Automating Rock Classification: A Vision Transformer Approach in Brazil's Ornamental Stone

Douglas Fiório Dias[1], Karin Satie Komati[1], Kelly Assis de Souza Gazolli[1]

[1]*Graduate program in Applied Computing (PPComp)*
Instituto Federal do Espírito Santo, Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, 29166-630, Serra-ES, Brazil
dfdiasbr@gmail.com, {kkomati, kasouza}@ifes.edu.br

Abstract. The ornamental stone industry in Brazil is distinguished by its vast variety of rock types, presenting a unique challenge for classification due to its inherent subjectivity and reliance on expert judgment. To address this issue, this study introduces a publicly accessible database, Ornamental Rocks dataset, comprising 12 distinct classes and 1,794 images, and evaluates ten image classification models for ornamental stones: AlexNet, VGG16, VGG19, DenseNet121, DenseNet169, ResNet50, ResNet101, Xception, Inception V3, and Vision Transformer (ViT) networks. Additionally, these models are tested on two other ornamental stone databases: the Rock Image Dataset, which contains 711 rock images divided into 9 classes, and the Ornamental Stone Slab dataset, with 34,630 images divided into 45 classes. Finally, the models are trained on a unified database encompassing all 66 classes of ornamental stones. Our empirical analysis indicates that the ViT model outperforms traditional architectures on the database created for this study, achieving an accuracy rate of 98.36%.

Keywords: Neural networks, Image classification, Rock Image Dataset, Ornamental stone slab

1 Introduction

When used for coverings and decorative purposes, natural rocks are referred to as ornamental rocks. Brazil holds a pivotal role within the ornamental stone industry. According to a 2022 report by the Brazilian Association of the Ornamental Rocks Industry, Brazil produced 10 million tons of ornamental rocks, with granite, marble, and quartzite being the predominant types [1]. The production data for 2022 reveals that granites and similar stones accounted for 40% of the total output.

Despite its significance to the national economy, the classification of ornamental rocks within Brazil relies heavily on subjective assessments, necessitating expertise from specialists in the field [2]. This subjectivity leads to discrepancies between buyers and sellers regarding the quality and categorization of materials. Consequently, there is a compelling case for exploring automated approaches to the classification of ornamental rock imagery. Such technological advancements could significantly reduce subjectivity in the classification process and enhance the comprehension of all stakeholders involved [3].

Since the early 1990s, considerable research has focused on developing methodologies for the automatic classification of rock imagery [2, 4–9]. Convolutional Neural Networks (CNNs) have been widely used in ornamental rock image classification. Recently, Vision Transformers (ViTs) have shown superior performance compared to CNNs in various image classification tasks [10]. This work proposes adopting the Vision Transformer (ViT) for this task, exploring its potential to enhance current methods.

This study outlines a structured approach to addressing the challenges of ornamental rock classification through the following strategies:

- The study will employ the Vision Transformer (ViT) neural network model for the classification of ornamental rocks using images from the developed database and the images from databases selected for this work. The performance of the ViT model will be evaluated in comparison with other neural network architectures, including AlexNet, VGG16, VGG19, Densenet121, Densenet169, ResNet50, ResNet101, Xception, and Inception V3.
- We present the Ornamental Rocks dataset, a comprehensive database consisting of images of three primary types of ornamental rocks: granite, marble, and quartzite. This database contains a total of 1,794 images, covering 12 rock classes. It is composed of photographs sourced directly from entities in the ornamental

stone sector, with the collection process managed by the researchers. Furthermore, the database is made publicly available.

- Evaluation on existing databases: Two existing databases were selected for this work due to their labeled images of ornamental stones: (i) 'Rock Image Datasets', with 711 images across 9 classes, and (ii) 'Ornamental Stone Slabs', containing 34,630 images divided into 45 classes. Additionally, a unified database composed of these three databases will be created. The total number of images among the classes in this database varies greatly, leading to an imbalance in the class group, which presents a challenge for machine learning models [11].

The article is structured as follows: Section 2 presents related works, Section 3 introduces the databases and the methods for image classification, and Section 4 shows the experimental results. Finally, Section 5 presents the main conclusions of the paper.

