illustrated by VGG-11 in Fig. 6(a), the model stacks multiple convolutional layers followed by fully connected layers to progressively extract hierarchical features. Despite being relatively large and computationally intensive, VGG's standardized design and strong performance have made it a foundational benchmark in deep learning for image processing.

3.1.2. ResNet models

ResNet, introduced by He et al. (2016), is a deep residual network designed to address training challenges in deep architectures, particularly the issues of vanishing and exploding gradients. As illustrated by ResNet-18 in Fig. 6(b), the core innovation lies in the use of residual blocks, which incorporate skip connections to enable identity mappings:

$$\text{output} = \text{input} + F(\text{input}) \tag{4}$$

where "input" is the input to the residual block, "F(input)" is the mapping function within the residual block, representing the transformed output of the input, and "output" is the output of the residual block. These skip connections facilitate gradient flow across layers, enabling the training of much deeper networks and improving both convergence and generalization.

ResNet has achieved outstanding results across various vision tasks, including image classification, object detection, and semantic segmentation. Its architecture has become a cornerstone in deep learning, influencing many subsequent model designs.

3.1.3. Densenet models

DenseNet, proposed by Huang et al. (2017), is a deep convolutional neural network characterized by dense connectivity, where each layer receives the concatenated outputs of all preceding layers. As shown in DenseNet-121 in Fig. 6(c), its architecture is built upon dense blocks, which promote efficient feature reuse and improve gradient flow:

$$\text{output} = H(\text{input}_1, \text{input}_2, \text{input}_3, \ldots \text{input}_{k-1}) \tag{5}$$

where H denotes a series of convolutional layers within the dense block, input_1, input_2, input_3, input_3, … input_{k-1} represent the outputs of all previous layers, and "output" is the output of the current layer. This design mitigates the vanishing gradient problem, enhances feature propagation, and reduces the number of parameters compared to traditional deep networks.

By integrating dense connections with standard components like batch normalization and activation functions, DenseNet achieves strong performance and stability across various image recognition tasks.

3.1.4. MobileNet models

MobileNet, introduced by Google (Howard, 2017), is a lightweight CNN architecture designed for efficient inference in resource-constrained environments such as mobile and embedded devices. As illustrated by MobileNet_v2 in Fig. 6(d), it employs depthwise separable convolutions, which split standard convolutions into depthwise and pointwise operations. This significantly

Fig. 4. Representative image samples from the DCID-35 dataset.

reduces computational cost and parameter count while maintaining competitive accuracy.

MobileNet also incorporates lightweight convolutional kernels and global average pooling to further reduce complexity. Its compact design and low-latency inference make it well-suited for edge deployment, influencing a wide range of mobile deep learning applications.

3.2. Vision Transformer

The Vision Transformer (ViT), proposed by Dosovitskiy (2020), adapts the Transformer architecture—originally designed for natural language processing—to image classification tasks. As shown in Fig. 7(a), ViT divides an input image into fixed-size, non-overlapping patches, flattens each patch into a 1D vector, and treats the resulting sequence as input tokens to a standard Transformer encoder. The attention operation is formulated as

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. This mechanism enables the model to compute a weighted sum of the value vectors V based on the similarity between the query and key vectors, allowing it to focus on the most relevant information across the entire image.

By learning relationships between distant regions, ViT achieves strong performance in image recognition tasks, including lithology identification, particularly where global spatial context is crucial.

3.3. MLP-Mixer

The MLP-Mixer, proposed by Tolstikhin et al. (2021), is a lightweight and efficient architecture that replaces self-attention with multi-layer perceptrons (MLPs) applied along both spatial and channel dimensions. As illustrated in Fig. 7(b), unlike the Vision Transformer (ViT), which models patch dependencies via self-attention, the MLP-Mixer uses alternating MLP layers to perform spatial mixing and channel mixing, enabling it to capture both local and global patterns:

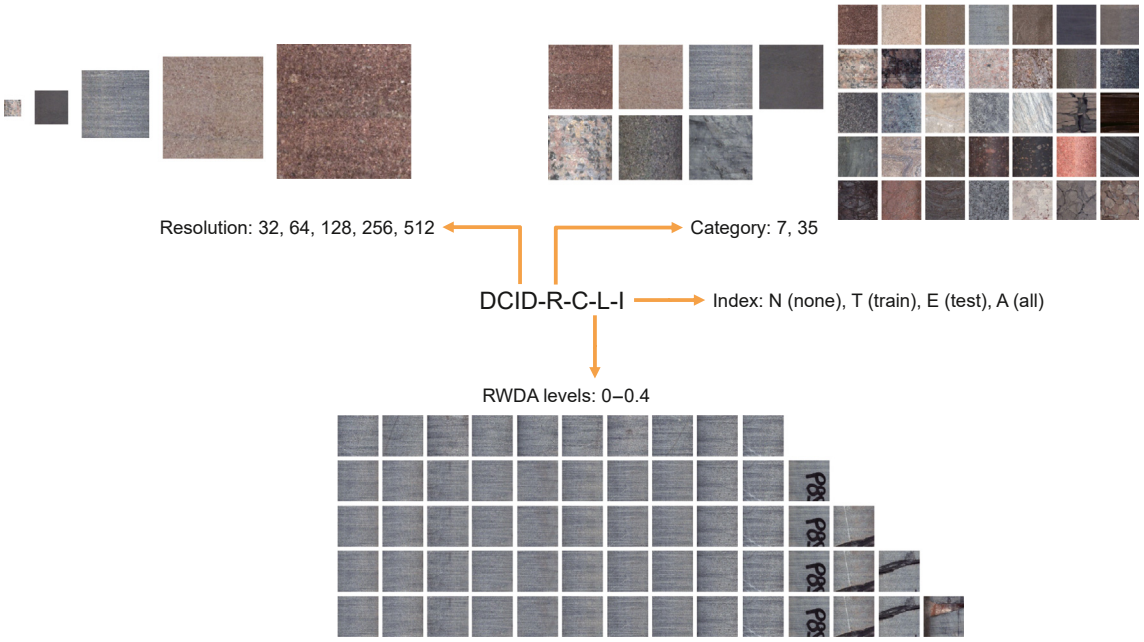


Fig. 5. The expanded Drill Core Image Dataset (DCID). The format DCID-R-C-L-I represents image resolution (R), number of categories (C), real-world data augmentation level (L), and image index (I).

Table 1
Description of lithology categories in the DCID dataset.

Category	Description
Red sandstone	Fine to medium-grained, containing altered red sandstone and mottled sandstone
Light sandstone	Fine to medium-grained, light to medium brown, with lithic grains, clay pellets, cross-bedded, interbedded shale layers
Gray-green sandstone	Fine to medium-grained, pale gray-green to green, iron-oxide staining, feldspathic and lithic fragments
Gray siltstone	Fine-grained gray siltstone forming the basement, minor magnetite and pyrite
Brown siltstone	Fine-grained brown siltstone, prominent carbonate-chlorite veins, parallel to bedding, reticulate pattern
Carbonaceous siltstone	Fine-grained, blackish-gray carbonaceous siltstone with coal
Mudstone	Rich in carbonaceous material, soft, dark gray
Granite	Coarse-grained, with veins of hornblende, alkali feldspar, feldspar crystals, and metadolerite
Megacrystic granite	Medium to very coarse-grained, grayish red, with megacrysts, hornblende, and biotite
Monzogranite	Medium-grained biotite granite, equigranular texture, reddish-orange, with biotite and muscovite
Fine-grained granite	Fine-grained, with hematite veinlets and large clasts of quartz
Pegmatite	Fine to coarse-grained, pegmatoidal, slight foliation, with saussurite, chlorite, oxidized hematite
Basalt	Amygdaloidal to massive, altered by chlorite, sericite, hematite, amphibole, epidote, K-feldspar
Fine-grained basalt	Fine-grained, dark green, with plagioclase and pyroxene
Gabbro	Coarse-grained, brecciated in parts, with minor disseminated fine pyroxene
Dolerite	Fine to coarse-grained, greenish-gray, with clinopyroxene and olivine
White marble	Calcitic marble, with potassium feldspar, quartz, diopside, calcite, ankerite, white to yellow
Blue marble	Diopside-bearing marble, with diopside, calcite, magnetite, interspersed with felsic rock
Dolomite	White and pale green banded, with chlorite, coarse quartz, feldspar, occasional augen structures
Black shale	Flaggy to massive, gray to black, interbedded with dolomite and siltstone
Red shale	Thin multiple beds, deep red, with white bleaching layers
Dark green calc-silicate	Brecciated tremolite-diopside calc-silicate, very dark green, variable black chlorite/biotite blotches
Brown calc-silicate	Garnet-dominant metasomatite, with minor hematite, biotite, and actinolite
Skarn	Massive cpx-kspar calc-silicate skarn, medium dark green, with large tremolite grains
Red-brown dacite	Massive to flow-banded, amygdaloidal, red-brown, amygdales with quartz, zoned quartz, amorphous silica
Black dacite	Massive, amygdaloidal, black, filled with quartz and fluorite
Rhyolite	Volcaniclastic, red-brown, with free quartz phenocrysts and high-lustre mica minerals
Banded iron formation	Well-laminated, dark gray to black, rich in magnetite
Hematite breccia	Dense and porous, some clasts brecciated, supported by red and gray hematite matrix
Altered breccia	Pink-red altered breccia, with minor sulphides in hematite-rich fractures and veins
Schist	Felsic schist, hematite-coated feldspars, minor silicification
Gneiss	Fine-grained, quartz-feldspar-biotite psammitic felsic gneiss
Porphyry	Fine to coarse-grained, grayish orange-pink, moderately to highly weathered
Brecciated mudstone	Detrital to brecciated, irregular fractures, uneven particle distribution, dark gray
Shale fragments	Intermingled red and black shale fragments

$$\begin{aligned}\mathbf{X}_{\text{spatial}} &= \text{MLP}_1(\mathbf{X}) \\ \mathbf{X}_{\text{channel}} &= \text{MLP}_2(\mathbf{X})\end{aligned} \tag{7}$$

where $\mathbf{X}$ is the input image patches, and MLP_1 and MLP_2 are the MLPs applied to the spatial and channel dimensions, respectively. This mixing is repeated across multiple layers to progressively refine feature representations.

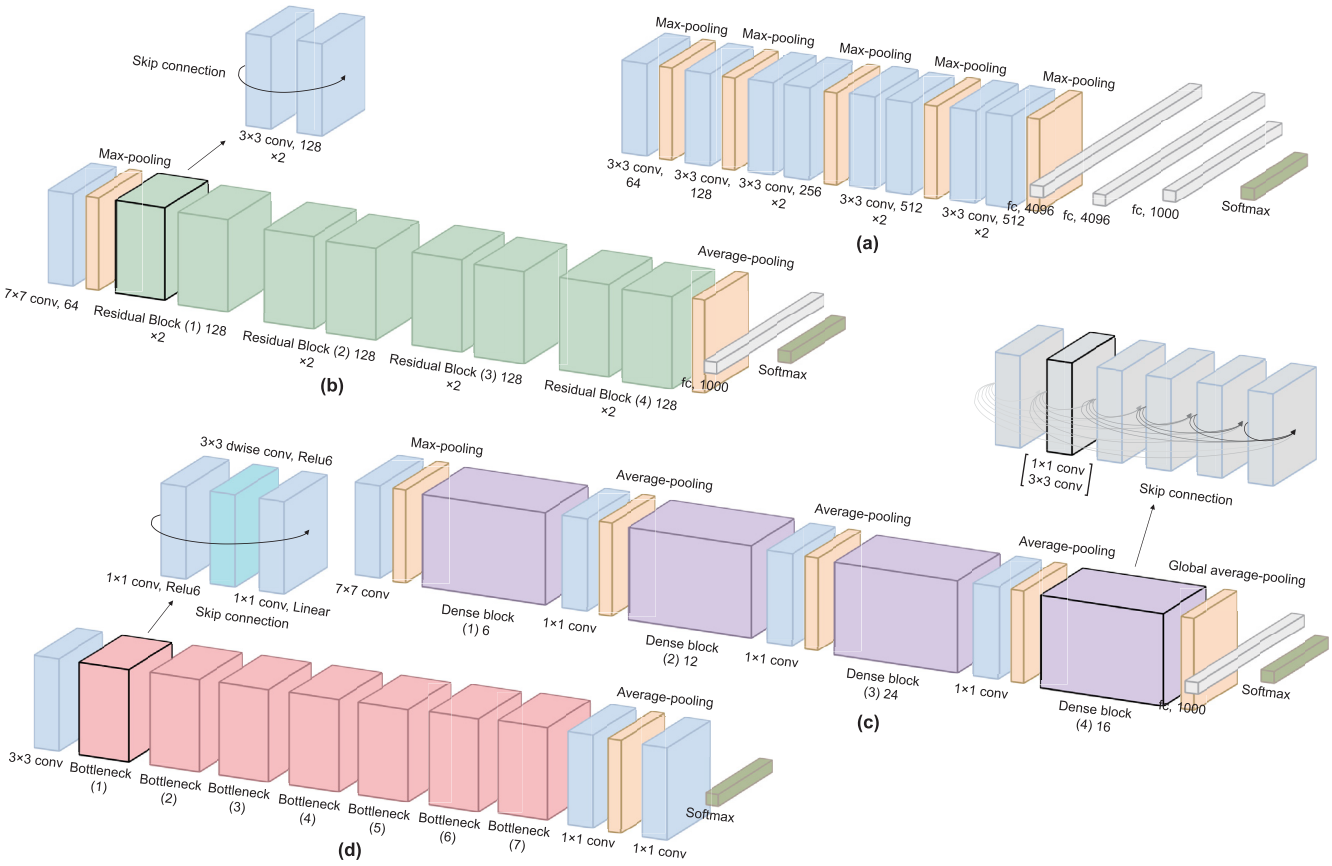


Fig. 6. Simplified sketches of convolutional neural networks. **(a)** VGG-11; **(b)** ResNet-18; **(c)** DenseNet-121; **(d)** MobileNet_v2.

