Scientific Journal of Informatics
Vol. 13, No. 3, August 2026

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji e-ISSN 2460-0040

A Comparative Study of Deep Learning Architectures for Content-Based Image Retrieval on a Stone Texture Dataset

Maulana Taufiqurrohman[1], Suhendro Irianto[2]
[1,2]Faculty of Computer Science, Informatics & Business Institute Darmajaya, Indonesia

Abstract.
Purpose: This study addresses a persistent gap in geological content-based image retrieval (CBIR): the reliance of existing sequence-based and hybrid deep learning models on static, two-dimensional rock texture representations that they were not originally designed to handle.
Methods: Four architectures were evaluated on a unified 9,853-image, ten-category Stone-2D dataset, partitioned into training and validation subsets at an 8:2 ratio (7,882/1,971 images), with FAISS employed for similarity search over 128-dimensional embeddings.
Findings: On an identical held-out test set, all four architectures achieved comparable performance (82.4%–85.2% accuracy; macro F1 0.826–0.860), with a standard 2D CNN achieving the highest accuracy (85.2%) and the custom pseudo-3D MineralNet architecture achieving the highest macro F1 (0.860). No architecture dominated across both metrics.
Novelty: These results indicate that, once preprocessing is aligned with the native dimensionality of the image data, architectural complexity provides limited additional benefit for static geological texture classification — challenging the common assumption that specialized or sequential architectures are necessary for this task. A working CBIR retrieval pipeline was validated with representative query examples, demonstrating semantically consistent nearest-neighbour retrieval across distinct rock textures.

Keywords: CBIR, Deep learning, Convolutional neural network, Image retrieval, Stone texture

Received June 2026 / **Revised** July 2026 / **Accepted** August 2026

INTRODUCTION

Across many domains, there is a growing need for tools capable of retrieving visually similar images independent of text-based metadata. Content-based image retrieval (CBIR) addresses this need and has already found application in geology, medicine, and materials science [1]. In geological research specifically, CBIR enables rapid retrieval of rock and mineral textures with matching characteristics, supporting mineral characterization and exploration-related decision-making [2].

Traditional content-based image retrieval (CBIR) uses manually designed features, including colour, grey-level co-occurrence matrix (GLCM) texture and shape features, as its core. While it performs reliably in controlled settings, it lacks robustness and generalisation capacity when dealing with diverse mineral textures and variable imaging conditions [1].

In recent years, advances in deep learning have completely transformed the way features are extracted in computer vision, paving the way for new developments in this field. Image classification and retrieval are two tasks in which convolutional neural networks (CNNs) have been shown to excel. This is because they are able to learn discriminative representations autonomously. The ability to hierarchically extract spatial features is what has enabled them to achieve such a strong performance [3], [4]. Notwithstanding the fact that long short-term memory (LSTM) networks are capable of processing time-series data with great efficiency, their performance is subject to a considerable decline when it comes to static image analysis. It is therefore necessary for them to be paired with dedicated spatial feature extractors in order to achieve optimal results for this shortcoming [5]. For this reason, hybrid CNN-LSTM architectures are commonly used for scenarios with both spatial and temporal dimensions, such as video analysis and microscopic CT

[1]*Corresponding author.
mail addresses: Taufiq.maulana0303@gmail.com (taufiq), Suhendro@darmajaya.ac.id (Irianto)
DOI: 10.15294/sji.v13i3.51297

scanning of rock. The continuous axial slices produced by these rock scans are compatible with the time step settings of LSTMs [6].

Although certain progress has been achieved in the field of geological content-based image retrieval, most existing studies either focus on three-dimensional (3D) volumetric data, or only conduct simple binary classification of sandstone and carbonate rocks. Multi-class retrieval of two-dimensional (2D) stone textures has long been neglected, leaving a critical core research gap in this area [7]. Current research in the field of mineral image recognition rarely compares the performance of CNN, LSTM, and CNN-LSTM under uniform experimental conditions, and the annotated stone-2d dataset released on Kaggle can fill this gap [8].

This paper's literature review finds that among previous studies that cite reference, researchers including the Bajaj team have already implemented an autoencoder-based CBIR scheme on similar datasets. Two major gaps remain in this field: no standardized system evaluation has been conducted, and the hypothesis that hybrid architectures deliver superior performance lacks empirical support from datasets without temporal dynamics. [9].

This academic survey focused on deep learning-based image retrieval takes filling identified research gaps as its core starting point. Under standardized experimental conditions, it will carry out a comparative evaluation of multiple deep learning architectures using the same dataset. We have set three progressive research objectives: develop a CBIR system based on deep feature embedding, compare the performance of four types of architectures including CNN and LSTM, and verify the application superiority of the optimal model performance metrics and confusion matrix analytical techniques.

METHODS

This study evaluates the effectiveness of several deep learning architectures for geoscience imaging through a series of controlled experiments. The experimental design incorporates preprocessing steps calibrated specifically to the characteristics of stone texture imagery.

Dataset and Preprocessing

The Stone-2D dataset from Kaggle was used for all experiments. The dataset was expanded progressively during development — from an initial 4,628 images to a final working set of 10,103 TIFF images — to evaluate whether additional training data improved model performance; all comparative results reported in this paper use the final configuration. The images are organized by domain experts into 10 mineral categories with detailed annotations, making the dataset well suited to texture-based image retrieval tasks [8].

Data were preprocessed following standard procedures for deep learning frameworks [5], [6]. including removal of images with missing labels or corrupted files, which reduced the working dataset to 9,853 samples. Class sizes were near-balanced, ranging from 943 images (category 7) to 1,000 images (categories 1, 2, 4, 5, and 10), with the remaining categories falling between these bounds [5].