2 Related Works

In 1995, researchers were exploring the use of machine learning to classify ornamental rocks. Hernandez et al. [4] conducted a comparative study between traditional clustering and classification algorithms and a Multi-layer Perceptron (MLP) neural network employing Backpropagation (BP). The study evaluated four algorithms: the a priori Euclidean Classifier, the Euclidean Classifier by supervised learning, the Bayesian a priori Classifier, and the Statistical Classifier by supervised learning. It specifically focused on the image classification of "Sierra de la Puerta" marble, analyzing the colors within each image using the RGB color system. The findings indicated that the neural network model achieved a classification closely aligned with the assessments of industry experts, albeit requiring a longer learning time.

Ferreira et al. [3] employed CNNs for the classification of granite slab images. Their approach involved segmenting the original images into smaller sections. Following the neural network's analysis and classification of these segments, a majority voting mechanism was employed to classify the original image. This method proved effective in classifying granite slabs across various resolutions and sizes. The study trained four neural networks: three inspired by the digit recognition challenge using the MNIST database, and one based on the CIFAR image recognition challenge. In experiments with 32x32 pixel images, without employing the majority voting strategy, the CIFAR-based network achieved an accuracy rate of 87.26%. This was significantly higher than other networks that relied solely on high-resolution image feature extraction, which recorded accuracies around 33%.

Pascual et al. [8] employed CNNs for image classification using the Rock Image Datasets. The study initially focused on applying a CNN to the dataset to evaluate its performance metrics. Utilizing a 3-layer CNN resulted in an impressive accuracy of 99.60%. Subsequently, the same network was applied to images captured by robots during rock exploration activities in the field, where the images were not acquired under controlled conditions. This new task was simplified into a binary classification problem where the images were classified into breccia and non-breccia. For field-acquired images, a modified CNN with 5 layers demonstrated an accuracy of 89.43%.

In 2021, Ouzounis et al. compared 15 convolutional neural networks on a dataset containing 489 marble images sourced from a production line in Northeast Greece. The networks analyzed included DenseNet, Inception, and ResNet (residual networks), among others. The study aimed to compare the networks and interpret the heat maps generated by Gradient-weighted Class Activation Mapping (Grad-CAM). The results indicated that CNNs achieved superior outcomes compared to models such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and MLP, particularly in classifying marble images based on texture. DenseNet201 outperformed the other networks analyzed, achieving an accuracy rate of 83.24%.

The work of Xu et al. [12] conducts classification using microscopic images of rock. Seven different convolutional neural networks were compared: Xception, MobileNet_v2, Inception_ResNet_v2, Inception_v3, DenseNet121, ResNet101_v2, and ResNet-101, utilizing the technique of transfer learning. The images were randomly selected with a ratio of 9:1 between the training dataset and the testing dataset, with the training dataset comprising 13,463 images and the testing dataset comprising 1,487 images. The results demonstrated that the Xception-based model achieved the highest performance, with an accuracy of 97.66% on the training dataset and 98.65% on the test dataset.

The authors, Yimeng Zhou, Louis Ngai Yuen Wong, and Keith Ki Chun Tse, developed a new CNN called HKUDES_Net [13], specifically designed to classify seven common types of rocks in Hong Kong, including fine-grained, medium-grained, and coarse-grained granites, fine and coarse ash tuffs, feldspar rhyolite, and granodiorite. The implementation of the "alerting level" in HKUDES_Net was crucial in eliminating overfitting, significantly improving the network's performance in rock image classification. The best results achieved by HKUDES_Net include an accuracy of 96.5%, a recall of 95.7%, and an f1-score of 96.1%. The study highlights the effectiveness of HKUDES_Net in handling similar textures and different grain sizes, outperforming ten benchmark CNNs and seven feature-based algorithms in terms of accuracy, recall, and f1-score.

3 Materials and Methods

3.1 Materials

Database 1 - Ornamental Rocks Dataset[1]: The database developed in this study was compiled with the assistance of Angramar Granitos e Mármores, a company based in Cachoeiro de Itapemirim-E.S., Brazil. Through field visits, 1,794 images were collected, covering 12 classes of rocks: three varieties of granite, three of marble, and six types of quartzite. Figure 1 shows the 12 classes in the database. The images were captured using the following equipment: Canon PowerShot SX40HS camera (12.1 MP); Motorola G20 smartphone (48 MP) and Samsung Galaxy S20 FE smartphone (12 MP).

