Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Proof of Concept for Dunham Classification of Carbonate Thin Sections Using Vision Transformers

Mohammad Aviandito

raden.aviandito@pertamina.com

Pertamina (Indonesia)

Research Article

Keywords:

Posted Date: April 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9244708/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Proof of Concept for Dunham Classification of Carbonate Thin Sections Using Vision Transformers

Mohammad Aviandito

Pertamina

Corresponding author: Mohammad Aviandito (raden.aviandito@pertamina.com)

Abstract

The classification of carbonate rocks using petrographic thin sections is a fundamental task in sedimentology and is traditionally performed by geologists using schemes such as Dunham's depositional-texture classification. However, this approach is time-consuming and may be prone to inter-observer variability, as reported in prior studies.

This paper presents a proof of concept for using a Vision Transformer (ViT) to classify carbonate thin-section images into Dunham classes. The dataset from Lokier and Al Junaibi (2016) is used for fine-tuning and consists of 14 carbonate samples imaged at multiple magnifications and polarization states and labeled through expert consensus. A pre-trained ViT model (vit-base-patch16-224-in21k) is fine-tuned using open-source tools, with an emphasis on reproducibility through transparent methodology and publicly available code.

The fine-tuned model achieves high accuracy on a held-out test split derived from the same dataset; however, this performance likely reflects overfitting due to the limited size of the dataset. Evaluation on an independent validation dataset sourced from the CarbonateWorld website yields a lower accuracy of 67%, consistent with known ambiguities in carbonate texture classification. Notably, when considering the top-2 predicted classes, validation accuracy improves to 85.71%, indicating that secondary prediction probabilities capture these ambiguities.

Overall, the results highlight both the potential of ViT-based models for carbonate classification and the importance of open, expert-labeled datasets. Rather than proposing a state-of-the-art classifier, this study demonstrates a reproducible and extensible workflow. By providing a clear methodology and code snippets, it enables readers to fine-tune ViT models for their own use cases.

1. Introduction

Classifying carbonate rock textures is critical for interpreting depositional environments, diagenesis, and reservoir quality. The Dunham classification scheme (Dunham, 1962) is widely used in sedimentary geology for carbonate rocks (Lokier & Al-Junaibi, 2016), defining classes such as mudstone, wackestone, packstone, grainstone, boundstone, and crystalline carbonate based on fabric and mud content. Application of this scheme

traditionally relies on expert petrographic analysis of thin-section images. However, human classification can be time-consuming and may suffer from inter-observer variability.

Recent advances in deep learning offer opportunities for automating image-based classification tasks. Most notably, Vision Transformers (ViT) (Wu, et al., 2020; Dosovitskiy, et al., 2021) have achieved state-of-the-art results in many computer-vision tasks by leveraging self-attention mechanisms. In this study, a ViT model is fine-tuned to classify carbonate thin sections into Dunham classes. Thin-section photographs from Lokier and Al-Junaibi (2016) are used as the training dataset, which contains 14 carbonate samples labeled by a group of expert geoscientists from both industry and academia. The ViT model `vit-base-patch16-224-in21k`, pre-trained on `ImageNet-21k`, is fine-tuned on this dataset. By leveraging an open-access dataset and the open-source Hugging Face Transformers library (Wolf, et al., 2020), the reproducibility of the pipeline is ensured. Training details and outcomes are reported, including code snippets for each step.

Due to the limitations of the training data, this study does not aim to develop a state-of-the-art model for carbonate classification. But instead, by including step-by-step code snippets, replication of this study, and experimentations/adaptation to other use cases within geoscience are encouraged.

2. Literature Review

Building on this context, it is necessary to review bodies of prior work that underpin this research. First, the conceptual and practical foundations of the Dunham carbonate classification scheme are examined. Second, the Lokier and Al-Junaibi (2016) study is reviewed in detail, as it provides both the empirical basis for this study and an important perspective on expert variability in carbonate classification. Third, previous studies that employ machine learning for carbonate identification, most of which use pre-Transformer architectures, are reviewed. Lastly, recent developments in Transformer-based architecture, particularly Vision Transformers, are discussed to situate their application within the broader literature on image classification. Together, these strands establish the geological and computational framework within which the proposed methodology is developed.

2.1. Dunham carbonate classification scheme

The Dunham classification was developed as a practical, texture-based framework for naming carbonate rocks, with the goal of emphasizing depositional characteristics most relevant to environmental interpretation (Dunham, 1962). The scheme infers depositional energy primarily from the degree of carbonate mud retention and was formulated based on observations from both petrographic thin sections and hand specimens. As such, it is inherently visual in nature and well suited to thin-section image analysis.

Classification within the Dunham system is defined by three key criteria: the supporting fabric of the original sediment, the presence or absence of carbonate mud (particles <20 μm), and evidence of organic binding at the time of deposition. Mud-supported carbonates with less than 10% grains are classified as mudstones, whereas mud-supported rocks with higher grain content are termed wackestones. Grain-supported carbonates containing 1% or more mud are classified as packstones, while those with less than 1% mud are designated as grainstones. Sediments that were biologically bound during deposition are classified as boundstones, and rocks in which the original depositional fabric has been obscured are categorized as crystalline limestones or dolomites. These textural class names were intended by Dunham to function as base terms, commonly modified by grain-type and mineralogical descriptors.

Some modifications of the scheme have been proposed, most notably by Embry and Klovan (1971). Recognizing the difficulties in using the Dunham classification to describe organically bound and coarse-grained limestones, Embry and Klovan expanded the original Dunham classification by subdividing boundstone into framestone, bindstone, and bafflestone, and by introducing the term floatstone for coarse-grained, matrix-supported carbonates. The full application of this scheme require outcrop- and/or core-scale observation. Owing to its systematic approach to carbonate classification and ease of use, the Dunham classification with the Embry and Klovan modification has emerged as the preferred choice among geoscientists over other classification schemes, e.g. Folk (1959) (Lokier & Al-Junaibi, 2016). Figure 1 graphically illustrates the Dunham classification with the Embry and Klovan modification.

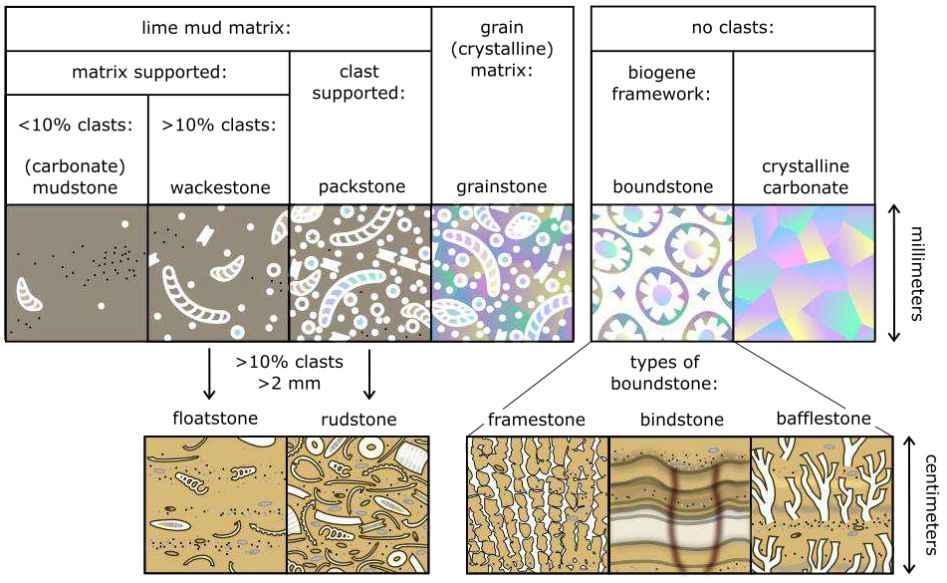


Figure 1: The Dunham (1962) classification with Embry & Klovan (1971) modification.
This figure by Wikipedia user Woudloper is used under a Creative Commons Attribution-ShareAlike 4.0 International License

Because Dunham classifications are defined by visually identifiable textural attributes such as grain support, mud content, and fabric preservation, they provide a target for image-based classification approaches.