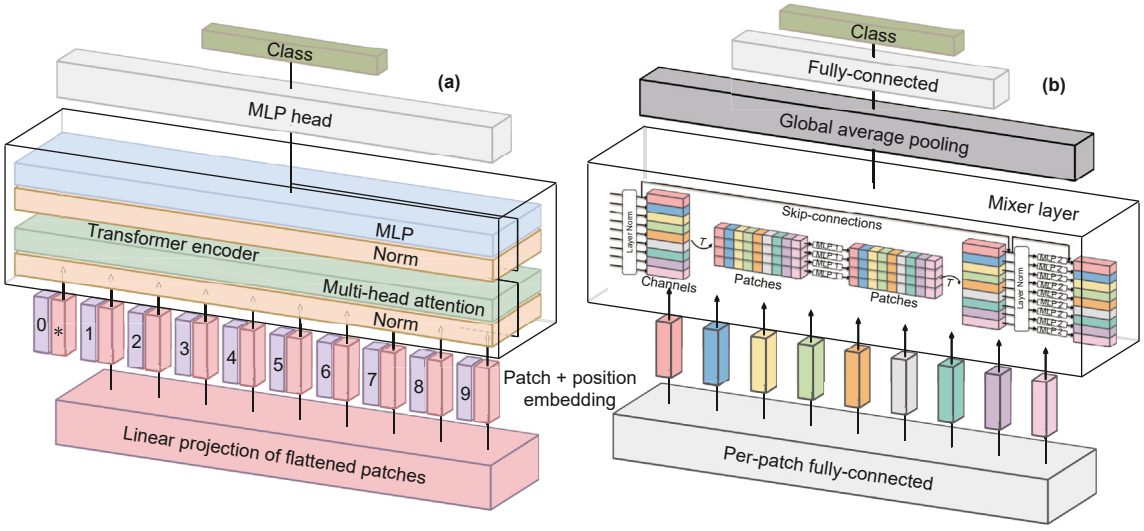


Fig. 7. Simplified sketches of Vision Transformer and MLP-Mixer. **(a)** Vision Transformer; **(b)** MLP-Mixer.

The MLP-Mixer achieves competitive image classification performance while offering lower computational complexity than ViT, making it well-suited for tasks like lithology identification in large-scale datasets where efficiency is critical.

3.4. Real-world data augmentation

Data augmentation is a widely adopted technique to increase dataset size and diversity, thereby improving model generalization and performance (Shorten and Khoshgoftaar, 2019). In lithology recognition, it plays a vital role by enabling models to learn more robust and generalized lithological features across varied geological conditions.

As illustrated in Fig. 8(a), traditional augmentation techniques—such as occlusion, cropping, rotation, and color transformations—generate synthetic variations of original data. While helpful, these methods are limited in simulating the complex noise and artifacts present in real-world scenarios.

To address this, we introduce real-world data augmentation (RWDA), shown in Fig. 8(b). RWDA leverages slightly defective images collected during the drilling core acquisition process, incorporating artifacts such as markings, cracks, incompleteness, and impurities. Compared to synthetic augmentations, RWDA offers greater authenticity and variability, thereby enhancing the model's robustness and generalization in practical applications.

3.5. Details of training

All experiments were conducted on an NVIDIA GeForce RTX 4080 GPU. Models were implemented in PyTorch with default settings to ensure consistency, and OpenCV was used for image preprocessing. To enable fair evaluation, no pre-trained weights were used during training, eliminating potential bias from prior learning and allowing for accurate performance comparison across methods.

3.5.1. Loss function and optimization algorithm

During training, we employed the cross-entropy loss function, a standard choice for classification tasks due to its effectiveness in image classification. It is defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\ln\left(\hat{y}_{ij}\right) \tag{8}$$

where y_{ij} denotes the ground truth label, $\hat{y}_{ij}$ is the predicted probability of class j for sample i, N is the number of samples, and C is the number of classes.

All models were optimized using the Adam optimizer (Kingma, 2014), which adaptively adjusts learning rates based on first- and second-order moment estimates of gradients:

$$\theta_{t+1} = \theta_t - \eta\cdot\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \tag{9}$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, η is the learning rate, and ϵ is a small constant for numerical stability. Adam combines the benefits of momentum and adaptive learning rate strategies, resulting in faster and more stable convergence.

3.5.2. Learning rate scheduler and batch size

We adopted a cosine annealing learning rate scheduler combined with a warm-up strategy (Jacobs, 1988) to enhance model convergence. In cosine annealing, the learning rate η_t varies over epochs according to

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right) \tag{10}$$

where $\eta_{\min}$ and $\eta_{\max}$ denote the minimum and maximum learning rates, T_{cur} is the current epoch, and $T_{\max}$ is the total number of epochs.

To stabilize early-stage training, a linear warm-up was applied during the first 5 epochs, increasing the learning rate from 0.0001 to 0.001. From epoch 6 to 50, the cosine annealing strategy was used to gradually reduce the learning rate to 0.00001.

The batch size is also critical in training performance. Larger batches accelerate computation but demand more memory, while smaller batches offer better generalization but slower convergence. Considering the trade-offs between computational resources and model complexity, we used a batch size of 64 for all experiments to ensure both stability and efficiency.

3.5.3. Model evaluation

Evaluating model performance is essential in lithology recognition. In this study, we adopt two evaluation levels: overall performance and per-class performance (Raschka, 2018). The overall performance is assessed using four standard classification metrics: accuracy, precision, recall, and F1 score:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\\
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}}\\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}\\
\text{F1 score} &= 2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}
\end{aligned} \tag{11}$$

where, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

For per-class performance, we employ the confusion matrix, which compares predicted and actual labels across all lithology classes. Diagonal elements represent correctly classified instances, while off-diagonal elements indicate misclassifications. This matrix provides an intuitive visualization of class-wise recognition accuracy and highlights model performance variations across lithologies.

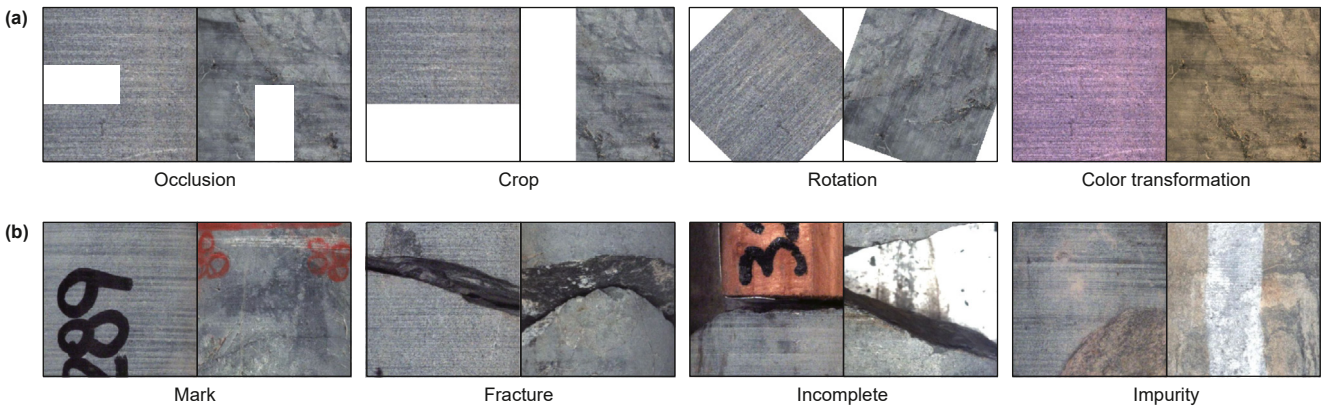


Fig. 8. Data augmentation. **(a)** Traditional data augmentation; **(b)** real-world data augmentation.

4. Experiments

4.1. Base case

In this study, we selected ResNet-18 as the baseline model and trained it on two foundational datasets: DCID-7 (DCID-256-7-0-N) and DCID-35 (DCID-256-35-0-N). Fig. 9 illustrates the loss and accuracy curves during training and testing. For DCID-7, both training and testing losses decrease significantly with increasing epochs, with the testing loss stabilizing at a low level, suggesting good generalization without evident overfitting. Training accuracy rises rapidly to nearly 1.0, indicating effective learning of the training data, while testing accuracy, though fluctuating, remains high (close to 1.0). These fluctuations may stem from sample diversity or batch difficulty in the test set.

For DCID-35, the testing loss converges toward the training loss in later training stages, reflecting strong generalization. Training accuracy increases steadily but slightly declines after peaking, hinting at potential overfitting. Testing accuracy remains stable with minor fluctuations and an overall upward trend. Notably, loss fluctuations are greater for DCID-35 than DCID-7, likely due to the increased complexity of the 35-category dataset, requiring more nuanced model adjustments.

Fig. 10(a) shows the confusion matrix for ResNet-18 on DCID-7, with strong performance across most classes (diagonal values near or at 1000, indicating high accuracy). Per Table 2, "Gray siltstone" had one sample misclassified as "Marble". For "Marble", misclassifications included three samples as "Basalt", one as "Gray siltstone", and one as "Mudstone". Fig. 10(b) presents the DCID-35 confusion matrix, with 200 samples perfectly classified, reflecting good accuracy. Table 2 indicates misclassifications: one "Pegmatite" as "Red sandstone", one "Gabbro" and one "Dark green calc-silicate" as "Skarn", one "Gray siltstone" as "Fine-grained basalt", one "Granite" as "Fine-grained granite", and one "Brecciated mudstone" as "Mudstone". Additionally, two "Mudstone" samples were misclassified as "Brecciated mudstone".

4.2. Comparison of model architectures

We compared six widely used model architectures for lithology recognition: VGG-19, ResNet-18, DenseNet-121, MobileNet_v2, ViT-Patch16, and Mixer-B16. These models, common in image classification, were tested on the DCID-7 and DCID-35 datasets with consistent training parameters. Performance was assessed via training and testing loss and accuracy.