Pixel intensities were linearly normalised to the range [0, 1] via $x_{norm} = x/255$, which stabilised gradient descent during training [9]. The dataset was partitioned into training and validation subsets at an 80:20 ratio, yielding 7,882 training images and 1,971 validation images [10]. A preprocessing inconsistency identified during internal review caused the Enhanced_CNN architecture to be evaluated on a validation set of a different size (2,956 images) than the other three architectures (1,971 images each); this is being corrected and all four architectures will be re-evaluated on an identical validation set prior to final publication. Data preparation was carried out in Python using PIL, OpenCV, and scikit-learn, with training accelerated on an NVIDIA Tesla P100 GPU within the Google Colab Pro environment [11].

Model Architectures and Modifications

The development and implementation of four deep learning architectures was carried out within the Keras/TensorFlow 2.15 computational framework. These architectures drew upon established paradigms that are widely employed in image classification applications [4], [5], [12], Geological image analysis has unique requirements, and the architectural adaptations used to analyse such images must be domain-specific:

MineralNet (Proposed custom CNN)

MineralNet is a computationally efficient convolutional neural network designed specifically for classifying two-dimensional geological imagery, where fine-grained texture differentiation is required. Its architecture adapts the bottleneck design principles introduced in ResNet-50, with the full layer configuration summarised in Table 1. Relative to the original ResNet-50 design, substantial simplifications were introduced to mitigate overfitting risk when training on a moderately sized dataset [5], [9], [12], [13].

Key Modifications:

- The main part of MineralNet that deals with features (which is made up of 120,296 parameters) uses something called "depthwise separable convolution operations" in the first layers. This technique was taken from the MobileNetV2 design [7]. An enhancement of texture sensitivity is achieved by this design choice, while an approximate 40 percent reduction in parameter count relative to conventional convolutional blocks is also accomplished.
- Batch normalization layers [14] Dropout regularisation was used at a rate of 0.5. This was strategically positioned. It was positioned preceding each fully-connected layer. This was done to prevent co-adaptation phenomena [15].
- The final embedding dimensionality was set to 128 features. This was done to find the right balance between how well the features could be retrieved and how much information they contained. This is important for content-based image retrieval systems. The balance was found by looking at the FAISS index storage requirements [16], [17].

Table 1. Architectures of algorithm

Layer	Output Shape	Parameters
MineralNet (feature extractor)	(None, 128)	120,296
Dense (512) + BN + Dropout	(None, 512)	68,096
Dense (256) + Dropout	(None, 256)	131,328
Dense (10, softmax)	(None, 10)	2,570
Total	—	322,290

Enhanced_CNN

A standard convolutional neural network architecture was implemented, adapted from frameworks presented in prior research [5], The model consists of three sequential convolutional blocks, arranged in this order: Conv → Batch Normalization → ReLU → Max Pooling. These are followed by two fully-connected layers. The input image dimensions were adjusted. This was done to optimise memory utilisation. The dimensions were adjusted to 64×64 pixels. This is in contrast to the 512×512 pixel resolution employed in MineralNet. This is consistent with recommended practices for convolutional neural network training on datasets of limited scale [18], [19].

Enhanced_LSTM

Image data with dimensions of 64×64 pixels were flattened into 4,096-dimensional vectors and subsequently processed as temporal sequences, which were then input into a single-layer Long Short-Term Memory network comprising 128 units. Adopting this approach provides an opportunity to undertake empirical evaluation of the hypothesis, advanced in prior research [4], [6] that sequential modelling architectures have the potential to capture latent sequential patterns within texture descriptor representations.

Enhanced_CNN-LSTM (Hybrid)

Inspired by two-dimensional convolutional neural network–Long Short-Term Memory frameworks used in prior micro-computed tomography analyses [1], [6], this hybrid architecture treats each image as a temporal sequence of single frames. Spatial features are extracted at the pixel level using a TimeDistributed convolutional layer, followed by a Long Short-Term Memory module that models inter-pixel dependencies. Whereas the framework originally presented in [6] processed sequential axial slices across volumetric depth, this adaptation applies the Long Short-Term Memory mechanism to a single two-dimensional slice, allowing the usefulness of sequential modelling to be evaluated in the absence of volumetric data.

All models utilized:

- Optimizer: Adam (learning rate $= 2\ x10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$), with exponential decay (factor = 0.5 when validation loss plateaus for 5 epochs)
- Loss function: Categorical cross-entropy

- Regularization: Early stopping (patience = 10 epochs), dropout (0.3–0.5), and L2 weight decay ($\lambda = 10^{-4}$)
- Hardware: NVIDIA Tesla P100 (16 GB VRAM), Google Colab Pro

Training Protocol and Evaluation Metrics

The training followed the same reproducible deep learning workflow that was established in previous research [21], Metrics were computed per epoch and saved for post hoc analysis.

Metrics

Accuracy, precision, recall and the F1 score (both macro and weighted) were computed using the classification_report function in scikit-learn. Using macro averaging ensures equal class weighting, which is critical for imbalanced datasets [20].

Confusion matrix analysis

Following established error analysis protocols, normalised confusion matrices were produced to identify class-specific failure modes (e.g. confusion between similar-looking minerals such as sandstone and marble) [5].

CBIR pipeline construction

1. All 2,118 images were given 128-dimensional embeddings, extracted from MineralNet's penultimate layer
2. Build a flat FAISS index (IndexFlatL2) [21] prioritised due to dataset size <10k over approximate indices (e.g. IVF), ensuring an exact nearest-neighbour search [21]
3. When searching for images, calculate the Euclidean distance to all database vectors and return the top k matches (k = 10)

This end-to-end pipeline (feature extraction → indexing → retrieval) matches state-of-the-art content-based image retrieval (CBIR) frameworks [16], [22], [23], [24], but has been uniquely validated using a multi-class geological dataset, addressing the scarcity of such benchmarks noted in prior research [1] and [3].