2.2. Lokier and Al-Junaibi study: the inconsistency of petrographic description of carbonates

Lokier and Al-Junaibi (2016) demonstrated that substantial inconsistency exists among geoscientists when applying carbonate classification schemes. In their study, they collected the opinions of 241 volunteers with more than 4,200 years of combined experience, who were asked to classify 24 carbonate samples using their preferred classification schemes. Notably, 89% of the classifications employed the Dunham system or its variants.

Their analysis showed that all carbonate samples examined exhibited wide variability in assigned textural names, with individual specimens receiving between 22 and 131 different classifications. This variability occurred not only across different classification schemes, but also within the same scheme, indicating that inconsistencies arise primarily from interpretative differences rather than from terminology alone.

Figure 2 shows the percentage of participants who assigned the most dominant classification to each of the 24 samples in the study. It illustrates that having more than half of the respondents choose a single class for the same sample, let alone achieving consensus, is difficult. The results suggest that consensus may be easier to achieve for grainstone classifications, whereas mudstone, wackestone, and packstone classifications are more divisive.

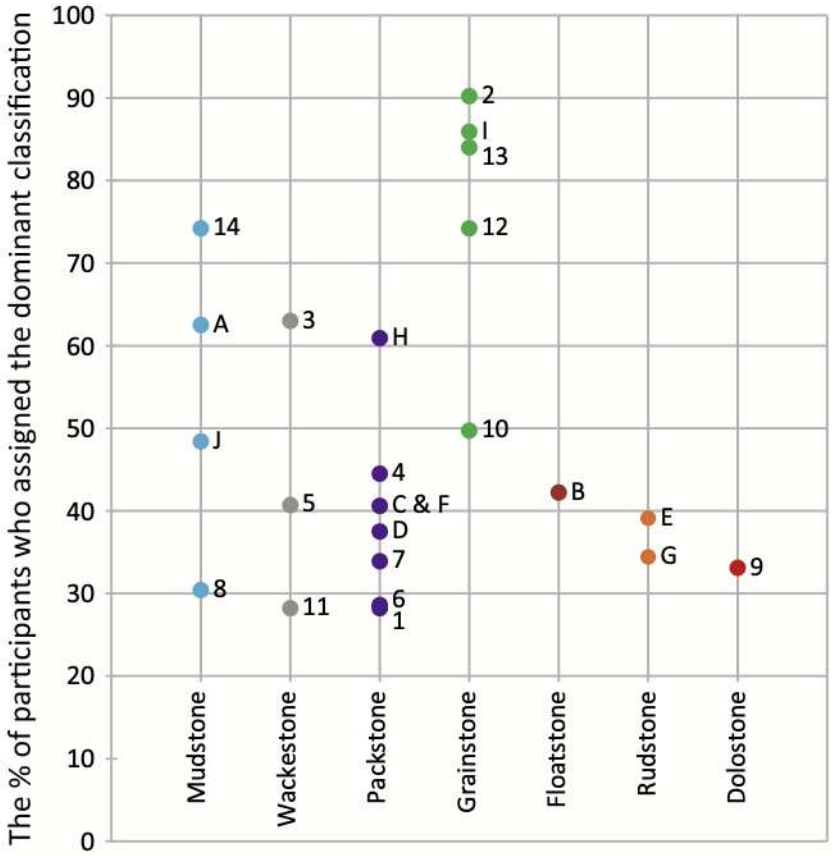


Figure 2: The dominant classification assigned to the 24 carbonate samples by 241 participants in the study (Lokier & Al-Junaibi, 2016)

The authors identified several dominant sources of inconsistency, including errors in determining whether a carbonate fabric is grain-supported or mud-supported, difficulties in estimating grain size and volumetric proportions, and ambiguity in classifying lithologies that display more than one depositional texture. These challenges were further influenced by differences in observer training, professional background, and analytical emphasis, i.e. reliance on thin-section petrography versus hand-specimen or core-scale observations.

Collectively, these findings underscore the subjective nature of carbonate classification. In this context, an automated image-based classification model based on expert consensus offers a potential means of reducing interpretative variability, thereby complementing expert judgment and improving reproducibility in carbonate facies analysis.

2.3. Vision Transformer (ViT)

Previous studies have applied Convolutional Neural Network (CNN) architectures to thin-section analysis. One of the pioneering studies by Koeshidayatullah et al. (2020), used CNN models trained on 4,000 carbonate petrographic images to classify components such as bioclasts, calcite cement, coated grains, dolomite, micrite, and porosity. More recently, Zhao et al. (2025) trained a TLCA-ResNet34 CNN model on 3,116 microscopic images to classify granite, gneiss, schist, dolomite, limestone, tuff, sandstone, gabbro, and basalt. Similarly, Aydın (2025) trained hybrid CNN models combining VGG16 and EfficientNetV2 on 2,634 images to classify a wide range of igneous, sedimentary, and metamorphic rocks. In contrast, this study explores the use of transformer-based architectures as an alternative to CNNs.

Transformers are a class of deep neural network architecture built around self-attention mechanism, originally introduced to model sequential data in natural language processing (NLP) without relying on recurrent operations. Instead of processing inputs step-by-step, the transformer processes all tokens in parallel, using multi-head self-attention to capture global dependencies between inputs in each layer. This architecture was first formalized in the “Attention Is All You Need” paper (Vaswani, et al., 2017), which demonstrated state-of-the-art performance on machine translation (Figure 3).

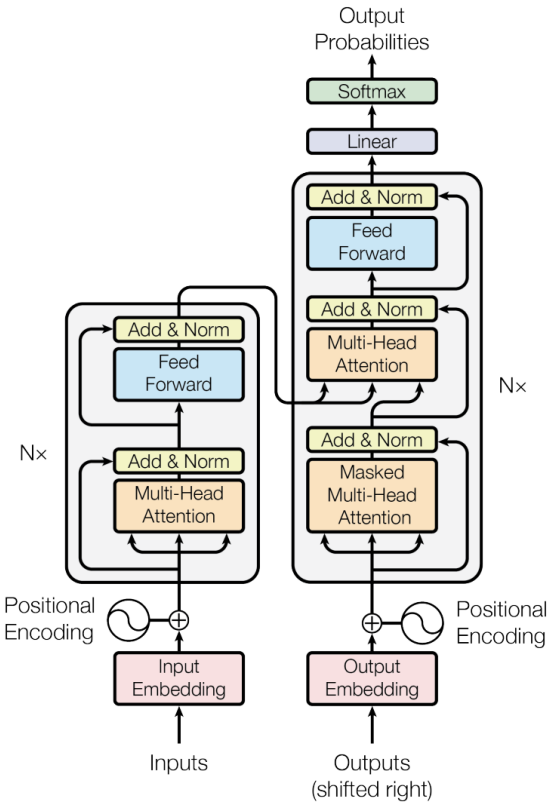


Figure 3: Transformer model architecture (Vaswani, et al., 2017)

The performance of transformer architecture in NLP spurred the adaptation to computer vision, most notably the Vision Transformer (ViT) (Wu, et al., 2020; Dosovitskiy, et al., 2021). Figure 4 shows the architecture of ViT, which repurposes the transformer encoder for visual data by decomposing an image into a sequence of fixed-size patches (e.g., 16 x 16 pixels). Each patch is flattened and projected into an embedding, and positional encodings are added before feeding the sequence to the transformer encoder. A learnable classification token summarizes the sequence, and its final representation is used for image classification. ViT demonstrated that pure transformer architectures can match or exceed the performance of state-of-the-art CNN on image classification benchmarks (Dosovitskiy, et al., 2021).