Fig. 11(a)–(d) shows results for DCID-7. Training loss decreases across all models with epochs, though VGG-19 exhibits the slowest decline, suggesting a lower learning rate. DenseNet-121, MobileNet_v2, ViT-Patch16, and Mixer-B16 maintain stable loss curves, while all models achieve near-1.0 training accuracy, with DenseNet-121 and MobileNet_v2 slightly outperforming others. Testing loss is lowest and most stable for DenseNet-121; ResNet-18, MobileNet_v2, and ViT-Patch16 show larger, diminishing fluctuations. VGG-19's higher initial testing loss stabilizes but remains

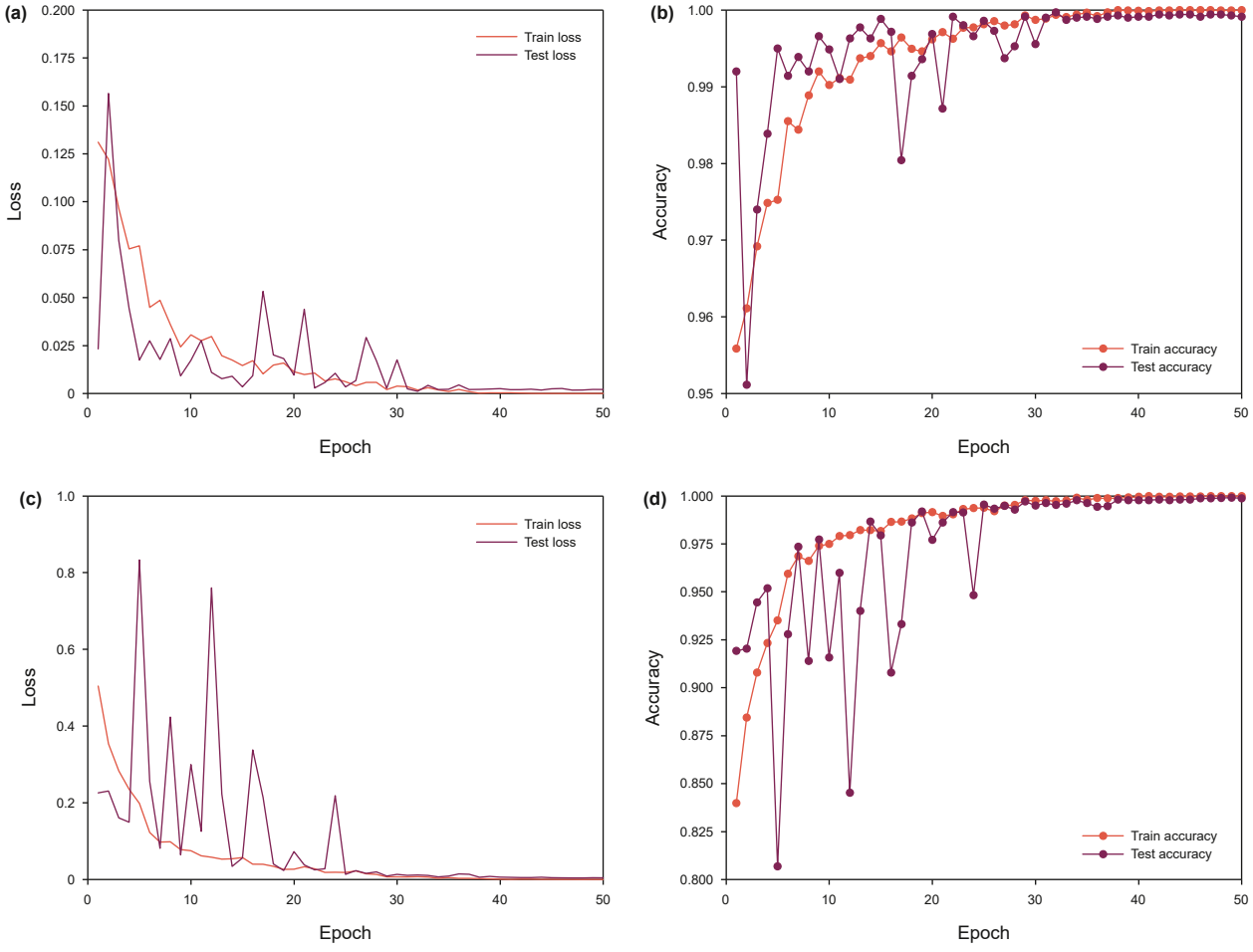


Fig. 9. Performance of ResNet-18 on DCID-7 and DCID-35. **(a)** Loss on DCID-7; **(b)** accuracy on DCID-7; **(c)** loss on DCID-35; **(d)** accuracy on DCID-35.

elevated, indicating weaker generalization. Testing accuracy is high across models, with MobileNet_v2 converging fastest and VGG-19 fluctuating most.

Fig. 11(e)–(h) present DCID-35 results, mirroring trends. Training loss drops quickly, with VGG-19 again slowest and DenseNet-121 achieving the lowest, smoothest descent. Training accuracy nears 1.0, though VGG-19 improves most slowly. Testing loss fluctuates more than on DCID-7, especially for MobileNet_v2, ResNet-18, and ViT-Patch16; VGG-19's remains high, reflecting poor generalization. Testing accuracy, though more variable, reaches high levels, with VGG-19 lagging slightly. ViT-Patch16 and Mixer-B16 show stable accuracy, with Mixer-B16 slightly outperforming some traditional models.

Fig. 12 compares accuracy across datasets. On DCID-7, all models excel, with DenseNet-121 and MobileNet_v2 hitting 1.0. On DCID-35, accuracy drops as categories increase, with ResNet-18 showing the smallest decline and VGG-19 the largest. All models retain strong performance, with ViT-Patch16 matching DenseNet-121 and MobileNet_v2, and surpassing VGG-19.

4.3. Comparison of model sizes

The performance of neural networks is closely influenced by model size. While deeper models with more parameters generally capture complex patterns more effectively, they also risk overfitting and demand greater computational resources. To explore this trade-off, we evaluated ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152 on the DCID-256-7-0-N and DCID-256-35-0-N datasets.

As shown in Fig. 13(a)–(d) and Fig. 13(e)–(h), all models exhibit steadily declining training loss. Although deeper models (e.g., ResNet-152) start with higher loss, they eventually reach similar accuracy levels as shallower ones. However, deeper networks show larger fluctuations in test loss, especially in the more complex 35-class dataset, suggesting a greater need for regularization.

All models achieve >99.7% accuracy, highlighting the strong feature extraction ability of ResNet architectures. Yet, as illustrated in Fig. 14, training cost increases substantially with model depth. ResNet-152, for example, requires over twice the training time and five times the memory of ResNet-18, with only minor accuracy gains. These diminishing returns are critical in scenarios with limited computational resources. ResNet-50 emerges as the most balanced option, offering competitive accuracy with moderate training time and memory usage—an effective compromise between performance and efficiency.

4.4. Impact of image resolution on model performance

Choosing an appropriate image resolution is critical in lithology recognition, as it affects both computational cost and model accuracy. High-resolution images (e.g., 256 × 256) may improve feature representation but increase training time and risk overfitting, while low-resolution images reduce resource demands but may lose critical information. To investigate this trade-off, we trained ResNet-18 on datasets with four resolutions: 256 × 256 (DCID-256-7/35), 128 × 128 (DCID-128-7/35), 64 × 64 (DCID-64-7/35), and 32 × 32 (DCID-32-7/35), keeping all other training parameters fixed.

As shown in Fig. 15(a)–(d) and Fig. 15(e)–(h), higher-resolution models started with lower initial losses and showed faster convergence. Testing accuracy was also positively correlated with resolution, particularly in the more complex 35-category dataset, where the performance gains from high-resolution inputs were more pronounced. These findings confirm that higher resolution provides richer lithological details, enhancing learning effectiveness. However, models trained on high-resolution images exhibited more pronounced fluctuations in test loss and accuracy, indicating a greater risk of overfitting, likely due to the inclusion of noise along with informative features.

As shown in Fig. 16, lower-resolution datasets significantly reduced training time, processing time per image, and total computation time. The computational savings were substantial between 256 × 256 and 64 × 64 but became less significant between 64 × 64 and 32 × 32. Dataset size also decreased dramatically with resolution, reducing storage and data transfer costs. Lower resolutions offer clear computational benefits with only a slight accuracy trade-off, especially in 7-class tasks. However, for more complex tasks involving fine-grained classification, higher resolutions are more advantageous despite their greater resource demands.

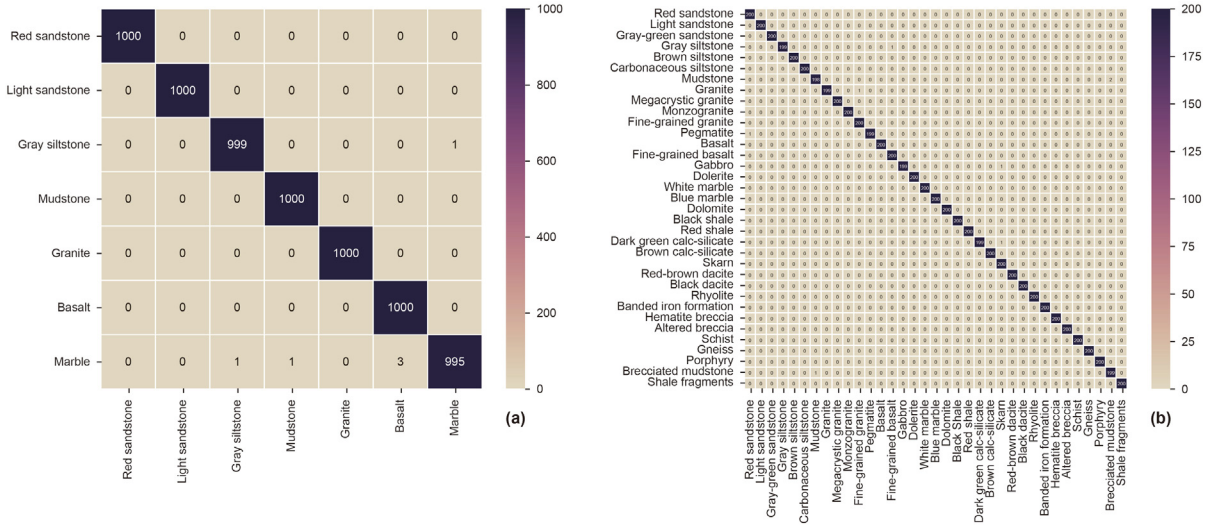
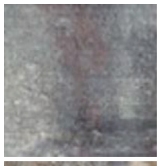
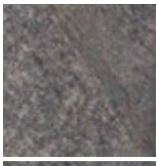
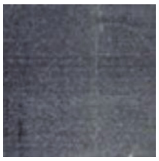
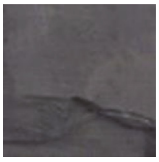


Fig. 10. Confusion matrices of ResNet-18 on DCID-7 and DCID-35. **(a)** DCID-7; **(b)** DCID-35.

Table 2
Misclassified samples in the base case.

DCID-7			DCID-35		
Image	True	Predicted	Image	True	Predicted
	Gray siltstone	Marble		Pegmatite	Red sandstone
	Marble	Basalt		Gabbro	Skarn
	Marble	Basalt		Dark green calc-silicate	Skarn
	Marble	Basalt		Brecciated mudstone	Mudstone
	Marble	Gray siltstone		Gray siltstone	Fine-grained basalt
	Marble	Mudstone		Mudstone	Brecciated mudstone
				Mudstone	Brecciated mudstone
				Granite	Fine-grained granite

4.5. Validation of real-world data augmentation

To evaluate the impact of real-world data augmentation (RWDA) on model robustness in lithology identification, we conducted targeted experiments by introducing slightly defective samples into the training data. This simulates real-world imperfections such as image artifacts, markings, or incompleteness. Experiments were performed on both high-resolution (256×256) and low-resolution (32×32) datasets to eliminate resolution bias. We first injected 40% defective samples into the testing sets (DCID-256/32–7/35-0.4-E) to mimic noisy real-world environments. Then, we trained models on datasets with varying RWDA levels (0%–40%, in 5% increments; DCID-256/32–7/35-L-T) and evaluated them against the fixed 40% RWDA test set.

As shown in Fig. 17(a) and (d), while training loss steadily decreased across all configurations, testing loss fluctuated significantly and showed little improvement without RWDA, indicating poor generalization in noisy conditions. Similarly, Fig. 17(b) and (e) show high training accuracy but notably lower and unstable testing accuracy, especially in the more complex 35-class dataset. This suggests that models trained without RWDA are sensitive to noise and fail to generalize.

In contrast, Fig. 17(c) and (f) demonstrate that increasing RWDA levels in training improves test accuracy, confirming enhanced model robustness under real-world conditions. However, this improvement is nonlinear—initial increments in RWDA yield noticeable gains, while higher levels provide diminishing returns, indicating a saturation effect in robustness enhancement.