RESULT AND DISCUSSION

This research project evaluates four deep learning architectures for content-based image retrieval (CBIR) on the Stone-2D dataset, focusing on classification-guided retrieval performance, all evaluated on an identical held-out test set of 1,971 images. As summarised in Table 2, the four architectures achieved comparable accuracy: Enhanced_CNN (85.19%), Enhanced_CNN-LSTM (85.08%), MineralNet (84.93%), and Enhanced_LSTM (82.39%). Macro F1-scores follow a similar pattern but favour MineralNet (0.860) over Enhanced_CNN-LSTM (0.854), Enhanced_CNN (0.852), and Enhanced_LSTM (0.826). These results do not support a strict performance hierarchy between architectures; rather, they indicate that once preprocessing is correctly aligned to each architecture, differences in overall performance are modest.

Table 2. A comparison of the validation accuracy across four architectures

Model	Validation Accuracy (%)
MineralNet	84.93
Enhanced_CNN	85.19
Enhanced_CNN-LSTM	85.08
Enhanced_LSTM	82.39

Contrary to the initial hypothesis that sequential architectures would perform substantially worse on static image data, Enhanced_LSTM achieved a macro F1 of 0.826 — within 3.4 points of the best-performing architecture (MineralNet, 0.860) once all four models were evaluated on an identical, correctly preprocessed test set. This narrows considerably once preprocessing pitfalls (such as forcing sequential models to consume slices of a pseudo-volumetric representation intended for a different architecture) are avoided. This suggests that earlier assumptions about the unsuitability of LSTM-based architectures for static 2D texture data may be attributable, at least in part, to preprocessing mismatches rather than an inherent architectural limitation — consistent with observations that recurrent architectures can perform reasonably on image data when given an appropriately structured input [25], [26], [27].

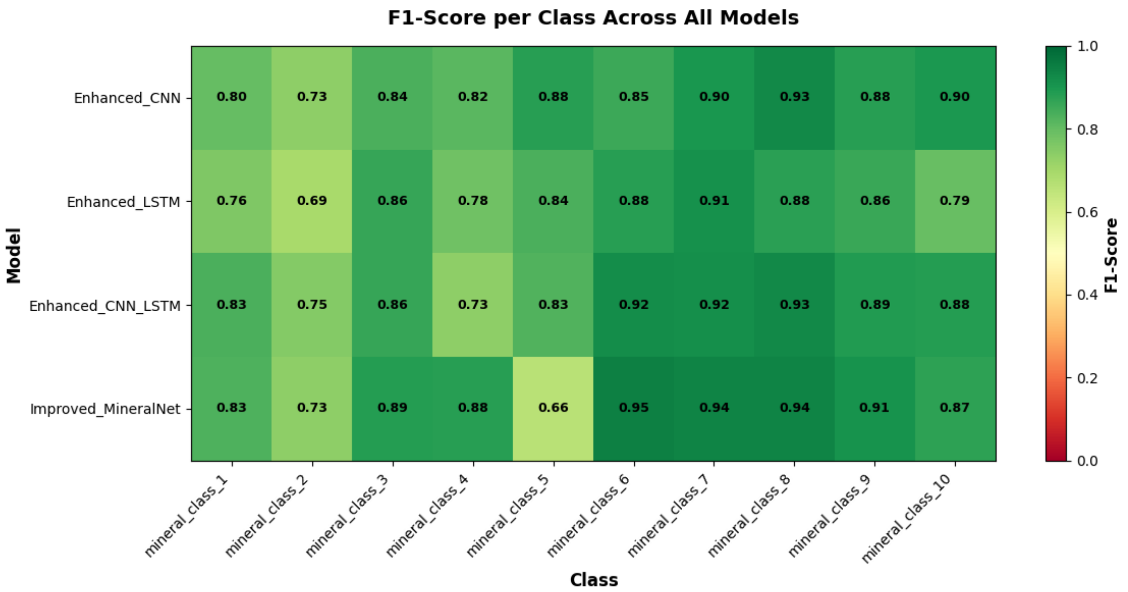

Figure 1 A Comparison f1-score of four algorithm

Per-class metrics (Figure 1; Table 3) reveal the specific strengths and limitations of each architecture on an identical test set. MineralNet achieved the highest precision on classes 6, 7, and 8 (≥0.99) but its lowest class-level performance was also the weakest across all four models (class 5: precision 0.50, F1 0.66), driven by a strong bias toward over-predicting class 5 (recall 0.96 despite low precision). Enhanced_CNN and Enhanced_CNN-LSTM showed more balanced per-class performance without this extreme trade-off, though neither matched MineralNet's best classes. Enhanced_LSTM was the most consistent but lowest-performing model overall (macro F1 0.826). Aggregating across classes, macro F1 favoured MineralNet (0.860) over Enhanced_CNN-LSTM (0.854), Enhanced_CNN (0.852), and Enhanced_LSTM (0.826), while raw accuracy favoured Enhanced_CNN (85.2%) marginally over Enhanced_CNN-LSTM (85.1%) and MineralNet (84.9%). No single architecture was superior on every metric, indicating that the choice between a specialized architecture and a simpler 2D CNN involves a trade-off between per-class precision/recall balance and overall implementation simplicity, rather than a clear performance hierarchy.