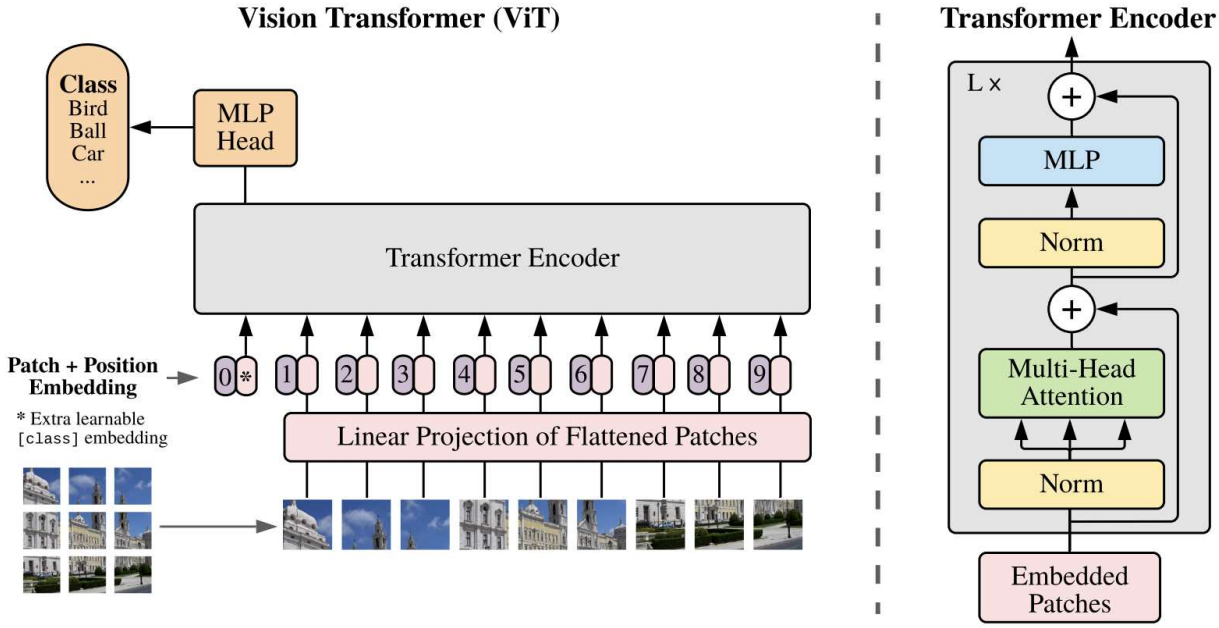


Figure 4: ViT model architecture (Dosovitskiy, et al., 2021)

The rise of transformer-based models in both NLP and computer vision has been supported by open-source tooling that democratizes access to these architectures. An example is the Hugging Face Transformers library (Wolf, et al., 2020), which provides a unified API for a wide range of pre-trained models. The library has enabled researchers to load, fine-tune, and deploy state-of-the-art architectures with minimal code. It is complemented by a model hub that hosts thousands of community-contributed and pre-trained models. For example, the ViT model `vit-base-patch16-224-in21k` (Wu, et al., 2020) fine-tuned on the Lokier and Al-Junaibi (2016) dataset, as demonstrated in this article, is available for use in the author's Hugging Face repository (Aviandito, 2025).

Together, these developments illustrate a shift in computer vision where attention-based architecture offered alternative to CNN architecture. ViT leverages the strengths of the

transformer architecture for image classification, while ecosystems such as Hugging Face ensure that such models are accessible and reusable across domains.

3. Methodology

The data preparation, model configuration, training, and inference steps will be outlined in this section. The core code is implemented in Python using Hugging Face libraries. Readers can consult the full code in the repository.

3.1. Data Preparation

For the fine-tuning process, thin-section photographs from the second phase of the Lokier and Al-Junaibi (2016) study are used. The supplementary archive of their study contains thin-section photographs from 14 samples (Sample 1 through Sample 14). Each sample includes three magnification levels (2x, 4x, and 10x) and two polarizations (plane-polarized light/PPL, and cross-polarized light/XPL), resulting in six JPEG images per sample and a total of 14 samples x 6 images = 84 images. Each of these samples was classified by the respondents according to their own preferred classification system, and the resulting labels are tabulated.

Figure 5 shows a preview of all samples at 2x magnification under PPL, labeled with the most dominant and the second most dominant classifications assigned by the respondents. Table 1 presents the tabulated classification results for the samples. For the fine-tuning process, the most dominant classification is used as the label for each sample, except for Sample 9, where the second most dominant classification is adopted to maintain consistency with the Dunham classification scheme as shown in Figure 1. As shown in Table 1, there is no sample labeled as boundstone, which represents a limitation of the model.