Figure 1. Rock Classes in the Ornamental Rocks dataset. From the top to bottom, left to right: granite-blackswan, granite-lucyinthesky, granite-nevascawhite, marble-dolomite-brancoparana, marble-dolomite-calacata, marble-shadow, quartzite-biancosuperiore, quartzite-oceanblue, quartzite-patagonia, quartzite-silvermoon, quartzite-tajmahal and quartzite-volupia.

#	Class	Total
0	granite-blackswan	145
1	granite-lucyinthesky	161
2	granite-nevascawhite	161
3	marble-dolomite-brancoparana	152
4	marble-dolomite-calacata	124
5	marble-shadow	155
6	quartzite-biancosuperiore	126
7	quartzite-oceanblue	159
8	quartzite-patagonia	164
9	quartzite-silvermoon	158
10	quartzite-tajmahal	127
11	quartzite-volupia	162
	TOTAL	**1794**

#	Class	Total
0	andesite	87
1	dolostone	73
2	granite	86
3	limestone	82
4	oolitic_limestone	83
5	peridotite	83
6	red_granite	70
7	rhyolite	78
8	volcanic_breccia	69
	TOTAL	**711**

Figure 2. Classes and their quantities in the Ornamental Rocks dataset and Rock Image Datasets

Database 2 - Rock Image Datasets [6, 8]: This database was created by Shu et al. [6] and is publicly available[2]. It contains 711 rock images, divided into 9 classes. In the Figure 2, the table on the right shows the classes and the number of images in each one.

Database 3 - Ornamental Stone Slabs[3]: This database contains 34,630 images, divided into 45 classes. The database was created by João Victor Costa Araujo and provided by Cajugram Granitos e Mármores do Brasil Ltda. Although Database 3 has a large number of images, it has a considerable imbalance between the classes, unlike the previous two databases, which have a relatively similar number of images across the classes. Figure 3 shows the distribution of images among the classes.

One of the challenges in correctly classifying images of ornamental stones lies in the similarity between some types of rocks. Figure 4 shows two classes of rocks obtained from the listed databases. It is possible to notice the great similarity between these classes, which can pose a difficulty for the models in correctly classifying the images.

[1] https://www.kaggle.com/datasets/douglasfiriodias/ornamental-rocks-dataset
[2] https://data.mendeley.com/datasets/7g7zpy9vcb/1
[3] https://www.kaggle.com/datasets/joovictorcostaaraujo/chapas-polidas-de-rochas-ornametais

#	Class	Total
0	giallo_fiorito	181
1	giallo_maracana	339
2	golden_storm	1185
3	icarai_yellow	292
4	ice_leke	113
5	ipanema_beige	2894
6	itaunas_white	1546
7	kalahari	665
8	maracuja_yellow	174
9	naica	727
10	nevada_black	3806
11	new_caledonia	556
12	olympios	209
13	ornamental	251
14	perla_venato	508
15	quartzito_green_da_vinci	174
16	quartzito_thannos	130
17	quartzito_venom	434
18	quartzito_verde_sauipe	106
19	rocky_mountain	191
20	santa_cecilia	1446
21	santa_cecilia_light	390
22	san_francisco_green	1404

#	Class	Total
23	sao_gabriel_black	610
24	shadow_white	1922
25	siena_white	4588
26	solarius	470
27	splendor_gold	171
28	tabaco_red	346
29	ubatuba_green	2965
30	vitoria_white	619
31	white_bellukha	122
32	white_ceara	453
33	white_cintilante	183
34	white_everest	295
35	white_extreme	388
36	white_himalaya	360
37	white_mirage	1219
38	white_olympus	1153
39	white_samoa	106
40	white_sea	108
41	white_serenata	109
42	white_superiore	246
43	whte_liberdade	196
44	xango_red	280
	Total	**34630**