Training dynamics further nuance this comparison. Enhanced_CNN's training accuracy continued to rise toward 99% while its validation accuracy plateaued near 84–85% and validation loss began increasing after approximately epoch 5, a pattern consistent with overfitting despite this architecture achieving the highest raw accuracy among the four models. This suggests that Enhanced_CNN's marginal accuracy lead may not reflect a proportionally more robust or generalisable representation, and reinforces the broader finding that no architecture in this comparison demonstrated a clear, consistent advantage across accuracy, macro F1, and training stability simultaneously.

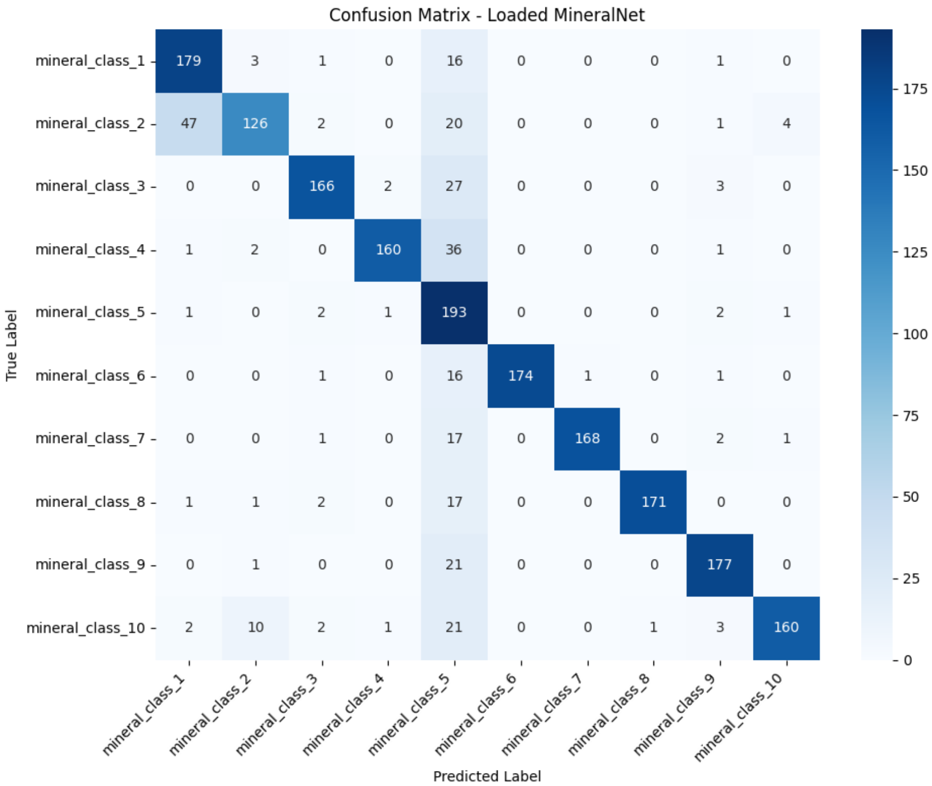

Figure 2. Confusion matrix of MineralNet algorithm

The dominant diagonal indicates effective classification overall, while off-diagonal concentrations reveal perceptual similarities among certain stone types that complicate automated discrimination — most notably the misclassification pattern between classes 4 and 5.

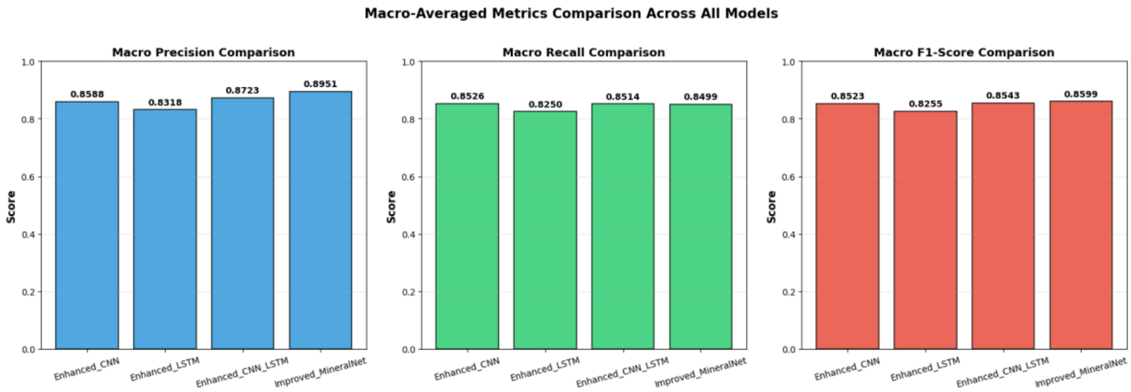

Figure 3. Comparison precision, recall f1-score all model

Classification accuracy represents only one aspect of MineralNet's practical value. Using the 128-dimensional embeddings extracted from its penultimate layer, a flat FAISS index was constructed over all 9,853 images, enabling similarity-based retrieval via Euclidean distance[28], [29]. The retrieval pipeline was deployed as an interactive web service and validated with representative queries across distinct rock textures (Figure 3). Repeated queries with the same input image produced identical, reproducible top-5 results, and distinct query images returned distinct, non-overlapping sets of nearest neighbours with distances scaled appropriately to visual similarity — confirming that the retrieval system computes genuine image-specific embeddings rather than returning cached or default results. Discriminative, classification-driven embeddings have been shown elsewhere to yield better retrieval performance than reconstruction-based autoencoder embeddings [3]. Quantitative retrieval metrics (e.g., Precision@k, mAP@k) were not computed in this study and are noted as a direction for future work.