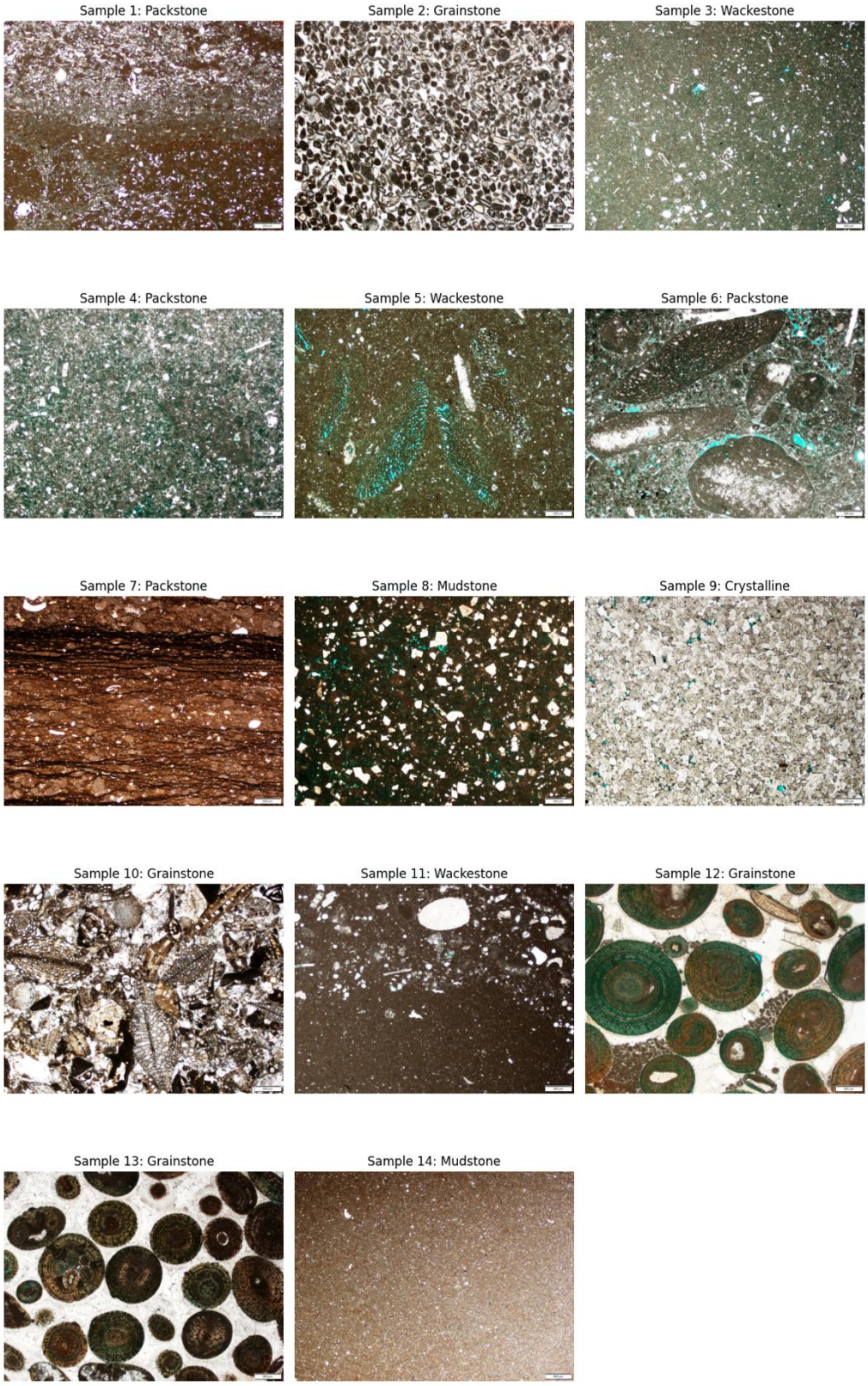


Figure 5: Preview of the dataset at 2x magnification under PPL, labeled with the most dominant classification. Modified from Lokier and Al-Junaibi (2016)

Table 1: The simplified result for each sample of the second phase of the Lokier and Al-Junaibi (2016) study. For the fine-tuning process, the most dominant classification is used as the label for each sample, except for Sample 9, where the second most dominant classification is adopted to maintain consistency with the Dunham classification scheme.

Sample	Number of respondents	Classification Used (% of respondents)		Classification Result (% of respondents)					
		Dunham	Other	Dominant texture	Second	Third	Fourth	Fifth	Sixth
1	177	90.96	9.04	Packstone (28.2%)	Wackestone (9.6%)	Wackestone-packstone (8.5%)	No texture given (6.8%)	Wackestone to packstone (6.2%)	
2	172	88.37	11.63	Grainstone (90.2%)					
3	170	91.76	8.24	Wackestone (63%)	Mudstone (17.7%)				
4	171	91.23	8.77	Packstone (44.5%)	Wackestone (17.5%)	Grainstone (7.6%)			
5	170	90.59	9.41	Wackestone (40.7%)	Floatstone (16.5%)	Packstone (10%)	Floatstone with wackestone matrix (5.9%)		
6	168	92.26	7.74	Packstone (28.6%)	Rudstone (15.5%)	Grainstone (10.1%)	Floatstone (8.9%)	Floatstone with packstone matrix (7.7%)	
7	168	84.52	15.48	Packstone (33.9%)	Wackestone (22%)	No texture given (7.7%)			
8	168	86.31	13.69	Mudstone (30.4%)	Wackestone (25%)	Packstone (14.9%)			
9	166	10.24	89.76	Dolomite (33.1%)	Crystalline (13.9%)	No texture given (11.4%)	Dolostone (10.8%)	Grainstone (6%)	
10	167	89.82	10.18	Grainstone (49.7%)	Rudstone (18%)	Packstone (7.2%)			
11	163	93.25	6.75	Wackestone (28.2%)	Mudstone to wackestone (14.1%)	Mudstone (base) and wackestone (top) (8.6%)	Mudstone/wackestone (6.1%)	Mudstone (5.5%)	Mudstone-wackestone (5.5%)
12	163	86.5	13.5	Grainstone (74.2%)	Oosparite (9.2%)				
13	163	86.5	13.5	Grainstone (84%)	Oosparite (7.4%)				
14	163	90.8	9.2	Mudstone (74.2%)	Wackestone (8%)				

The supplementary archive directory structure is then rearranged into a format suitable for Hugging Face's `imagefolder` dataset loader, which expects a `/<label>/<filename>` structure. Using Python and OS commands, the directory is rearranged from:

```
Supplemental material 1 - Thin section photomicrographs/
├─ Sample 1/
│   ├─ 1 x2 PPL.jpg
│   ├─ 1 x2 XPL.jpg
│   └─ ...
├─ Sample 2/
└─ ...
```

to subfolders labeled Dunham class according to the consensus classification in Table 1 as follows:

```
train/
├─ Packstone/
│   ├─ images from Samples 1, 4, 6, 7
├─ Grainstone/
│   ├─ images from Samples 2, 10, 12, 13
├─ Wackestone/
│   ├─ images from Samples 3, 5, 11
└─ ...
```

Finally, the images are loaded using `datasets.load_dataset("imagefolder")` and the train-test split is done using the `train_test_split` method. The dataset was split into training and test sets using a 90:10 ratio (test_size=0.1), ensuring that model evaluation could be performed on held-out data. The ViT model outputs inference results as numerical IDs; therefore, for interpretability, ID-to-label and label-to-ID mappings must be defined.

3.2. Model selection & adaptation

The choice of a pre-trained model is crucial when working with limited domain-specific data, as it provides learned feature representations that can be fine-tuned for the target task with significantly fewer training examples than would be required for training from scratch (Chollet, 2017). This study employed the `vit-base-patch16-224-in21k` checkpoint as the foundation model (Wu, et al., 2020). This variant was pre-trained on `ImageNet-21k`, a large-scale dataset containing 14 million images across 21,000 classes, providing rich visual representations transferable to specialized domains (Ridnik, Ben-Baruch, Noy, & Zelnik-Manor, 2021). The model architecture consists of:

- Patch size: 16 x 16 pixels

- Input resolution: 224 x 224 pixels
- Hidden dimension: 768
- Number of attention heads: 12
- Number of transformer layers: 12

The pre-trained ViT model was adapted for the Dunham classification task by replacing the classification head with a new linear layer matching the number of target classes. The `AutoModelForImageClassification.from_pretrained()` function from the Transformers library facilitated this adaptation (Wolf, et al., 2020), accepting parameters for the number of labels and the label mappings created during dataset preparation.

3.3. Data Augmentation and Preprocessing

The dataset inherently contains a form of data augmentation through the inclusion of multiple magnifications and polarization modes for each sample. This multi-view representation of each thin section provides the model with diverse perspectives despite the limited number of unique samples.