Figure 3. Classes and their quantities in the Ornamental Stone Slabs

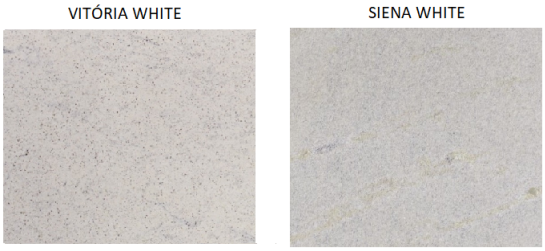

Figure 4. Similar Classes in the databases

3.2 Methods

In this study, ten neural network models were used: the Vision Transformer [10], which utilizes a Transformer-based architecture; and CNNs including AlexNet [14], Inceptionv3 [15], Xception [16], ResNet50 and ResNet101 [17], DenseNet121 and DenseNet169 [18], along with VGG16 and VGG19 [19]. CNNs are designed to learn spatial hierarchies of features through backpropagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers. The convolution layers apply a set of learnable filters to the input image, enabling the network to detect various features such as edges, textures, and shapes. Pooling layers reduce the dimensionality of the data, making the computation more efficient while retaining important information. Finally, the fully connected layers interpret the extracted features to make a final classification or prediction. The hierarchical structure of CNNs allows them to capture complex patterns and structures in the data, making them particularly powerful for image classification, object detection, and other tasks involving high-dimensional data.

The differences between various CNN architectures mainly involve the number of layers, the types of layers, the sequence of these layers, and the specific parameters used within them. Some architectures, like VGG, utilize a straightforward approach with a deep stack of convolutional layers of the same size, while others, like ResNet, introduce shortcut connections to allow gradients to flow more easily during training. Inception networks employ a more complex structure that combines multiple convolutional operations with different kernel sizes at each layer, enabling the network to capture information at various scales. DenseNet connects each layer to every other layer in a feed-forward fashion to improve information flow and gradient propagation. The Inception architecture captures features at multiple scales within each module using various convolutional filters, while Xception replaces Inception modules with depthwise separable convolutions for greater efficiency and performance.

The Transformer network is a learning architecture designed for sequential data that does not require processing in a specific order, initially applied to the field of natural language processing [20]. It comprises an encoder-decoder structure for handling sequences of elements, fundamentally relying on the concept of self-attention mechanisms. The intuition behind self-attention is that not all words in a sentence are given equal importance. Hence, self-attention is a mechanism that establishes relationships between words within the same sentence, aiming to fo-

cus on some words while others receive minimal attention. This mechanism can aid in the disambiguation process for words that have the same spelling but different meanings.

The Transformer architecture has been adapted for image recognition tasks [10]. In the architecture for a ViT network, initially, the input image is divided into smaller fixed-size patches, which are then embedded into a vector that feeds into a transformer encoder. This encoder consists of alternating layers of Multi-Head Self-Attention (MSA) and MLP blocks. Similar to the Transformer network, the MSA plays a crucial role in the process of analyzing images within ViT network. The images, segmented into smaller fixed-size patches, function analogously to words (or tokens) in a Transformer network. Linearizing the image patches treats each as an individual input unit. Consequently, the self-attention mechanism is applied to these image parts, enabling the network to learn about the relationships between them.

The core concept is to view images as a sequence of patches (rectangular segments of the original image) and then employ the architecture to generate a compact representation of the image. This process unfolds in two stages: the first involves extracting features from the image patches, and the second aggregates these features into a global representation of the image. In the initial stage, the Transformer architecture extracts features from the image patches in a sequence of vectors. Each patch is regarded as a word in the sequence, and the architecture generates a contextualized representation of each, considering the others. This implies that the attention-equipped architecture can capture contextual information from various parts, thereby producing a richer and more informative representation of each segment. In the subsequent stage, the features derived from the patches are amalgamated into a singular global representation of the image. The outcome of this aggregation is a linear layer, where each subgroup represents a 1x1 matrix. This global representation then serves as input for a classifier layer, which assigns a label to the image.

4 Experiments and Results

For training and testing the models, the three databases were used separately, and a unified database was created by joining the three databases, totaling 37,135 images and 66 classes of ornamental rocks, divided in the proportion 70/15/15 for training, validation, and testing. Since the 10 models used in this study have a maximum image input size of 384x384, an initial image pre-processing phase was carried out to obtain images of size 400x400 in all 66 classes, which contributed to a shorter processing time for the models.

All models were trained with the same parameters, including the learning rate and number of epochs. The AdamW [21] optimizer was employed for its effectiveness, complemented by the OneCycleLR [22] policy to dynamically adjust the learning rate based on training progression. Figure 5 shows the Confusion Matrix for the ViT model, considering Database 1. Figure 6 shows the results for the models, for each database.