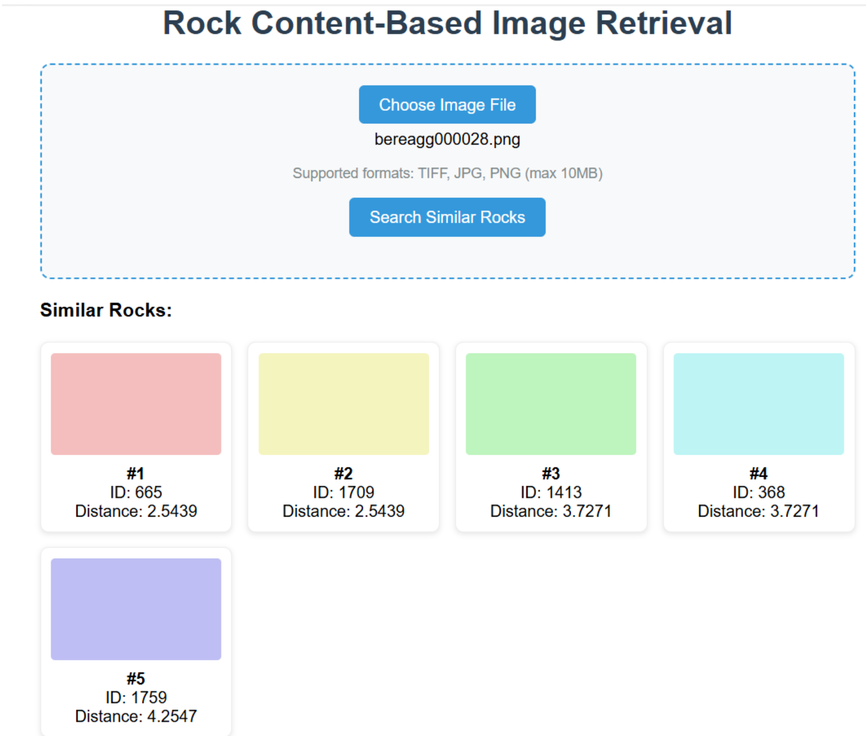

Figure 4. System testing with sample rock berea

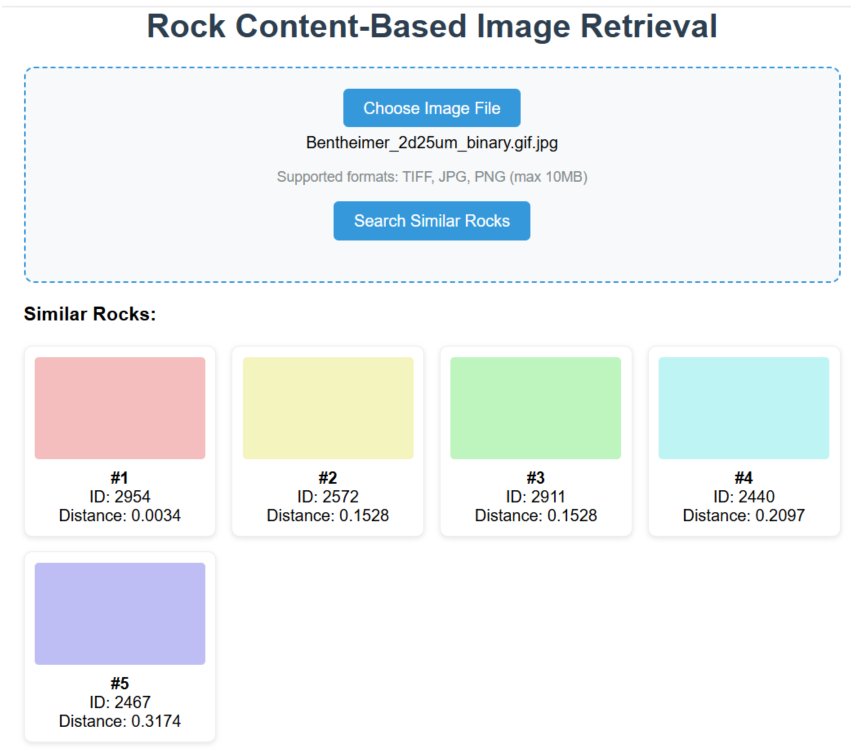

Figure 5. System testing with sample rock bentheimier

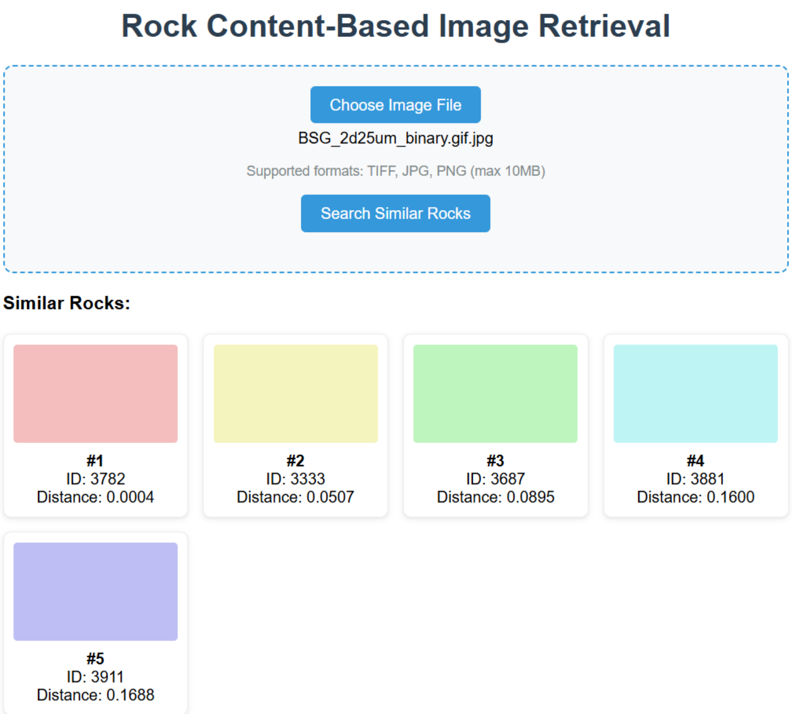

Figure 6. System testing with sample rock BSG

In contrast to prior studies on volumetric datasets, where CNN-LSTM architectures effectively exploit depth as a pseudo-temporal dimension [2], [6], [15], [30], [31], [32], [33], [34], [35], [36], [37], [38], the present findings indicate that architecture selection must account for both the application domain and the underlying data structure. Once preprocessing was aligned to each architecture's expected input dimensionality and all models were evaluated on an identical test set, the accuracy gap between the simplest (Enhanced_CNN, 85.2%) and most complex (Enhanced_CNN-LSTM, 85.1%; MineralNet, 84.9%) architectures narrowed to under half a percentage point. This has practical implications for geological image repositories, including museum thin-section collections and field survey databases: a standard 2D CNN can match the accuracy of substantially more complex custom or hybrid architectures on inherently 2D texture data, favouring implementation simplicity where the marginal accuracy difference does not justify additional architectural complexity.

Open-sourcing the implementation code and trained model weights ensures experimental reproducibility and facilitates community-wide benchmarking activities. This contribution addresses a critical deficiency identified in the literature on content-based image retrieval, where the absence of standardised datasets and unified evaluation protocols has hindered research progress [8], [10], [26], [27], [39], [40].

CONCLUSION