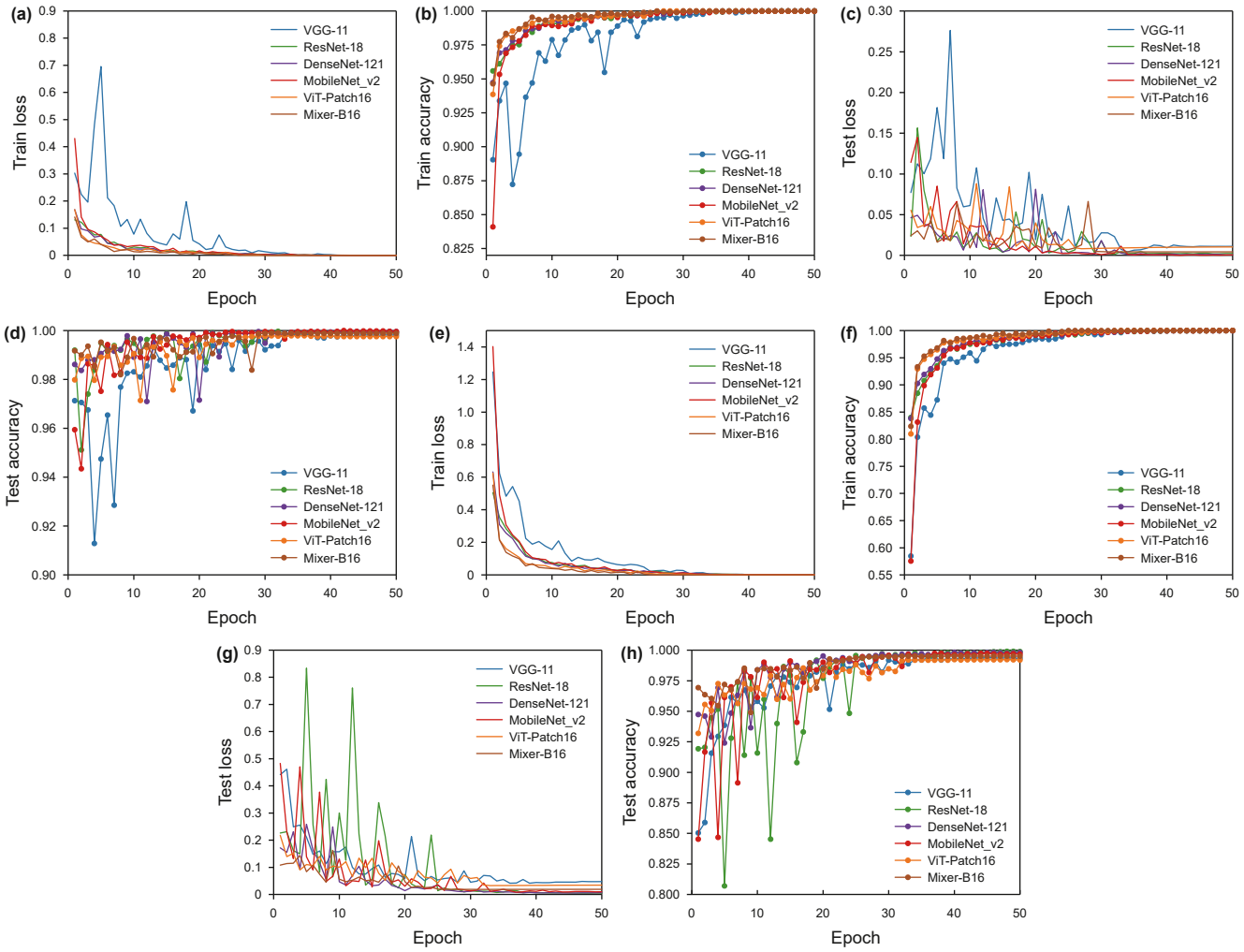


Fig. 11. Performance of different model architectures on DCID-7 and DCID-35. **(a)** Train loss on DCID-7; **(b)** train accuracy on DCID-7; **(c)** test loss on DCID-7; **(d)** test accuracy on DCID-7; **(e)** train loss on DCID-35; **(f)** train accuracy on DCID-35; **(g)** test loss on DCID-35; **(h)** test accuracy on DCID-35.

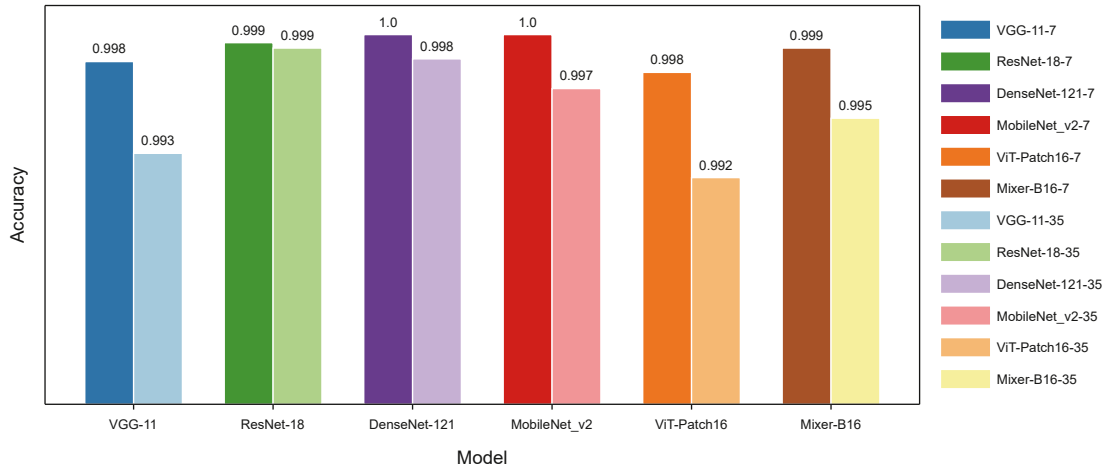


Fig. 12. Accuracy of different model architectures on DCID-7 and DCID-35.

4.6. Impact of image resolution in simulated real-world environments

This section examines how image resolution affects model performance under simulated real-world conditions. Using ResNet-18, we applied 40% RWDA to both training and testing datasets (DCID-R-C-0.4-A) and trained the model on images with four resolutions: 256 × 256, 128 × 128, 64 × 64, and 32 × 32.

As shown in Fig. 18(a)–(d) and Fig. 18(e)–(h), training loss decreases rapidly across all resolutions, indicating that the model

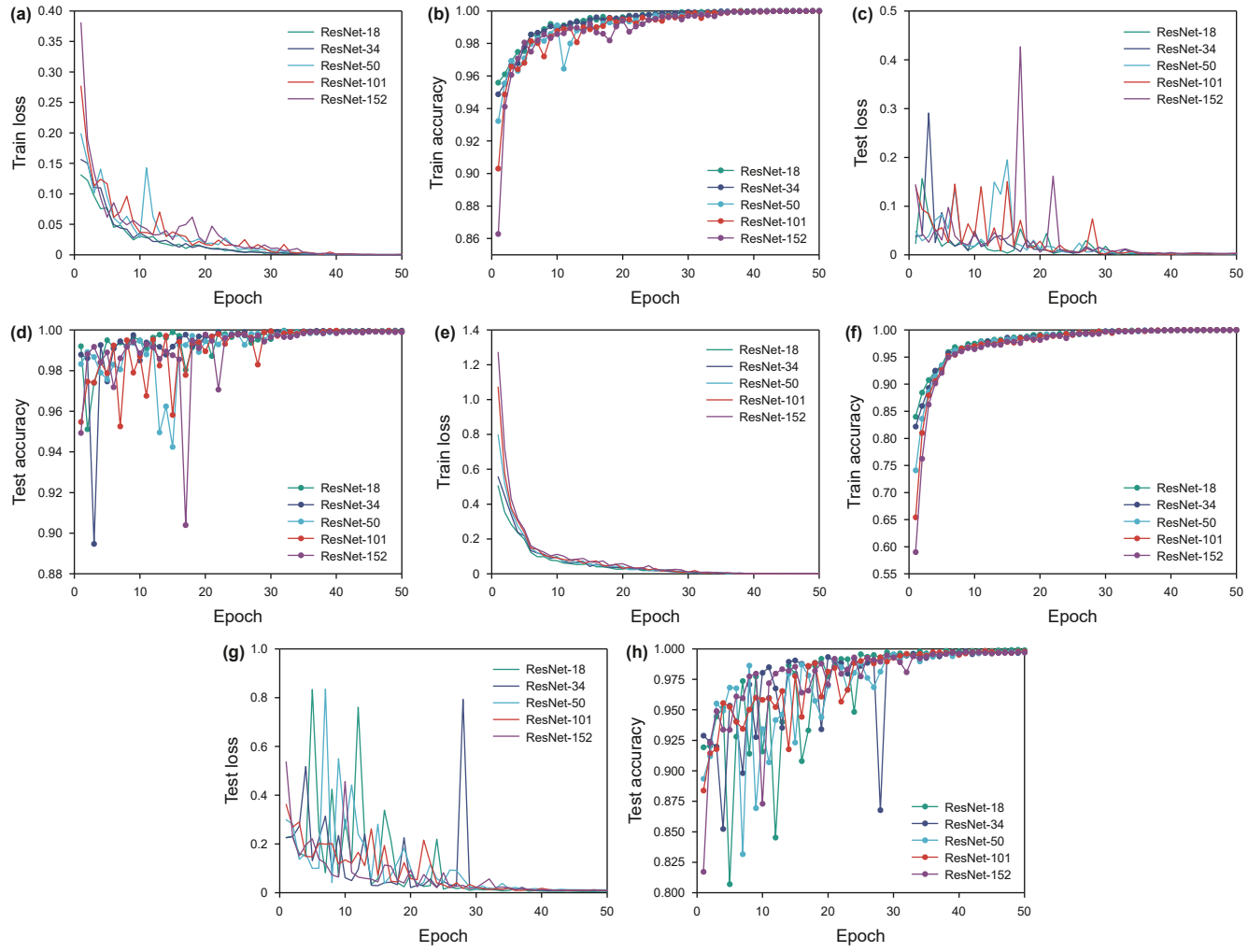


Fig. 13. Performance of models with different sizes on DCID-7 and DCID-35. **(a)** Train loss on DCID-7; **(b)** train accuracy on DCID-7; **(c)** test loss on DCID-7; **(d)** test accuracy on DCID-7; **(e)** train loss on DCID-35; **(f)** train accuracy on DCID-35; **(g)** test loss on DCID-35; **(h)** test accuracy on DCID-35.

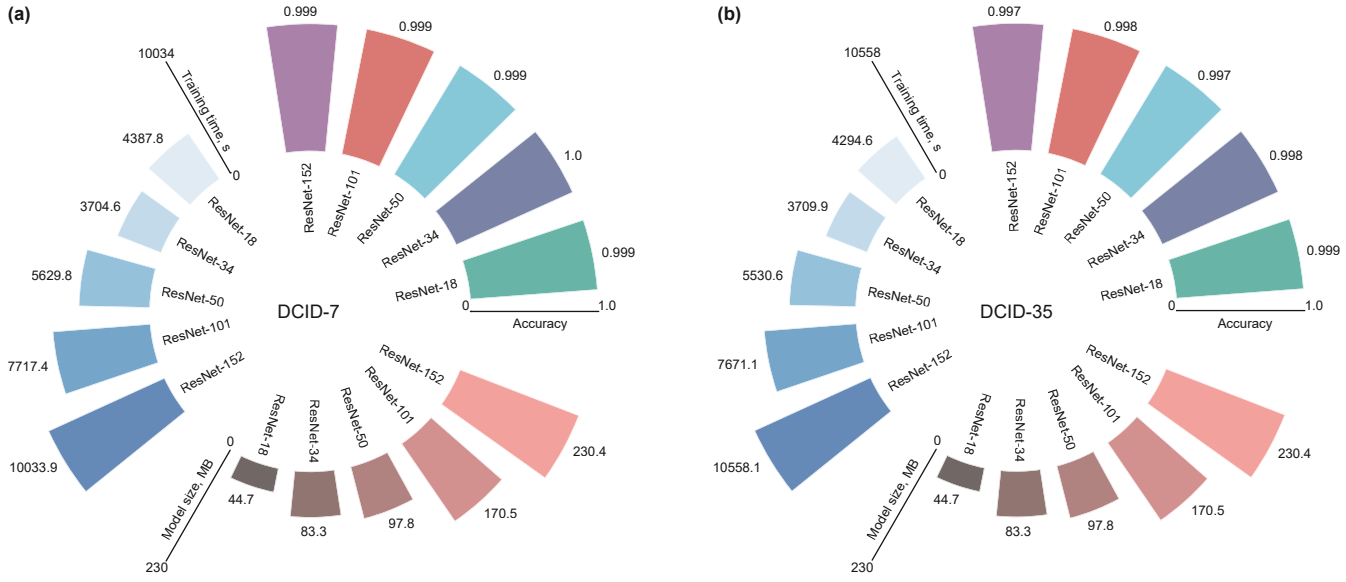


Fig. 14. Comparison of model accuracy, training time, and size across different architectures. **(a)** DCID-7; **(b)** DCID-35.