Additional augmentation was applied through a transformation pipeline using PyTorch's `torchvision.transforms` module (Paszke, et al., 2019). The pipeline consisted of three sequential operations:

1. `RandomResizedCrop`: This transformation randomly crops a portion of the input image with variable area and aspect ratio, then resizes the crop to the required 224 x 224 pixel input dimensions. This technique introduces scale and translation invariance, helping the model learn to recognize diagnostic features regardless of their position or size within the image.
2. `ToTensor`: Converts PIL Image objects to PyTorch tensors, transforming pixel values from the [0, 255] integer range to [0.0, 1.0] floating-point range.
3. `Normalize`: Standardizes pixel values using the mean and standard deviation computed on the ImageNet dataset during pre-training. This normalization ensures that the input distribution matches what the pre-trained model expects.

3.4. Training Configuration

The model is fine-tuned using the free tier of Google Colaboratory (Google Research, 2025), which provides access to a GPU accelerator such as an NVIDIA Tesla T4 when available. In the free tier, runtime memory (RAM) is typically around ~12 GB, and GPU memory depends on the allocated GPU type (e.g., ~16 GB VRAM on a T4 if assigned). Free Colab sessions are subject to time and resource constraints, including auto-disconnects after extended use and variable GPU availability.

To balance learning efficiency with generalization under such constraint, and given the absence of prior benchmarks for the specific case of carbonate thin-section identification, the hyperparameters are set moderately as follows:

- Learning rate: 5×10^{-5} (a moderate rate suitable for fine-tuning pre-trained models)
- Training batch size: 8 images per device
- Gradient accumulation steps: 2 (effective batch size of 16)
- Number of epochs: 20
- Warmup ratio: 0.1 (gradual learning rate increase during the first 10% of training)
- Optimizer: AdamW with β_1=0.9, β_2=0.999, $\varepsilon=1\times10^{-8}$

The model was evaluated at the end of each training epoch using the held-out test set. Accuracy was selected as the primary evaluation metric, defined as the number of correct predictions divided by the total number of predictions. An early stopping mechanism was implemented through the `load_best_model_at_end=True` parameter, which retained the model checkpoint achieving the highest test accuracy.

The Hugging Face `Trainer` class orchestrated the training process, abstracting away low-level details such as device management, gradient computation, and checkpoint saving (Wolf, et al., 2020). The `DefaultDataCollator` was employed to batch samples together, handling the conversion of lists of transformed images into properly shaped tensors. The `AutoImageProcessor` associated with the pre-trained model was passed as the tokenizer parameter, ensuring that any model-specific preprocessing requirements were consistently applied during both training and inference.

The fine-tuned model is uploaded to the author's Hugging Face repository (Aviandito, 2025) and is available for use through the Hugging Face Transformers library.

4. Results

This section presents the evaluation results on the test split. An additional evaluation using a validation dataset unseen by the model is presented in the following section.

4.1. Training result and evaluation on the test split

The model demonstrated rapid convergence during training (Table 2), with test set accuracy improving from 44.44% after the first epoch to 88.89% by the fourth epoch. The training loss decreased consistently from 1.5764 in epoch 1 to 0.2465 in epoch 20, while validation loss declined from 1.5329 to 0.2949 over the same period. The concurrent decrease in both training and validation losses, without clear signs of divergence, suggests that the model was able to learn discriminative features within the Lokier and Al-Junaibi (2016) dataset.

After the eighth epoch, the model achieved 100% accuracy on the test split. However, such performance should be interpreted with caution, as the test split consists of a limited number of images drawn from the same set of samples used for training. Consequently, this result does not necessarily indicate robust generalization to unseen carbonate thin sections. The model may be overfitted to the specific textural characteristics, sample set, and imaging conditions represented in the dataset, and its performance may degrade when applied to a broader range of carbonate lithologies or varying image acquisition settings.

Table 2: Training and evaluation metrics by epoch

Epoch	Training Loss	Test Loss	Accuracy
1	1.576400	1.532907	0.444444
2	1.399100	1.425288	0.555556
3	1.279200	1.285073	0.777778
4	1.011900	1.162515	0.888889
5	0.991600	1.047071	0.888889
6	0.920200	0.983578	0.777778
7	0.699400	0.864909	0.888889
8	0.526000	0.711034	1.000000
9	0.538300	0.612692	1.000000
10	0.512800	0.533678	1.000000
11	0.431200	0.488675	1.000000
12	0.382700	0.436507	1.000000
13	0.345200	0.389063	1.000000
14	0.316400	0.367653	1.000000
15	0.289900	0.355460	1.000000
16	0.287800	0.319668	1.000000
17	0.288400	0.305635	1.000000
18	0.263300	0.310725	1.000000

Epoch	Training Loss	Test Loss	Accuracy
19	0.266900	0.316350	1.000000
20	0.246500	0.294909	1.000000

4.2. Validation

To evaluate model performance more objectively given the limited size of the training and test datasets, an evaluation on a validation dataset of carbonate thin sections unseen by the model is required. It is difficult to find publicly available datasets consisting of consensus-labeled carbonate thin sections similar to the training data; therefore, thin-section photographs from the CarbonateWorld website (Della Porta & Wright, 2025) are used as a validation dataset. To demonstrate the ease of use of a pre-trained model via the Hugging Face Transformers library, the fine-tuned model from the author's Hugging Face repository (Aviandito, 2025) is used in the validation process. A total of 28 thin-section photographs labeled as grainstone, mudstone, packstone, and wackestone were downloaded from the site and are shown in Figure 6, along with their classification result by the model.