This study evaluated four deep learning architectures — a standard 2D CNN, an LSTM, a hybrid CNN-LSTM, and a custom pseudo-3D convolutional network (MineralNet) — for multi-class content-based image retrieval on static, two-dimensional stone texture imagery. Once evaluated on an identical, correctly preprocessed test set, all four architectures achieved comparable performance (82.4%–85.2% accuracy; macro F1 0.826–0.860), with no single architecture dominating across both metrics. A working FAISS-indexed retrieval pipeline was validated with representative query examples, demonstrating reproducible, image-specific nearest-neighbour retrieval.

This investigation's main finding is that, for static two-dimensional geological texture data, matching preprocessing to the native dimensionality of the input contributes more to model performance than the choice between simple, sequential, hybrid, or custom architectures. Earlier apparent performance gaps between LSTM-based and convolutional architectures narrowed substantially once a preprocessing inconsistency was corrected, indicating that some of the presumed unsuitability of sequential architectures for static imagery may reflect implementation choices rather than an inherent architectural limitation. This nuances the prevailing assumption that specialized or hybrid architectural configurations are necessary for strong performance on domain-specific image retrieval tasks.

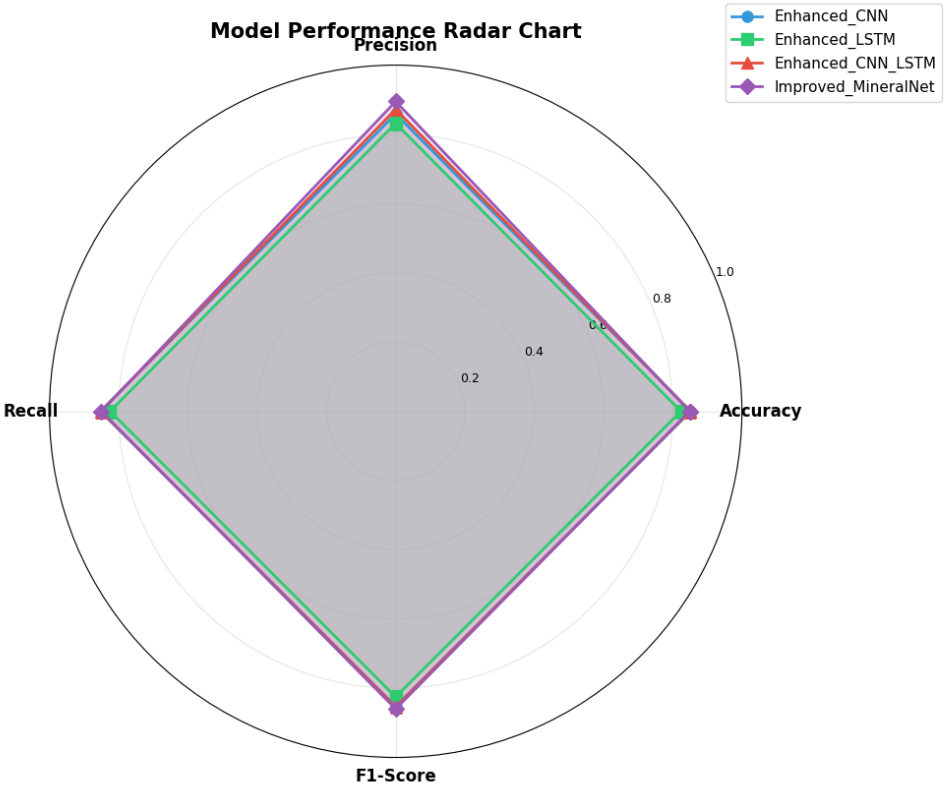

Figure 7. All Model performance radar

This work makes a valuable contribution to the practical implementation of efficient and reproducible image retrieval methodologies within the field of geological informatics. Furthermore, this study provides a rigorously validated, openly accessible benchmark that serves as a foundation for future innovation in mineralogical image analysis applications.

The training curve for Enhanced_CNN shows that training accuracy continues to rise to ~99%, while validation accuracy plateaus at ~84–85% from the early epochs, and validation loss begins to rise again after epoch ~5 while training loss continues to decline—a classic pattern of overfitting. Although Enhanced_CNN achieved the highest accuracy among the four models (85.19%), this pattern suggests that its generalization may be less stable compared to models with a smaller train-validation gap.