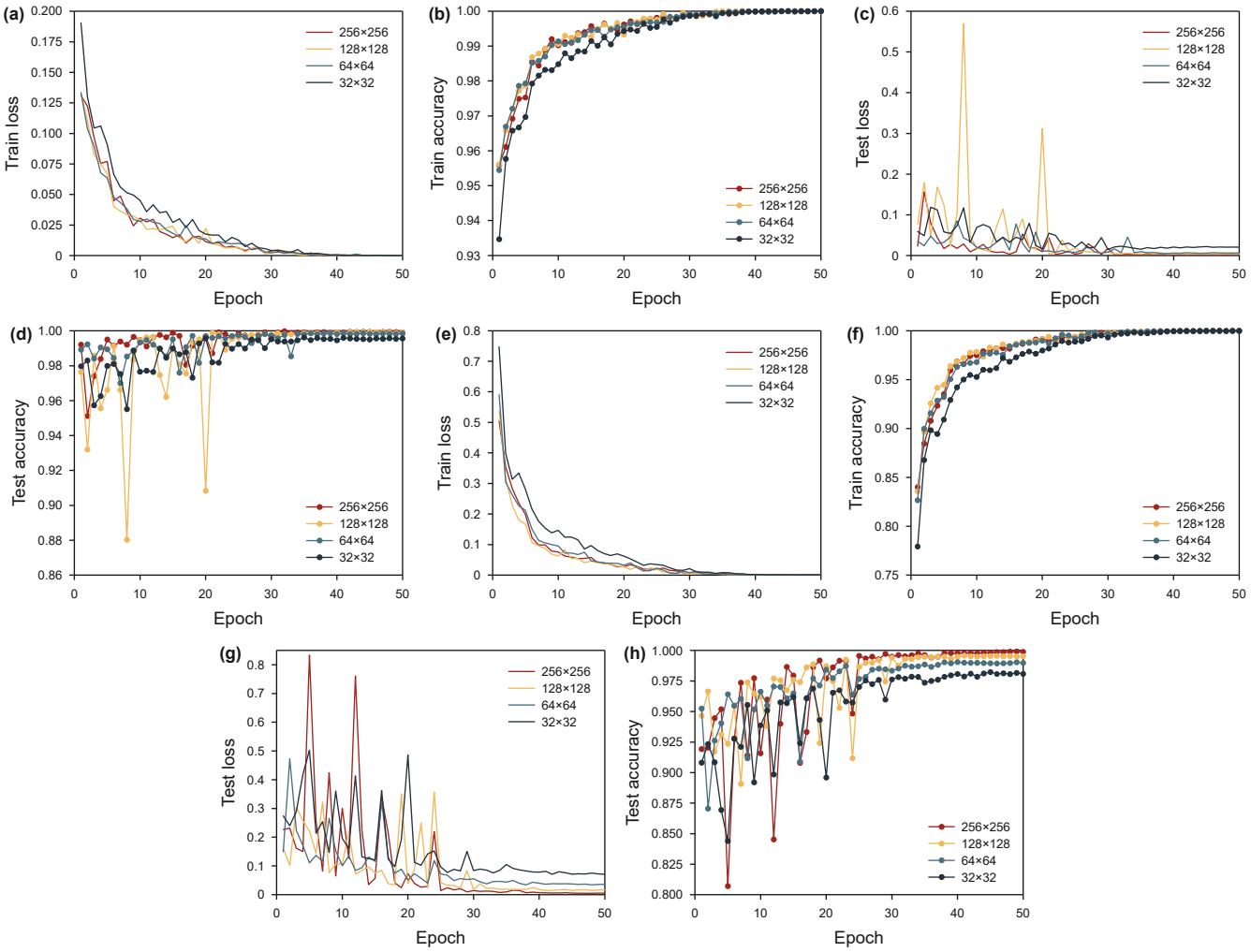


Fig. 15. Model performance across different image resolutions. **(a)** Train loss on DCID-7; **(b)** train accuracy on DCID-7; **(c)** test loss on DCID-7; **(d)** test accuracy on DCID-7; **(e)** train loss on DCID-35; **(f)** train accuracy on DCID-35; **(g)** test loss on DCID-35; **(h)** test accuracy on DCID-35.

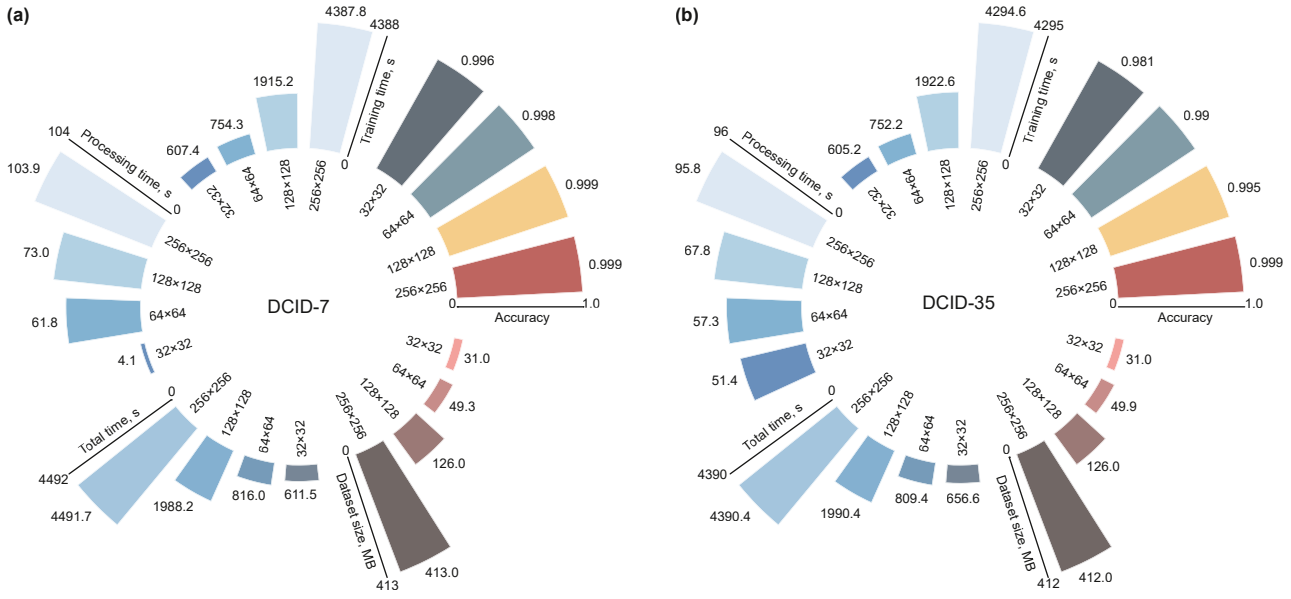


Fig. 16. Accuracy, training time, inference time, total runtime, and dataset size across different image resolutions. **(a)** DCID-7; **(b)** DCID-35.

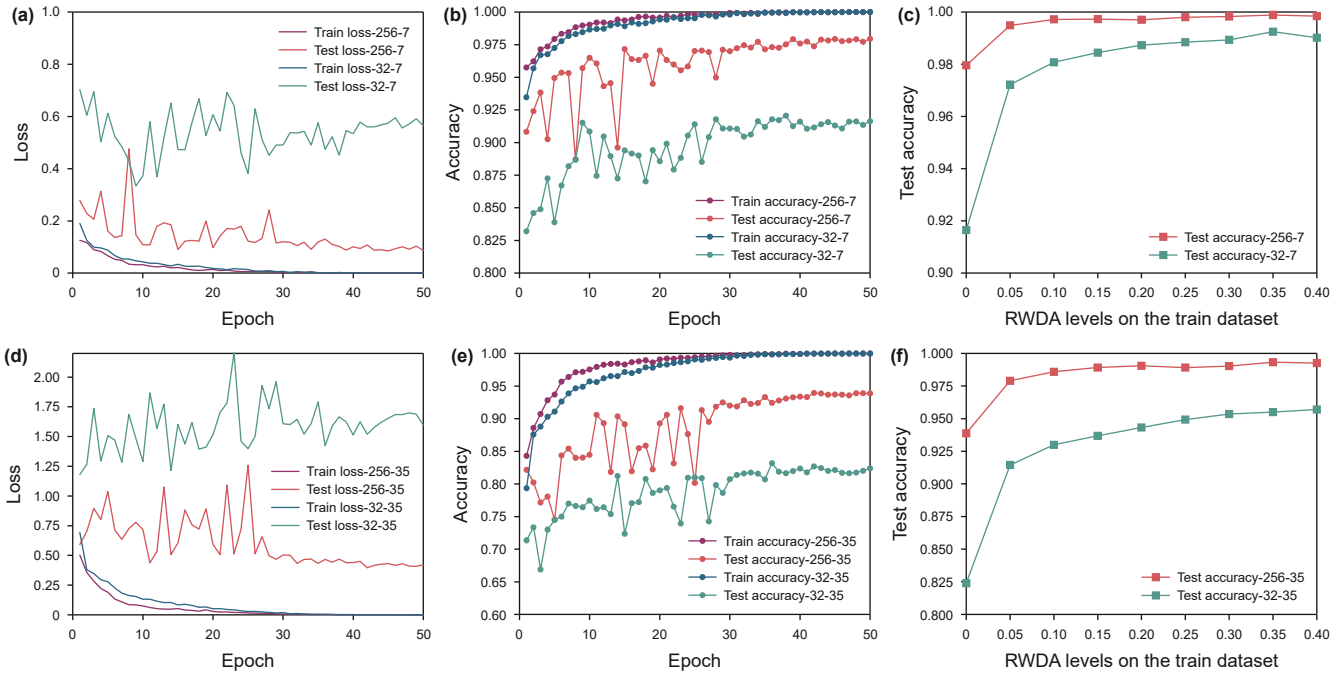


Fig. 17. Impact of different RWDA levels on model performance. **(a)** Training and testing loss at 0% RWDA on DCID-256/32−7; **(b)** training and testing accuracy at 0% RWDA on DCID-256/32−7; **(c)** test accuracy under varying RWDA levels on DCID-256/32−7; **(d)** training and testing loss at 0% RWDA on DCID-256/32−35; **(e)** training and testing accuracy at 0% RWDA on DCID-256/32−35; **(f) t**est accuracy under varying RWDA levels on DCID-256/32−35.

remains capable of learning despite the presence of defective samples. Loss values are consistently higher for the 35-class dataset, reflecting the added complexity of multi-class learning under noisy conditions. Training accuracy improves quickly, suggesting effective adaptation, while high-resolution models achieve slightly higher accuracy more rapidly. Testing loss and accuracy exhibit larger fluctuations with high-resolution inputs, especially in the 35-class case, suggesting a greater risk of overfitting to defective features. In contrast, low-resolution models show more stable testing curves, likely due to their inherent smoothing effect, which filters out fine-grained noise along with some detailed features.

As illustrated in Fig. 19, accuracy decreases with lower resolution, but the severity depends on task complexity. In the 7-class dataset, the decline is modest, indicating that key features are still retained at lower resolutions. However, in the 35-class dataset, accuracy drops more sharply—especially from 64 × 64 to 32 × 32—highlighting the need for higher resolution to distinguish fine-grained classes. Resolution plays a crucial role in complex classification tasks. While high-resolution images enhance accuracy under challenging conditions, the gains diminish beyond a certain point. For simpler tasks, 128 × 128 may offer a good balance between efficiency and performance, whereas for more complex tasks, higher resolutions remain essential.