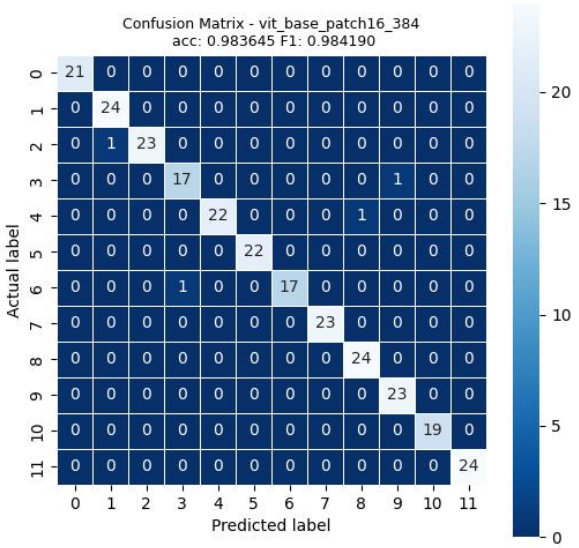

Figure 5. Confusion Matrix for the ViT model in the Database 1.

Database 1 - 12 classes			Database 2 - 9 classes			Database 3 - 45 classes			Unified Database - 66 classes		
Modelo	**Acc (%)**	**F1 (%)**	**Modelo**	**Acc (%)**	**F1 (%)**	**Modelo**	**Acc (%)**	**F1 (%)**	**Modelo**	**Acc (%)**	**F1 (%)**
ViT	**98.36%**	**98.42%**	**ViT**	**98.30%**	**98.24%**	ResNet50	94.02%	93.86%	**ViT**	**92.99%**	**92.82%**
Resnet50	92.61%	92.78%	Vgg19	94.52%	93.95%	ResNet101	93.89%	93.80%	ResNet101	92.81%	92.82%
Resnet101	90.62%	90.82%	Resnet101	94.94%	93.49%	**ViT**	**92.00%**	**92.05%**	ResNet50	93.02%	92.68%
densenet169	90.08%	90.42%	Vgg16	93.02%	92.01%	xception	91.75%	91.39%	densenet169	92.23%	92.33%
Inception v3	90.34%	90.29%	densenet169	89.39%	88.46%	inception_v3	91.21%	91.35%	inception_v3	90.95%	91.11%
Vgg16	87.09%	87.11%	AlexNet	88.22%	86.96%	densenet169	91.38%	91.10%	vgg16	89.47%	89.42%
AlexNet	86.76%	86.49%	Inception v3	84.58%	82.68%	densenet121	90.90%	90.14%	densenet121	89.31%	89.05%
Vgg19	85.96%	86.11%	Resnet50	83.92%	81.69%	AlexNet	89.86%	89.26%	xception	89.34%	89.04%
xception	82.31%	82.44%	densenet121	72.21%	69.78%	vgg16	88.69%	88.88%	vgg19	86.05%	84.87%
densenet121	75.73%	75.82%	xception	58.27%	57.24%	vgg19	86.91%	86.68%	AlexNet	84.17%	83.70%

Figure 6. Results for accuracy and F1 metrics on test data.

The performance of the Vision Transformer (ViT) network is notable when compared to convolutional networks. In the Database 1, the ViT network achieved a substantial accuracy rate of 98.36% and F1-measure of 98.42%, significantly outperforming the other networks. The second-best accuracy on this database was 92.61%.

In the other databases, the ViT network consistently ranked among the top three models. However, its lowest performance was on Database 3, which includes 45 classes. In the unified database, which combines all three databases and includes 66 classes, the ViT achieved an accuracy of 92.99%, very close to the highest accuracy of 93.02% obtained by the ResNet50 model. Additionally, the F1 score of the ViT on the unified database (92.82%) was slightly higher than that of the ResNet50 network (92.68%).

The ResNet50 and ResNet101 networks also demonstrated strong performance across the databases. These models, along with the ViT network, were almost always among the top three networks. Notably, in Database 2, the ResNet50 network had its worst performance, with an accuracy of 83.92%.

It is important to note the high imbalance present in Database 3. Barulina et al. [11] presented a study on the impacts of neural networks trained on unbalanced databases, highlighting issues such as the tendency of networks to incorrectly predict images from minority classes and potential overfitting. In our work, the models were trained on the database as is, without trying to balance the classes.