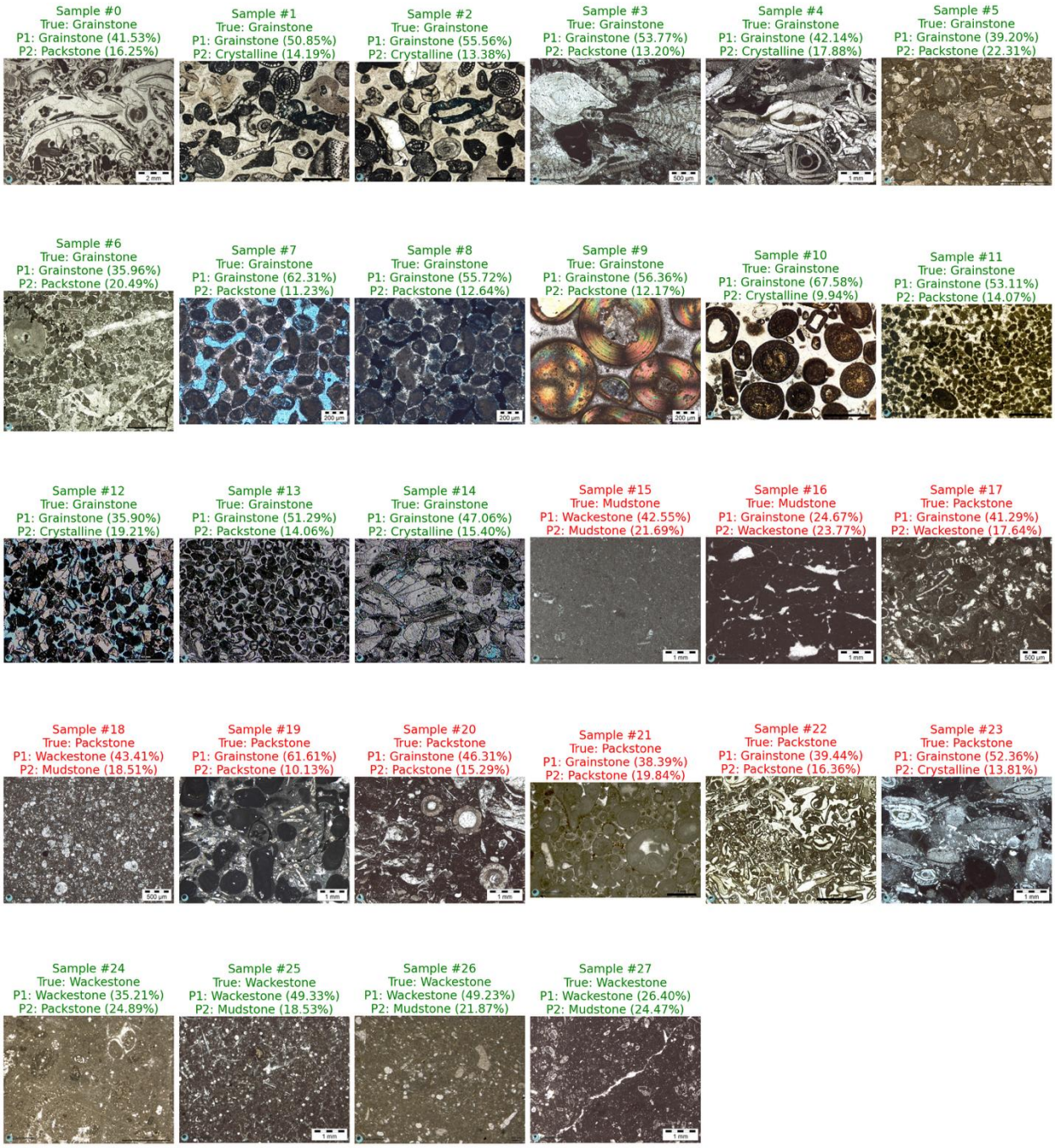


Figure 6: Grainstone, mudstone, packstone, and wackestone samples from CarbonateWorld website (Della Porta & Wright, 2025) used as a validation dataset. Model classification results are shown. Green indicates correct predictions, while red indicates incorrect predictions.

Evaluation on the validation dataset yielded an accuracy of 67.85%, which is lower than the accuracy achieved on the test split derived from the Lokier and Al-Junaibi (2016) dataset. To further examine the model's performance, the confusion matrix (a plot of true versus predicted labels) is shown in Figure 7.

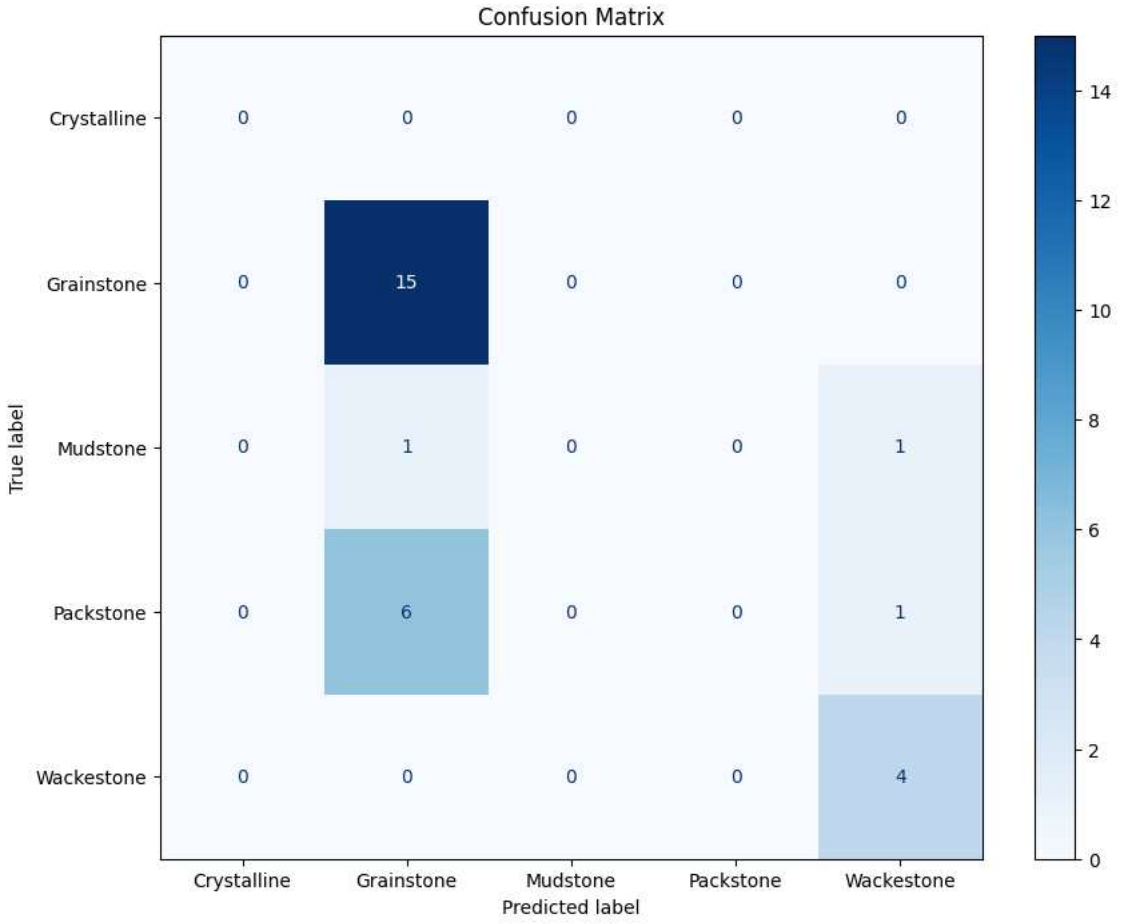


Figure 7: Confusion matrix of the validation dataset, showing significant misclassification of packstones as grainstones

The matrix shows that the model predicted only grainstone and wackestone, despite the presence of mudstone and packstone in the validation dataset. The model correctly classified all grainstones (15 out of 15 images) and wackestones (4 out of 4 images). However, it substantially misclassified packstones as grainstones, with 6 out of 7 packstone images mislabeled as grainstone.