4.7. Impact of lighting variations on model performance

To assess the robustness of lithology identification models under varying real-world lighting conditions, we conducted experiments using the DCID-32-7 and DCID-32-35 datasets (Li et al., 2021). As shown in Fig. 20, the datasets were augmented to simulate lighting variations through adjustments in brightness, contrast, saturation, hue, and their combination. Brightness and contrast were varied by ±50% to simulate intensity and shadow variations. Saturation was adjusted by ±50% to mimic color richness; hue by ±10% to simulate shifts in light source temperature. The Combination setting introduced all adjustments simultaneously, replicating complex lighting conditions. We trained ResNet-18 models on each adjusted dataset, using identical training parameters except for the applied augmentations, and compared results against models trained on the original dataset (Origin).

As shown in Fig. 21, models trained on the Origin dataset consistently achieved the lowest training loss and highest test accuracy for both datasets. Models trained with single-factor variations (brightness, contrast, saturation) showed only moderate performance drops. In contrast, the Combination setting led to the highest loss and steepest accuracy decline, indicating significant challenges under complex lighting conditions.

The performance degradation was more pronounced on DCID-32-35, likely due to its higher class count and fewer samples per class. Fig. 22 summarizes final test accuracies: the origin-trained model reached 0.996 (DCID-32-7) and 0.981 (DCID-32-35), while the combination condition resulted in the lowest scores (0.810 and 0.647, respectively). Among single-factor adjustments, contrast had the least impact, suggesting greater robustness to contrast changes compared to brightness or saturation.

4.8. Rapid model evaluation using small-sized datasets

This section investigates the effectiveness of using small-sized datasets for training and evaluating lithology identification models. These datasets reduce training time, processing speed, and memory usage, making them suitable for rapid model development and comparative analysis.

We used four datasets: DCID-32-7-0-N and DCID-32-35-0-N (without RWDA) to evaluate performance under ideal conditions, and DCID-32-7-0.4-A and DCID-32-35-0.4-A (with 40% RWDA) to simulate real-world noise. Model performance was assessed using accuracy on clean data (ACC-7, ACC-35), robustness under noise (NDA-7, NDA-35), along with training time (TT), throughput (TP), inference speed (IS), and model size (MS).

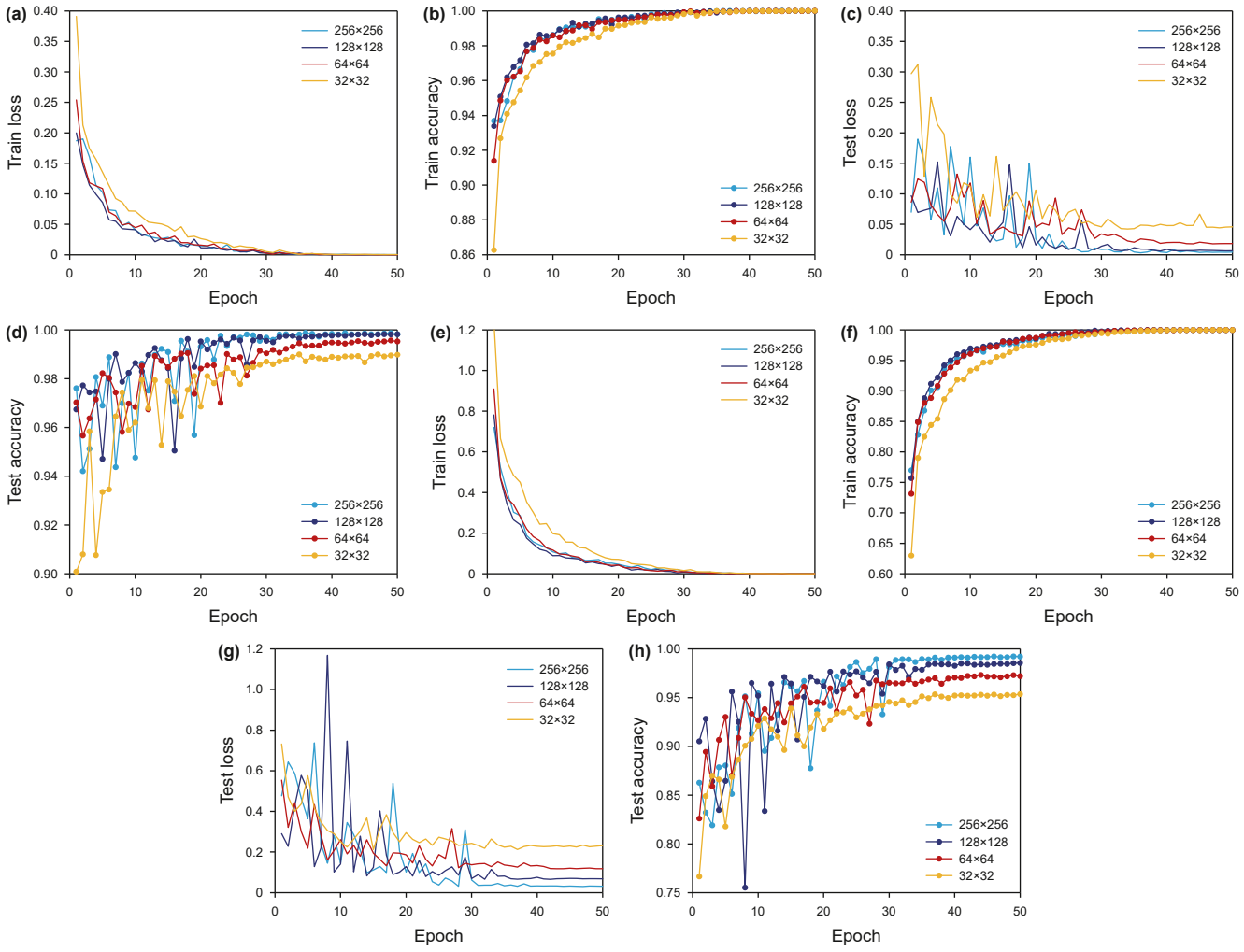


Fig. 18. Model performance on RWDA datasets with different image resolutions. **(a)** Train loss on DCID-7; **(b)** train accuracy on DCID-7; **(c)** test loss on DCID-7; **(d)** test accuracy on DCID-7; **(e)** train loss on DCID-35; **(f)** train accuracy on DCID-35; **(g)** test loss on DCID-35; **(h)** test accuracy on DCID-35.

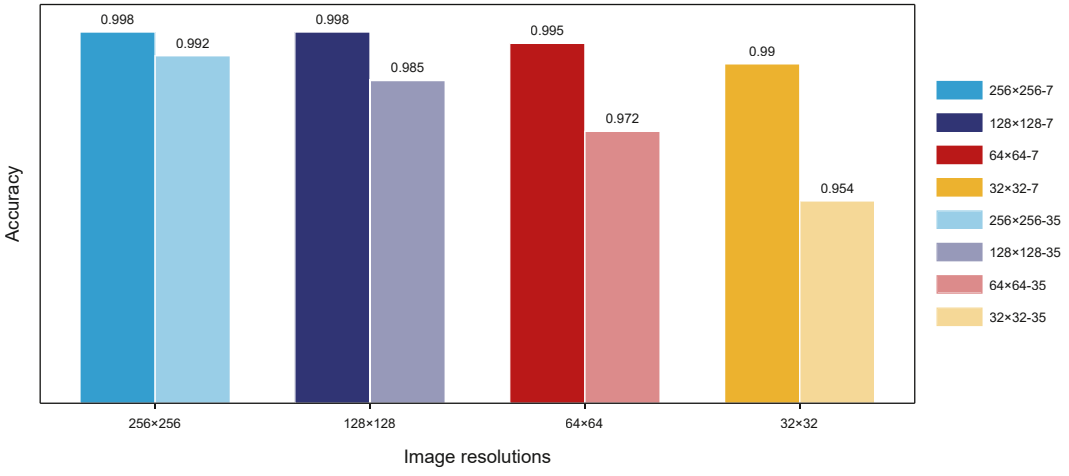


Fig. 19. Accuracy across datasets with RWDA at different resolutions.

As shown in Fig. 23 and Table 3, ResNet-18 performed best overall, combining high accuracy and robustness (ACC7 = 0.997, NDA35 = 0.951) with the fastest training and inference speeds. DenseNet-121 achieved the highest accuracy on the 35-class dataset (ACC35 = 0.986), but required more time and resources. VGG-11 showed strong clean-data performance but struggled under noise and had the longest training time. MobileNet_v2 was the most lightweight, offering fast inference and small size, though

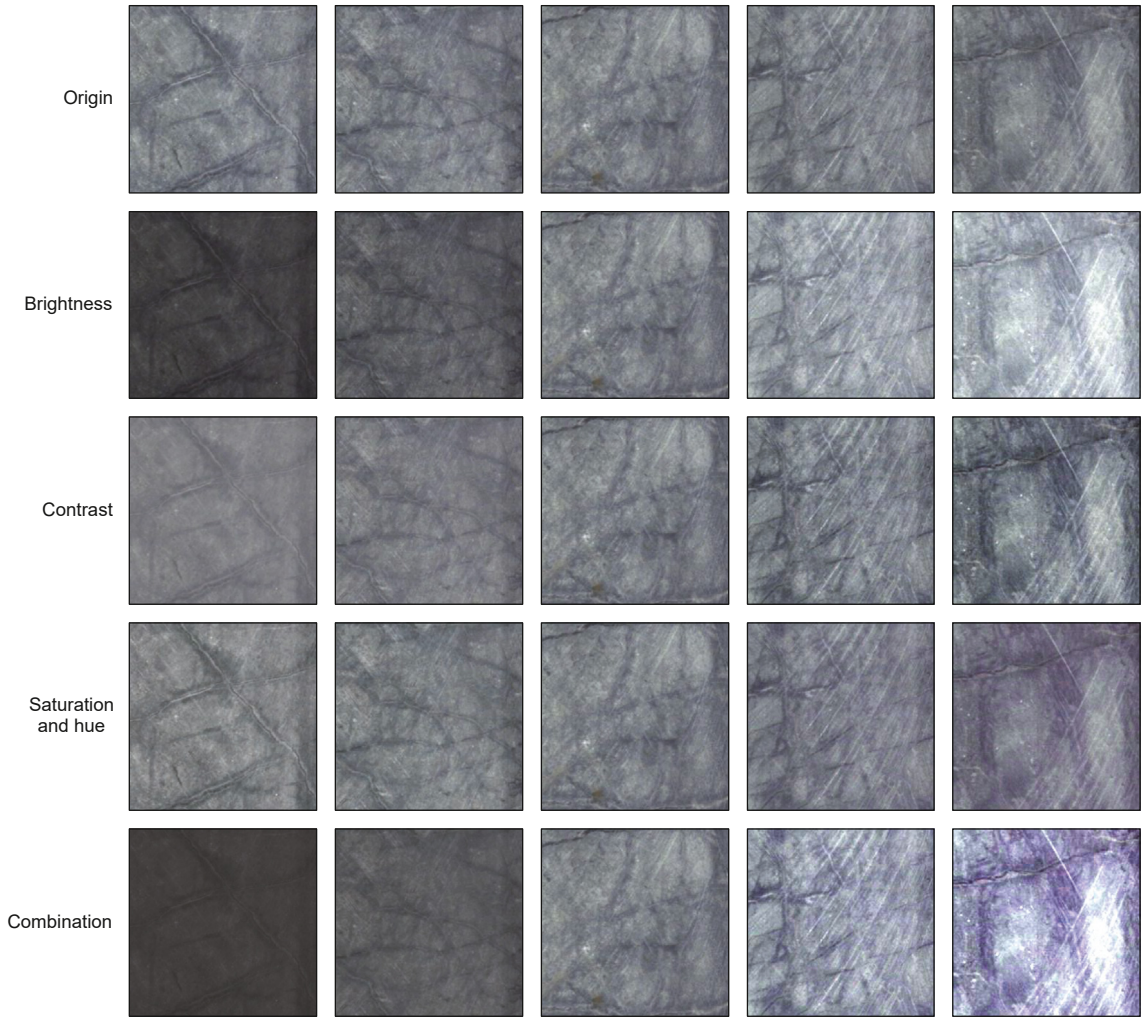


Fig. 20. Examples of lighting environment variations applied to the dataset (Marble). Brightness: 50%–150%. Contrast: 50%–150%. Saturation: 50%–150%. Hue: −10%–10%. Combination: Simultaneous adjustments of brightness, contrast, saturation, and hue.

with slightly reduced robustness. ViT-Patch16 and Mixer-B16 achieved good accuracy but were less efficient and more sensitive to noise. These results confirm that small-sized datasets can effectively benchmark model performance, revealing trade-offs between accuracy, robustness, and efficiency across architectures.