5 Conclusion

The present study sought to analyze the performance of a Vision Transformer (ViT) network in classifying images of ornamental rocks. Using a database created for this study and two other databases, the ViT was trained and its performance was compared to the performance of 9 other neural network models. The ViT outperformed the other models in the database with 12 classes created for this study, and was among the top three models in all studies in the other databases. This shows the robustness and the potential of using a model with transformer architecture for image classification.

As future research proposals, the performance of the Vision Transformer can be compared to models such as ConViT (Convolutional Vision Transformer) [23], which combines characteristics of convolutional networks with the Transformer architecture. It is also planned to develop a mobile version.

Acknowledgements. The authors would like to thank FAPES/UnAC (No. FAPES 1228/2022 P 2022-CD0RQ, No. SIAFEM 2022-CD0RQ) for the financial support provided through the Sistema UniversidaES. Professor Komati thanks CNPq for the DT-2 grant (nº 302726/2023-3) and project nº407742/2022-0, also thanks FAPES for project nº 1023/2022 P:2022-8TZV6.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] ABIROCHAS. Balance of exports and imports of ornamental rocks in 2022., 2023.

[2] J. Martinez-Alajarin, J. Luis-Delgado, and L. Tomas-Balibrea. Automatic system for quality-based classification of marble textures. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, n. 4, pp. 488–497, 2005.
[3] A. Ferreira and G. Giraldi. Convolutional neural network approaches to granite tiles classification. *Expert Systems with Applications*, vol. 84, pp. 1–11, 2017.
[4] V. Hernandez, P. C. Perez, L. G. Perez, L. T. Balibrea, and H. P. Pina. Traditional and neural networks algorithms: applications to the inspection of marble slab. In *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, volume 5, pp. 3960–3965. IEEE, 1995.
[5] M. López, J. Martínez, J. Matías, J. Taboada, and J. Vilán. Functional classification of ornamental stone using machine learning techniques. *Journal of Computational and Applied Mathematics*, vol. 234, n. 4, pp. 1338–1345. Proceedings of the Thirteenth International Congress on Computational and Applied Mathematics (ICCAM-2008), Ghent, Belgium, 7–11 July, 2008, 2010.
[6] L. Shu, K. McIsaac, G. R. Osinski, and R. Francis. Unsupervised feature learning for autonomous rock image classification. *Computers & Geosciences*, vol. 106, pp. 10–17, 2017.
[7] X. Ran, L. Xue, Y. Zhang, Z. Liu, X. Sang, and J. He. Rock classification from field image patches analyzed using a deep convolutional neural network. *Mathematics*, vol. 7, n. 8, 2019.
[8] A. D. P. Pascual, L. Shu, J. Szoke-Sieswerda, K. McIsaac, and G. Osinski. Towards natural scene rock image classification with convolutional neural networks. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pp. 1–4. IEEE, 2019.
[9] X. Liu, H. Wang, H. Jing, A. Shao, and L. Wang. Research on intelligent identification of rock types based on faster R-CNN method. *Ieee Access*, vol. 8, pp. 21804–21812, 2020.
[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
[11] M. Barulina, S. Okunkov, I. Ulitin, and A. Sanbaev. Sensitivity of modern deep learning neural networks to unbalanced datasets in multiclass classification problems. *Applied Sciences*, vol. 13, n. 15, 2023.
[12] Z. Xu, W. Ma, P. Lin, and Y. Hua. Deep learning of rock microscopic images for intelligent lithology identification: Neural network comparison and selection. *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 14, n. 4, pp. 1140–1152, 2022.
[13] Y. Zhou, L. N. Y. Wong, and K. K. C. Tse. Novel rock image classification: the proposal and implementation of hkudes_net. *Rock Mechanics and Rock Engineering*, vol. 56, n. 5, pp. 3825–3841, 2023.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, vol. 25, 2012.
[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
[16] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.
[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds, *Advances in neural information processing systems*, volume 30. Curran Associates, Inc., 2017.
[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
[22] L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. arxiv. *arXiv preprint arXiv:1708.07120*, vol. 6, 2017.
[23] S. D'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In M. Meila and T. Zhang, eds, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2286–2296. PMLR, 2021.