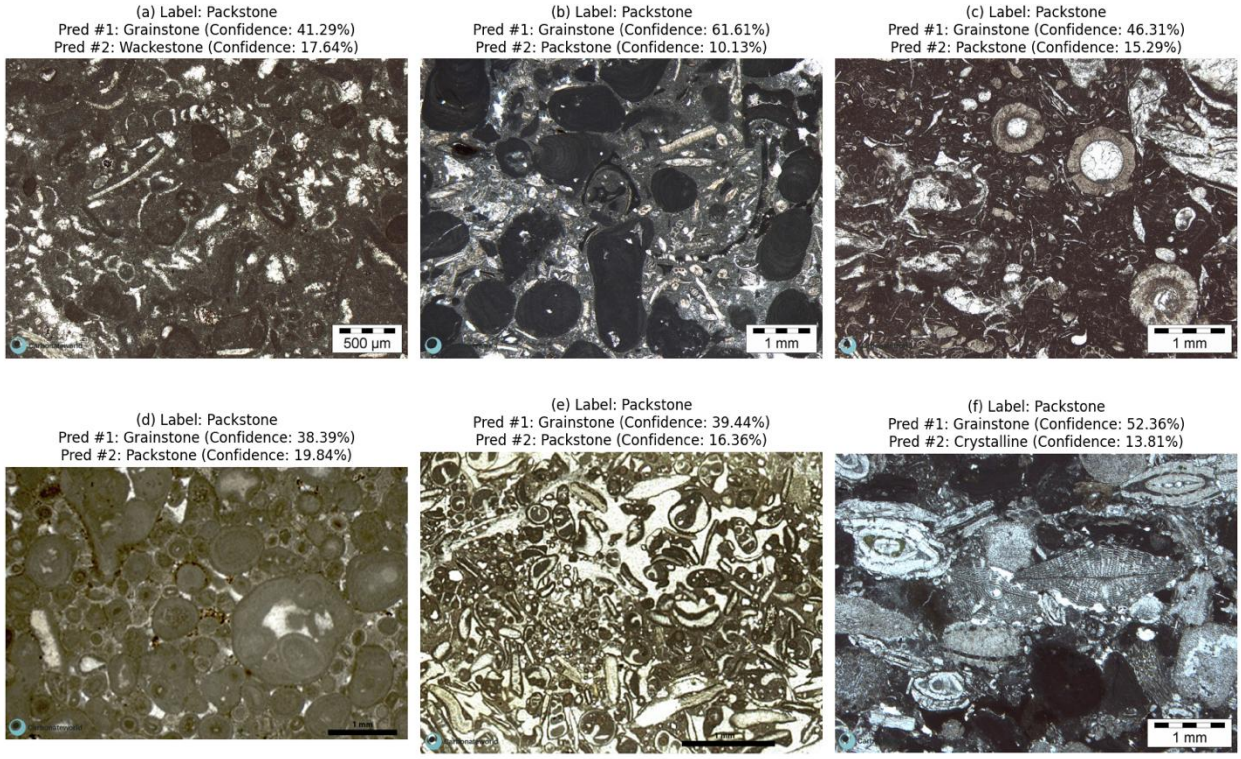


Figure 8: Plots of six packstone samples misclassified as grainstone by the model. Note that the second most probable class in samples (b), (c), (d), and (e) is packstone, which corresponds to the true label. Images are sourced from the CarbonateWorld website. (Della Porta & Wright, 2025)

For further examination, the six packstone images misclassified as grainstones are shown in Figure 8. The classification head of the ViT model outputs a probability distribution across all classes and selects the class with the highest probability as the predicted label. To better understand the model's behavior on the misclassified packstones, the top two predicted class probabilities are also displayed in the figure. Notably, in four out of the six misclassified packstones (samples b, c, d, and e in Figure 8), the class with the second-highest probability is packstone.

This behavior suggests a follow-up analysis in which the top-2 accuracy is evaluated on the entire validation dataset. Instead of relying solely on the class with the highest predicted probability, accuracy is calculated by also considering the class with the second-highest probability. If either the top-ranked or second-ranked class matches the true label, the prediction is counted as correct. Using this approach, the top-2 accuracy reaches 85.71%, representing a substantial improvement over the top-1 accuracy of 67.85%. Table 3 summarizes the full classification results.

Table 3: Full classification results on the grainstone, mudstone, packstone, and wackestone samples from CarbonateWorld website (Della Porta & Wright, 2025)

Sample #	True Label	Top 1 Predicted	Top 1 Confidence	Top 2 Predicted	Top 2 Confidence	Is Top 1 Correct	Is Top 2 Correct
0	Grainstone	Grainstone	41.53%	Packstone	16.25%	1	1
1	Grainstone	Grainstone	50.85%	Crystalline	14.19%	1	1
2	Grainstone	Grainstone	55.56%	Crystalline	13.38%	1	1
3	Grainstone	Grainstone	53.77%	Packstone	13.20%	1	1
4	Grainstone	Grainstone	42.14%	Crystalline	17.88%	1	1
5	Grainstone	Grainstone	39.20%	Packstone	22.31%	1	1
6	Grainstone	Grainstone	35.96%	Packstone	20.49%	1	1
7	Grainstone	Grainstone	62.31%	Packstone	11.23%	1	1
8	Grainstone	Grainstone	55.72%	Packstone	12.64%	1	1
9	Grainstone	Grainstone	56.36%	Packstone	12.17%	1	1
10	Grainstone	Grainstone	67.58%	Crystalline	9.94%	1	1
11	Grainstone	Grainstone	53.11%	Packstone	14.07%	1	1
12	Grainstone	Grainstone	35.90%	Crystalline	19.21%	1	1
13	Grainstone	Grainstone	51.29%	Packstone	14.06%	1	1
14	Grainstone	Grainstone	47.06%	Crystalline	15.40%	1	1
15	Mudstone	Wackestone	42.55%	Mudstone	21.69%	0	1
16	Mudstone	Grainstone	24.67%	Wackestone	23.77%	0	0
17	Packstone	Grainstone	41.29%	Wackestone	17.64%	0	0
18	Packstone	Wackestone	43.41%	Mudstone	18.51%	0	0

19	Packstone	Grainstone	61.61%	Packstone	10.13%	0	1
20	Packstone	Grainstone	46.31%	Packstone	15.29%	0	1
21	Packstone	Grainstone	38.39%	Packstone	19.84%	0	1
22	Packstone	Grainstone	39.44%	Packstone	16.36%	0	1
23	Packstone	Grainstone	52.36%	Crystalline	13.81%	0	0
24	Wackestone	Wackestone	35.21%	Packstone	24.89%	1	1
25	Wackestone	Wackestone	49.33%	Mudstone	18.53%	1	1
26	Wackestone	Wackestone	49.23%	Mudstone	21.87%	1	1
27	Wackestone	Wackestone	26.40%	Mudstone	24.47%	1	1

5. Discussion

A primary limitation of this study lies in the size and scope of the training dataset. The Lokier and Al-Junaibi (2016) dataset comprises a small number of carbonate samples, albeit carefully curated and labeled through expert consensus. As such, the results presented here should be interpreted as a proof-of-concept demonstrating the feasibility of applying Vision Transformers to carbonate thin-section classification, rather than as evidence of broad generalization across carbonate lithologies.

In practice, organizations operating in industry settings are well positioned to overcome this limitation by fine-tuning the model on their own proprietary databases of carbonate thin sections classified by in-house expert geoscientists. Such datasets often span a wider range of depositional environments and imaging conditions, allowing models to learn more robust and transferable representations. In contrast, independent researchers face a structural constraint due to the scarcity of publicly available, consensus-labeled petrographic datasets.

Visual inspection of the six misclassified packstone images suggests that these samples exhibit textural characteristics that are transitional between the two classes, such as low mud content and strong grain support. This criterion is inherently difficult to assess visually and subject to interpretative variability, as previously documented by Lokier and Al-Junaibi (2016). The model's behavior reflects this ambiguity. In four of the six misclassified cases, packstone appears as the second most probable class, indicating that the model captures relevant features for both categories but struggles to separate them. This observation suggests that secondary prediction probabilities provide meaningful information and could be incorporated into probabilistic or uncertainty-aware classification workflows, particularly for borderline textures. Such an approach aligns well with the interpretive nature of carbonate classification, where expert judgment often acknowledges gradational boundaries rather than discrete classes.

Several strategies could be explored to improve discrimination between packstone and grainstone. These include targeted data augmentation focused on subtle mud-content variations, class-specific reweighting or resampling to address imbalance, and explicit handling of class imbalance as suggested by Koeshidayatullah et al. (2020). Increasing the number of representative samples for these closely related classes would likely yield the greatest benefit.