5. Discussion

5.1. Transfer learning

This study demonstrates how lithology identification tasks can be effectively conducted using the constructed datasets, including experiments across different model architectures, parameter sizes, image resolutions, and lighting conditions. While the results offer valuable insights into model performance and influencing factors, further research is needed to enhance generalizability and robustness.

A key limitation is the lack of validation on external datasets, which is essential for reliable performance assessment. Due to resource constraints, such validation was not feasible in this work. To address this, we plan to release all constructed datasets, model training procedures, and trained parameters to the public, enabling future benchmarking and reproducibility.

As shown in Fig. 24, the trained model parameters can be directly reused for future testing on newly developed datasets or for deployment in real-world lithology identification scenarios. They may also serve as a foundation for transfer learning, facilitating model adaptation to new domains and further assessing generalization capabilities.

5.2. Multimodality

Integrating geospatial data—such as borehole logs and geological models—is vital for improving the accuracy and practical relevance of lithology identification, especially in exploration settings. Spatial context complements image-based features, enhancing model interpretability and performance (Liu et al., 2024; Saidi et al., 2024). Our current work focuses on image-based datasets. Due to data limitations, we have not yet incorporated logging or spatial data. However, developing multimodal datasets that combine core images, logs, and spatial information is a key direction for future research.

Advances in multimodal deep learning, including early, late, and hybrid fusion strategies, offer promising ways to integrate diverse data sources (Jabeen et al., 2023; Kieu et al., 2024). These techniques can capture complementary features, improving model accuracy, robustness, and generalization. Future efforts will focus on

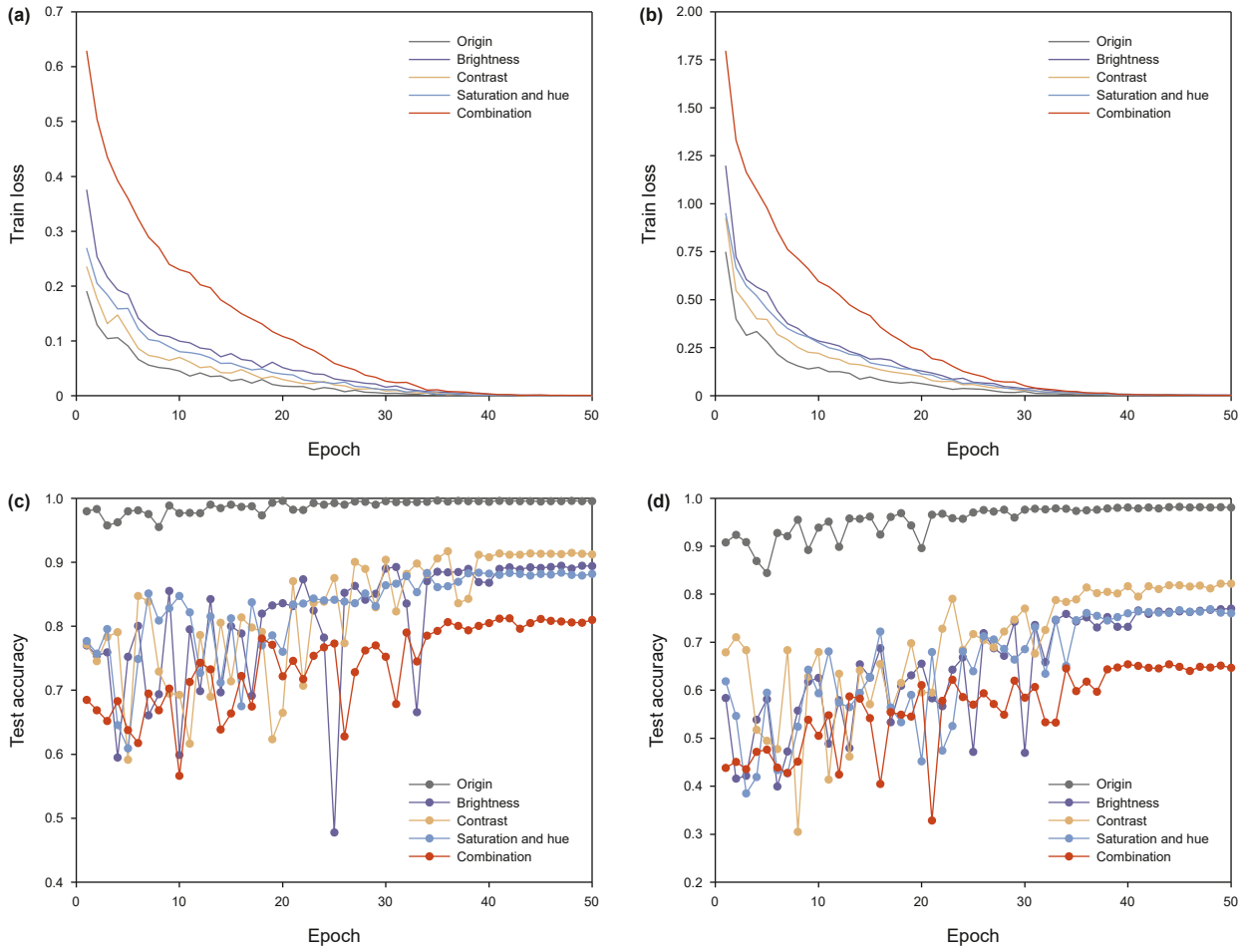


Fig. 21. Training and testing performance under lighting variations. **(a)** Train loss on DCID-7; **(b)** train loss on DCID-35; **(c)** test accuracy on DCID-7; **(d)** test accuracy on DCID-35.

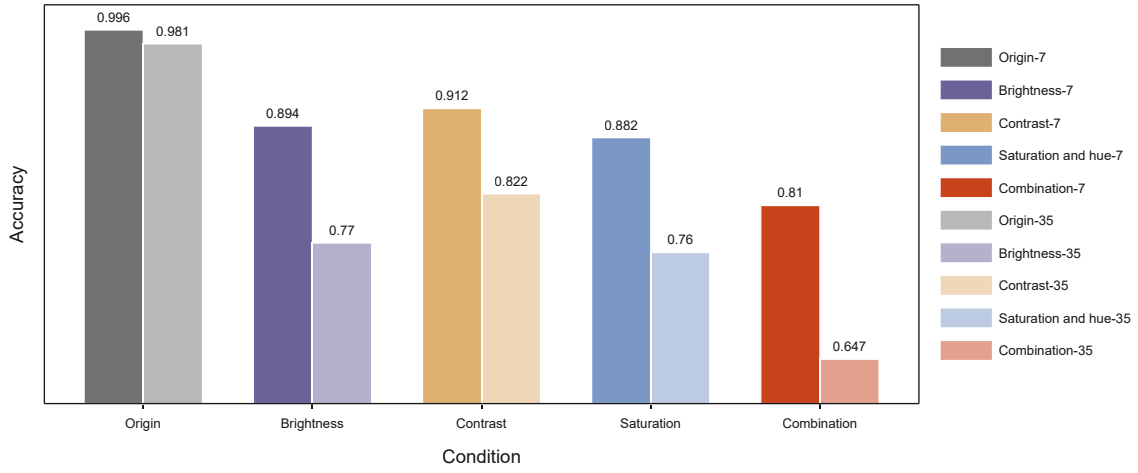


Fig. 22. Test accuracy under different lighting conditions.

constructing integrated datasets and applying fusion models to address real-world geological complexity more effectively.

6. Conclusions

This study presents the development and open release of the Drill Core Image Dataset (DCID)—the first publicly available core image dataset for lithology identification. Using DCID, we benchmarked a range of models, including CNNs (VGG, ResNet, DenseNet, MobileNet) and Transformer-based architectures (ViT, MLP-Mixer), and evaluated their performance under varying model sizes, image resolutions, lighting conditions, and simulated real-world noise. To address the lack of high-quality lithology datasets, we introduced a real-world data augmentation (RWDA) strategy based on slightly

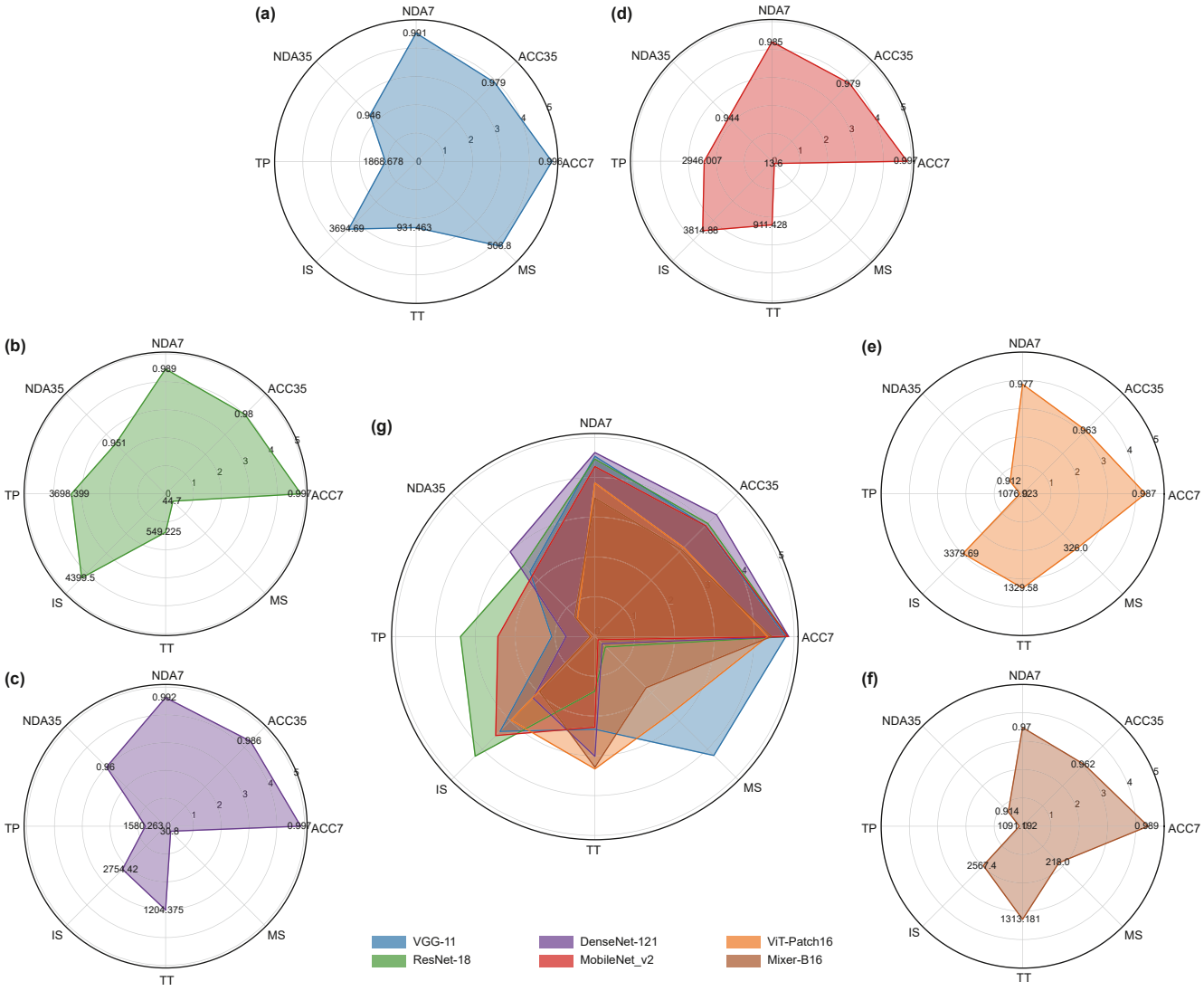


Fig. 23. Radar charts of different models. **(a)** VGG-11; **(b)** ResNet-18; **(c)** DenseNet-121; **(d)** MobileNet_v2; **(e)** ViT-Patch16; **(f)** Mixer-B16; **(g)** comparison.

Table 3
Performance of different models on low-resolution datasets.