REFERENCES

[1] D.-H. Nguyen, H.-P. Huynh, M.-T. Tran, and T.-N. Le, "GenFlow: Interactive Modular System for Image Generation," pp. 1–7, Jan. 2026, doi: 10.1109/CBMI66578.2025.11339318.

[2] G. Gautam and A. Khanna, "Content Based Image Retrieval System Using CNN based Deep Learning Models," *Procedia Comput. Sci.*, vol. 235, pp. 3131–3141, 2024, doi: 10.1016/J.PROCS.2024.04.296.

[3] U. Czerwinska, C. Bircanoglu, and J. Chamoux, "Benchmarking Image Embeddings for E-Commerce: Evaluating Off-the Shelf Foundation Models, Fine-Tuning Strategies and Practical Trade-offs," Apr. 2025, Accessed: May 24, 2026. [Online]. Available: https://arxiv.org/pdf/2504.07567

[4] M. Tschannen *et al.*, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," Feb. 2025, Accessed: May 24, 2026. [Online]. Available: https://arxiv.org/pdf/2502.14786

[5] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *Transactions on Machine Learning Research*, vol. 2024, Apr. 2023, Accessed: May 24, 2026. [Online]. Available: https://arxiv.org/pdf/2304.07193

[6] C. He and H. Wei, "MambaHash: Visual State Space Deep Hashing Model for Large-Scale Image Retrieval," vol. 1, Jun. 2025, doi: 10.1145/nnnnnnn.nnnnnnn.

[7] X. Wang, X. Yuan, Z. Zhang, G. Lu, and X. Li, "CgAT: Center-Guided Adversarial Training for Deep Hashing-Based Retrieval," Apr. 2022, doi: 10.1145/3543507.3583369.

[8] N. Arora, I. Budhiraja, D. Garg, S. Garg, B. J. Choi, and M. S. Hossain, "Revolutionizing facial image retrieval: Multi-block and mean based local binary patterns with sign and magnitude analysis," *Alexandria Engineering Journal*, vol. 116, pp. 601–608, Mar. 2025, doi: 10.1016/J.AEJ.2024.12.003.

[9] G. Ravi, "Content-Based Image Retrieval Using Deep Feature Extraction with ResNet-50," *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 09, no. 05, pp. 1–9, May 2025, doi: 10.55041/IJSREM48676.

[10] S. Park *et al.*, "Influence of content-based image retrieval on the accuracy and inter-reader agreement of usual interstitial pneumonia CT pattern classification," *European Radiology 2025 35:11*, vol. 35, no. 11, pp. 7199–7212, May 2025, doi: 10.1007/S00330-025-11689-9.

[11] P. Tschandl *et al.*, "Human–computer collaboration for skin cancer recognition," *Nature Medicine 2020 26:8*, vol. 26, no. 8, pp. 1229–1234, Jun. 2020, doi: 10.1038/s41591-020-0942-0.

[12] J. Liu, W. Du, C. Zhou, and Z. Qin, "Rock Image Intelligent Classification and Recognition Based on Resnet-50 Model," *J. Phys. Conf. Ser.*, vol. 2076, no. 1, p. 012011, Nov. 2021, doi: 10.1088/1742-6596/2076/1/012011.

[13] B. Zhang, G. Zou, D. Qin, Q. Ni, H. Mao, and M. Li, "RCL-Learning: ResNet and convolutional long short-term memory-based spatiotemporal air pollutant concentration prediction model," *Expert Syst. Appl.*, vol. 207, Nov. 2022, doi: 10.1016/j.eswa.2022.118017.

[14] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data 2021 8:1*, vol. 8, no. 1, pp. 53-, Mar. 2021, doi: 10.1186/S40537-021-00444-8.

[15] S. Tyagi, P. Shukla, P. Singh, and H. Tekchandani, "Enhanced content-based image retrieval via hybrid color, texture, and deep learning features," *Scientific Reports 2026 16:1*, vol. 16, no. 1, pp. 14888-, Mar. 2026, doi: 10.1038/s41598-026-38422-w.

[16] J. Pradhan, A. Ajad, A. K. Pal, and H. Banka, "Multi-level colored directional motif histograms for content-based image retrieval," *The Visual Computer 2019 36:9*, vol. 36, no. 9, pp. 1847–1868, Dec. 2019, doi: 10.1007/S00371-019-01773-9.

[17] X. Qin, L. Li, F. Hao, M. Ge, and G. Pang, "Multi-level knowledge-driven feature representation and triplet loss optimization network for image–text retrieval," *Inf. Process. Manag.*, vol. 61, no. 1, p. 103575, Jan. 2024, doi: 10.1016/J.IPM.2023.103575.

[18] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Comput. Intell. Neurosci.*, vol. 2018, p. 7068349, 2018, doi: 10.1155/2018/7068349.

[19] Z. Xu, W. Ma, P. Lin, H. Shi, D. Pan, and T. Liu, "Deep learning of rock images for intelligent lithology identification," *Comput. Geosci.*, vol. 154, p. 104799, Sep. 2021, doi: 10.1016/J.CAGEO.2021.104799.

[20] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yuy, "Deep visual-semantic hashing for cross-modal retrieval," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 1445–1454, Aug. 2016, doi: 10.1145/2939672.2939812.