An important direction for future work is the incorporation of explainable AI (XAI) techniques to better understand model decision-making. Vision Transformers are particularly amenable to interpretability through attention visualization, which can highlight image regions that contribute most strongly to classification decisions. Methods such as attention rollout and attention-weight visualization have been proposed to

aggregate attention weights across transformer layers and identify influential regions in the input image (Abnar & Zuidema, 2020; Chefer, Gur, & Wolf, 2021).

In the context of petrographic thin-section analysis, attention visualization could help assess whether the model focuses on meaningful features, such as grain/mud support, when assigning Dunham classes. Such insight would be particularly valuable for borderline cases, including the observed confusion between packstone and grainstone, where subtle textural differences drive classification decisions. Incorporating XAI techniques would not only enhance interpretability and trust in automated classification systems but also provide a feedback mechanism for geoscientists to evaluate whether model predictions align with established petrographic reasoning (Samek, Montavon, Lapuschkin, Anders, & Müller, 2021).

While this study focuses on whole-image classification, carbonate petrography often involves identifying and quantifying specific components within a thin section. Extending this work from image-level classification to object detection or segmentation represents a natural progression. Prior studies, such as Koeshidayatullah et al. (2020), have demonstrated the utility of object detection for identifying bioclasts, calcite cement, coated grains, dolomite, micrite, and porosity, and similar approaches could be explored using transformer-based architectures.

Finally, this work underscores the growing importance of geoscientific data ownership and management in the era of modern machine learning. As model architectures and training workflows become increasingly democratized and computational hardware becomes more accessible, the primary determinant of model performance shifts toward the quality, diversity, and curation of the underlying training data. This research highlights the importance of developing and releasing open, expert-labeled image datasets, which can serve not only as training resources but also as standardized benchmarks for comparing model performance across studies.

Open competitions organized by professional societies, e.g. a competition hosted by the Society of Petrophysicists and Well Log Analysts (SPWLA) in 2020 to predict sonic log from “triple combo” petrophysical measurements (Yu, et al., 2021), demonstrate how open datasets can catalyze methodological innovation and cross-disciplinary collaboration. Similar efforts focused on petrographic image data would significantly accelerate progress in automated rock classification and foster more transparent and cumulative research in computational geoscience.

6. Conclusion

This study has shown that a Vision Transformer can be fine-tuned to classify carbonate thin-section images according to Dunham’s scheme using the Lokier and Al-Junaibi (2016) dataset. The key aspects of this work include leveraging open-access, expert-

labeled data for training, employing open-source libraries to ensure ease of use, and fine-tuning a widely used pre-trained Vision Transformer model. The methodology is fully documented, with complete code provided for each stage, including data loading, augmentation, model training, and inference, and is intended as a methodological demonstration rather than a performance benchmark.

The results suggest that texture-based carbonate classes are learnable by modern image-based models. However, given the limited dataset size, further fine-tuning and testing on additional samples are recommended before deployment. Future work could include expanding the training set using labels derived from expert consensus.

In summary, this study demonstrates a reproducible pipeline for automating petrographic classification. By combining the Dunham framework with Vision Transformers, this work seeks to bridge sedimentology with contemporary AI experimentation and highlights the value of open science in advancing geologic image analysis.

Appendix A: Supplementary data

- The code can be found online at: https://github.com/aviandito/vit-dunham-carbonate-classifier/tree/main
- The fine-tuned model can be found online at: https://huggingface.co/aviandito/vit-dunham-carbonate-classifier
- The training dataset can be found online at Lokier and Al-Junaibi (2016) Data S1: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fsed.12293&file=sed12293-sup-0001-SupInfo.zip
- The validation dataset can be found online at CarbonateWorld website: https://carbonateworld.com/

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used ChatGPT and Claude to proofread the manuscript and Gemini to assist with the code debugging. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Declaration of Ethics, Consent to Participate, and Consent to Publish

Ethics, Consent to Participate, and Consent to Publish declarations are not applicable.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Abnar, S., & Zuidema, W. (2020). Quantifying Attention Flow in Transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4190–4197). Association for Computational Linguistics: https://doi.org/10.18653/v1/2020.acl-main.385.

Aviandito, M. (2025). *aviandito/vit-dunham-carbonate-classifier.* Retrieved from Hugging Face: https://doi.org/10.57967/HF/7237

Aydın, İ., Şener, T. K., Kılıç, A. D., & Derviş, H. (2025). Classification of Thin-Section Rock Images Using a Combined CNN and SVM Approach. *Minerals, 15*(9), 976. Retrieved from https://doi.org/10.3390/min15090976

Chefer, H., Gur, S., & Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 397-406). Montreal: http://doi.org/10.1109/ICCV48922.2021.00045.

Chollet, F. (2017). *Transfer learning and fine-tuning .* Retrieved from TensorFlow Documentation: https://www.tensorflow.org/tutorials/images/transfer_learning

Google Research. (2025). *Colaboratory Frequently Asked Questions*. Retrieved from Google Research: https://research.google.com/colaboratory/faq.html

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale . *International Conference on Learning Representations.* Vienna: https://doi.org/10.48550/arXiv.2010.11929.

Dunham, R. J. (1962). Classification of carbonate rocks according to depositional texture. *American Association of Petroleum Geologists Memoir Vol. 1*, pp. 108-121.

Folk, R. L. (1959). Practical petrographic classification of limestones. *AAPG Bulletin, 43*, 1-38.

Embry, A., & Klovan, J. (1971). A Late Devonian Reef Tract on Northeasterm Banks Island, NWT. *Bulletin of Canadian Petroleum Geology, 19*, 730-781.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Chintala. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver: https://doi.org/10.48550/arXiv.1912.01703.

Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021). ImageNet-21K Pretraining for the Masses . *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021).* Retrieved from https://doi.org/10.48550/arXiv.2104.10972

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining Deep Neural Networks and beyond: A Review of Methods and Applications. *Proceedings of the IEEE, 109*(3), 247-278. Retrieved from https://doi.org/10.1109/JPROC.2021.3060483

Lokier, S. W., & Al-Junaibi, M. (2016). The petrographic description of carbonate facies: are we all speaking the same language? *Sedimentology, 63*, 1843-1885. Retrieved from https://doi.org/10.1111/sed.12293

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, (pp. 6000-6010). Long Beach, CA. Retrieved from https://doi.org/10.48550/arXiv.1706.03762

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Xu, C. (2020). *Transformers: State-of-the-Art Natural Language Processing*. Retrieved from 10.18653/v1/2020.emnlp-demos.6

Yu, Y., Xu, C., Misra, S., Li, W., Ashby, M., Pan, W., . . . Iza. (2021). Synthetic Sonic Log Generation With Machine Learning: A Contest Summary From Five Methods. *Petrophysics, 62*(4), 393-406. Retrieved from https://doi.org/10.30632/PJV62N4-2021a4

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., . . . Vajda, P. (2020). Visual Transformers: Token-based Image Representation and Processing for Computer Vision. Retrieved from https://doi.org/10.48550/arXiv.2006.03677

Zhao, Z., Tan, S., Chen, H., Wang, P., Zhang, Q., Wei, H., & Zhang, Z. (2025). Classification of microscopic images of rock thin sections based on TLCA-ResNet34. *Applied Computing and Geosciences, 27*, 100272. Retrieved from https://doi.org/10.1016/j.acags.2025.100272

Della Porta, G., & Wright, P. (2025). *Images*. Retrieved from CarbonateWorld: https://carbonateworld.com/

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- vitdunhamnotebook.ipynb