Model	ACC7	ACC35	NDA7	NDA35	TP, img/s	IS, img/s	TT, s	MS, MB
VGG-11	0.996	0.979	0.991	0.946	1868.678	3694.69	931.463	506.8
ResNet-18	0.997	0.980	0.989	0.951	3698.399	4399.50	549.225	44.7
DenseNet-121	0.997	0.986	0.992	0.960	1580.263	2754.42	1204.375	30.8
MobileNet_v2	0.997	0.979	0.985	0.944	2946.007	3814.88	911.428	13.6
ViT-Patch16	0.987	0.963	0.977	0.912	1076.923	3379.69	1329.579	326.0
Mixer-B16	0.989	0.962	0.970	0.914	1091.192	2567.40	1313.181	218.0

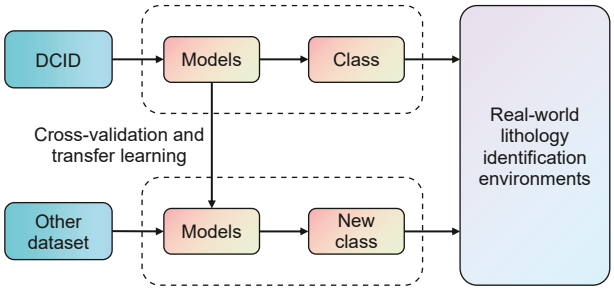


Fig. 24. Cross-validation, transfer learning and application with DCID.

defective images. Experiments demonstrated that RWDA improves model robustness. Additionally, we showed that low-resolution datasets can be effectively used for rapid model evaluation and iteration. All trained models and parameters are publicly available, supporting future research in transfer learning and real-world deployment. Future work should further validate RWDA in practical settings, expand DCID to include more lithology classes, and integrate additional data types—such as borehole logs and geospatial models—to enable multimodal lithology identification with enhanced accuracy and generalization.

CRediT authorship contribution statement

Jia-Yu Li: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Ji-Zhou Tang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Xian-Zheng Zhao:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Bo Fan:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation. **Wen-Ya Jiang:** Conceptualization, Methodology, Validation, Investigation, Writing – review & editing. **Shun-Yao Song:** Supervision, Project administration, Investigation, Funding acquisition. **Jian-Bing Li:** Supervision, Software, Formal analysis, Data curation. **Kai-Da Chen:** Writing – original draft, Visualization, Software. **Zheng-Guang Zhao:** Writing – review & editing, Software.

Declaration of competing interest

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

Acknowledgements

The authors would like to express their gratitude to the Geological Survey of South Australia for hosting the Drill Core Reference Library. Their open-source online data repositories have been invaluable for the research conducted in this study.

The authors gratefully acknowledge the support from the National Natural Science Foundation of China (Nos. U24B2034, U2139204), the China Petroleum Science and Technology Innovation Fund (2021DQ02-0501), and the Science and Technology Support Project of Langfang (2024011073).

References

Alzubaidi, F., Mostaghimi, P., Swietojanski, P., et al., 2021. Automated lithology classification from drill core images using convolutional neural networks. J. Petrol. Sci. Eng. 197, 107933. https://doi.org/10.1016/j.petrol.2020.107933.

Bai, X.F., Zhang, F.Y., Zou, H.Y., et al., 2025. Enhanced domain tuned Yolo-driven intelligent fault identification method: application in selection and construction of gas storage. Well Logging Technol. 49 (1), 47–56. https://doi.org/10.16489/j.issn.1004-1338.2025.01.006.

Baraboshkin, E.E., Ismailova, L.S., Orlov, D.M., et al., 2020. Deep convolutions for in-depth automated rock typing. Comput. Geosci. 135, 104330. https://doi.org/10.1016/j.cageo.2019.104330.

Borsaru, M., Zhou, B., Aizawa, T., et al., 2006. Automated lithology prediction from PGNAA and other geophysical logs. Appl. Radiat. Isot. 64, 272–282. https://doi.org/10.1016/j.apradiso.2005.07.012.

Busch, J., Fortney, W., Berry, L., 1987. Determination of lithology from well logs by statistical analysis. SPE Form. Eval. 2, 412–418. https://doi.org/10.2118/14301-PA.

Chen, G., Li, J., 2022. Cubenet: array-based seismic phase picking with deep learning. Seismol. Soc. Am. 93, 2554–2569. https://doi.org/10.1785/0220220147.

Dong, S.Q., Zhong, Z.H., Cui, X.H., et al., 2023. A deep kernel method for lithofacies identification using conventional well logs. Pet. Sci. 20, 1411–1428. https://doi.org/10.1016/j.petsci.2022.11.027.

Dosovitskiy, A., 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv Preprint, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929.

Dubois, M.K., Bohling, G.C., Chakrabarti, S., 2007. Comparison of four approaches to a rock facies classification problem. Comput. Geosci. 33, 599–617. https://doi.org/10.1016/j.cageo.2006.08.011.

Fu, D., Su, C., Wang, W., et al., 2022. Deep learning based lithology classification of drill core images. PLoS One 17, e0270826. https://doi.org/10.1371/journal.pone.0270826.

Fu, G.M., Yan, J.Y., Zhang, K., et al., 2017. Current status and progress of lithology identification Technology. Prog. Geophys. 32, 26–40. https://doi.org/10.6038/pg20170104 (in Chinese).

Galdames, F.J., Perez, C.A., Estévez, P.A., et al., 2017. Classification of rock lithology by laser range 3D and color images. Int. J. Miner. Process. 160, 47–57. https://doi.org/10.1016/j.minpro.2017.01.008.

Geological Survey of South Australia, n.d. Hylogger Data. Government of South Australia, Adelaide. http://www.energymining.sa.gov.au/minerals/geoscience/geoscientific_data/hylogger.

Government of South Australia — Department for Energy and Mining. https://map.sarig.sa.gov.au/.

He, J., La Croix, A.D., Wang, J., et al., 2019. Using neural networks and the Markov chain approach for facies analysis and prediction from well logs in the precipice sandstone and evergreen formation, surat basin, Australia. Mar. Petrol. Geol. 101, 410–427. https://doi.org/10.1016/j.marpetgeo.2018.12.022.

He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Howard, A.G., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv Preprint, arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., et al., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243.

Huang, L., Qi, Y., Chen, W., et al., 2023. A geomechanical modeling method for shale oil reservoir cluster well area based on GridSearchCV. Well Logging Technol. 47, 421–431. https://doi.org/10.16489/j.issn.1004-1338.2023.04.005.

Imamverdiyev, Y., Sukhostat, L., 2019. Lithological facies classification using deep convolutional neural network. J. Petrol. Sci. Eng. 174, 216–228. https://doi.org/10.1016/j.petrol.2018.11.023.

Izadi, H., Sadri, J., Bayati, M., 2017. An intelligent system for mineral identification in thin sections based on a cascade approach. Comput. Geosci. 99, 37–49. https://doi.org/10.1016/j.cageo.2016.10.010.

Jabeen, S., Li, X., Amin, M.S., et al., 2023. A review on methods and applications in multimodal deep learning. ACM Trans. Multimed Comput. Commun. Appl 19, 1–41. https://doi.org/10.1145/3545572.

Jacobs, R.A., 1988. Increased rates of convergence through learning rate adaptation. Neural Netw. 1, 295–307. https://doi.org/10.1016/0893-6080(88)90003-2.

Kieu, N., Nguyen, K., Nazib, A., et al., 2024. Multimodal Co-learning meets remote sensing: taxonomy, state of the art, and future works. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. https://doi.org/10.1109/JSTARS.2024.3378348.

Kingma, D.P., 2014. Adam: a method for stochastic optimization. arXiv Preprint, arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, C., Guo, C., Han, L., et al., 2021. Low-light image and video enhancement using deep learning: a Survey. IEEE Trans. Pattern Anal. Mach. Intell. 44, 9396–9416. https://doi.org/10.1109/TPAMI.2021.3126387.

Liu, H., Xia, S., Fan, C., et al., 2024. 3D geo-modeling framework for multisource heterogeneous data fusion based on multimodal deep learning and multipoint statistics: a case study in SouthSouth China sea. EGUsphere 1–44. https://doi.org/10.5194/egusphere-2024-44, 2024.

Liu, X., Meng, S.W., Liang, Z.Z., et al., 2023. Microscale crack propagation in shale samples using focused ion beam scanning electron microscopy and three-dimensional numerical modeling. Pet. Sci. 20, 1488–1512. https://doi.org/10.1016/j.petsci.2022.10.004.

Marmo, R., Amodio, S., Tagliaferri, R., et al., 2005. Textural identification of carbonate rocks by image processing and neural network: methodology proposal and examples. Comput. Geosci. 31, 649–659. https://doi.org/10.1016/j.cageo.2004.11.016.

Raschka, S., 2018. Model evaluation, model selection, and algorithm selection in machine learning. arXiv Preprint, arXiv:1811.12808. https://doi.org/10.48550/arXiv.1811.12808.

Ren, Q., Zhang, D., Zhao, X., et al., 2022. A novel hybrid method of lithology identification based on K-Means++ algorithm and fuzzy decision tree. J. Petrol. Sci. Eng. 208, 109681. https://doi.org/10.1016/j.petrol.2021.109681.

Saidi, S., Idbraim, S., Karmoude, Y., et al., 2024. Deep-learning for change detection using multi-modal fusion of remote sensing images: A review. Remote Sens. 16, 3852. https://doi.org/10.3390/rs16203852.

Shi, H., Xu, Z., Lin, P., et al., 2023. Refined lithology identification: methodology, challenges and prospects. Geoener. Sci. Eng. 231, 212382. https://doi.org/10.1016/j.geoen.2023.212382.

Shorten, C., Khoshgoftaar, T.M., 2019. A Survey on image data augmentation for deep learning. J. Big Data 6, 1–48. https://doi.org/10.1186/s40537-019-0197-0.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv Preprint, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556.

Tang, J.Z., Zhang, Z., Xie, J., et al., 2024. Re-evaluation of CO_2 storage capacity of depleted fractured-vuggy carbonate reservoir. Innovat. Energy 1 (2), 100019. https://doi.org/10.59717/j.xinn-energy.2024.100019.

Thomas, A., Rider, M., Curtis, A., et al., 2011. Automated lithology extraction from core photographs. First Break 29.

Tian, Y.J., Pan, H.X., Liu, X.C., et al., 2013. Lithofacies recognition based on extreme learning machine. Appl. Mech. Mater. 241, 1762–1767. https://doi.org/10.4028/www.scientific.net/AMM.241-244.1762.

Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., et al., 2021. MLP-mixer: an all-MLP architecture for vision. Adv. Neural Inf. Process. Syst. 34, 24261–24272. https://doi.org/10.48550/arXiv.2105.01601.

Wieling, I.S., 2013. Facies and permeability prediction based on analysis of core images. Master Thesis. Delft University of Technology. https://resolver.tudelft.nl/uuid:9b6bd4b0-1001-4d9b-a6eb-7761bc3b2309.

Xie, Y., Zhu, C., Zhou, W., et al., 2018. Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. J. Petrol. Sci. Eng. 160, 182–193. https://doi.org/10.1016/j.petrol.2017.10.028.

Xu, Z., Liu, F., Lin, P., et al., 2021. Non-destructive, in-situ, fast identification of adverse geology in tunnels based on anomalies analysis of element content. Tunn. Undergr. Space Technol. 118, 104146. https://doi.org/10.1016/j.tust.2021.104146.

Zhang, P., Sun, J., Jiang, Y., et al., 2017. Deep learning method for lithology identification from borehole images. In: 79th EAGE Conference and Exhibition 2017. European Association of Geoscientists & Engineers, pp. 1–5. https://doi.org/10.3997/2214-4609.201701164.

Zhang, Z., Tang, J., Fan, B., et al., 2024. An intelligent lithology recognition system for continental shale by using digital coring images and convolutional neural networks. Geoener. Sci. Eng. 239, 212909. https://doi.org/10.1016/j.geoen.2024.212909.

Zhao, F., Yang, Y., Kang, J., et al., 2023. CE-SGAN: classification enhancement semi-supervised generative adversarial network for lithology identification. Geoener. Sci. Eng. 223, 211562. https://doi.org/10.1016/j.geoen.2023.211562.

Zhao, X., Jin, F., Liu, X., et al., 2022. Numerical study of fracture dynamics in different shale fabric facies by integrating machine learning and 3-D lattice method: a case from cangdong sag, bohai bay basin, China. J. Petrol. Sci. Eng. 218, 110861. https://doi.org/10.1016/j.petrol.2022.110861.