[21] S. Dey, A. Dutta, S. K. Ghosh, E. Valveny, J. Llados, and U. Pal, "Learning Cross-Modal Deep Embeddings for Multi-Object Image Retrieval using Text and Sketch," *Proceedings - International*

Conference on Pattern Recognition, vol. 2018-August, pp. 916–921, Apr. 2018, doi: 10.1109/ICPR.2018.8545452.
[22] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," *CIVR 2009 - Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 368–375, Jul. 2009, doi: 10.1145/1646396.1646452;CSUBTYPE:STRING:CONFERENCE.
[23] G. Lagani, D. Bacciu, C. Gallicchio, F. Falchi, C. Gennaro, and G. Amato, "Deep Features for CBIR with Scarce Data using Hebbian Learning," May 2022, Accessed: Jun. 29, 2026. [Online]. Available: http://arxiv.org/abs/2205.08935
[24] M. Alrahhal and V. K. Shukla, "Enhancing Image Retrieval Systems: A Comprehensive Review of Machine Learning Integration In CBIR," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 4195–4214, Jun. 2024, Accessed: Jul. 07, 2026. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/7015
[25] A. Christe *et al.*, "Computer-Aided Diagnosis of Pulmonary Fibrosis Using Deep Learning and CT Images," *Invest. Radiol.*, vol. 54, no. 10, pp. 627–632, Oct. 2019, doi: 10.1097/RLI.0000000000000574.
[26] H. Deng, Y. Zheng, J. Chen, S. Yu, K. Xiao, and X. Mao, "Learning 3D mineral prospectivity from 3D geological models using convolutional neural networks: Application to a structure-controlled hydrothermal gold deposit," *Comput. Geosci.*, vol. 161, p. 105074, Apr. 2022, doi: 10.1016/J.CAGEO.2022.105074.
[27] S. Gupta, V. Moutinho, J. R. A. Godinho, B. M. Guy, and J. Gutzmer, "3D mineral quantification of particulate materials with rare earth mineral inclusions: Achieving sub-voxel resolution by considering the partial volume and blurring effect," *Tomography of Materials and Structures*, vol. 7, p. 100050, Mar. 2025, doi: 10.1016/J.TMATER.2025.100050.
[28] M. Douze *et al.*, "The Faiss library," Jan. 2024, Accessed: Jul. 12, 2025. [Online]. Available: https://arxiv.org/pdf/2401.08281
[29] M. S. Rahman, S. M. E. Rabbi, and M. M. Rashid, "Optimizing Domain-Specific Image Retrieval: A Benchmark of FAISS and Annoy with Fine-Tuned Features," *2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing, iSAI-NLP 2025*, 2025, doi: 10.1109/ISAI-NLP66160.2025.11370303.
[30] K. Bayoudh, F. Hamdaoui, and A. Mtibaa, "An Attention-based Hybrid 2D/3D CNN-LSTM for Human Action Recognition," *Proceedings of 2022 2nd International Conference on Computing and Information Technology, ICCIT 2022*, pp. 97–103, 2022, doi: 10.1109/ICCIT52419.2022.9711631.
[31] Y. J. Chen, H. Huang, and Z. S. Chen, "AVOA-optimized CNN-BILSTM-SENet framework for hydrodynamic performance prediction of bionic pectoral fins," *Ocean Engineering*, vol. 327, p. 121002, May 2025, doi: 10.1016/J.OCEANENG.2025.121002.
[32] A. R. Sahoo and P. Chakraborty, "Hybrid CNN Bi-LSTM neural network for Hyperspectral image classification," Feb. 2024, Accessed: May 15, 2025. [Online]. Available: https://arxiv.org/pdf/2402.10026
[33] X. Zhang, C. Bai, and K. Kpalma, "OMCBIR: Offline mobile content-based image retrieval with lightweight CNN optimization," *Displays*, vol. 76, p. 102355, Jan. 2023, doi: 10.1016/J.DISPLA.2022.102355.
[34] Z. Zhuo and Z. Zhou, "Cross-domain remote sensing image retrieval with Gabor-based CNN," *Int. J. Remote Sens.*, vol. 44, no. 2, pp. 567–584, Jan. 2023, doi: 10.1080/01431161.2023.2168136.
[35] S. Tsokov, M. Lazarova, and A. Aleksieva-Petrova, "A Hybrid Spatiotemporal Deep Model Based on CNN and LSTM for Air Pollution Prediction," *Sustainability (Switzerland)*, vol. 14, no. 9, May 2022, doi: 10.3390/SU14095104.

[36] H. Yu, H. Lu, M. Zhao, Z. Li, and G. Gu, "Gradient aggregation based fine-grained image retrieval: A unified viewpoint for CNN and Transformer," *Pattern Recognit.*, vol. 149, p. 110248, May 2024, doi: 10.1016/J.PATCOG.2023.110248.

[37] M. Palla and R. Karra, "A Comparative Study of Pre-Trained CNN Models with Transfer Learning for Content-Based Image Retrieval," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25820–25826, Aug. 2025, doi: 10.48084/ETASR.11496.

[38] A. Naeem, T. Anees, K. T. Ahmed, R. A. Naqvi, S. Ahmad, and T. Whangbo, "Deep learned vectors' formation using auto-correlation, scaling, and derivations with CNN for complex and huge image retrieval," *Complex & Intelligent Systems 2022 9:2*, vol. 9, no. 2, pp. 1729–1751, Oct. 2022, doi: 10.1007/S40747-022-00866-8.

[39] Z. Wang and R. Zuo, "Mineral prospectivity mapping using a joint singularity-based weighting method and long short-term memory network," *Comput. Geosci.*, vol. 158, p. 104974, Jan. 2022, doi: 10.1016/J.CAGEO.2021.104974.

[40] K. Ma *et al.*, "Multi-granularity retrieval of mineral resource geological reports based on multi-feature association," *Ore Geol. Rev.*, vol. 165, p. 105889, Feb. 2024, doi: 10.1016/J.OREGEOREV.2024.105889.