Self-supervised Transfer Learning for Medical Image Classification

Lorenzo Baietti
lorenzo.baietti@studenti.unipd.it

Francesco Carlesso
francesco.carlesso.1@studenti.unipd.it

Abstract

Self-supervised learning is a central area of research in computer vision. In this work, we use a pre-trained model that represents one of its main approaches, Predictive Learning. We focus on the respective state-of-the-art BEiT architecture, and evaluate the effectiveness of the learned representations on a medical image classification task, against some benchmarks. Our results demonstrate the potential of this paradigm for learning meaningful and generalizable features, highlighting its relevance when using transfer learning on a domain-shifted downstream task.

1. Introduction

With the increasing availability of large-scale image datasets, computer vision models have made significant progress in supervised learning tasks. However, the dependence on labeled datasets presents a major limitation in many practical applications, considering the cosmic amount of visual data available and the fact that manual annotation can be very costly, especially when domain experts are required. To mitigate this issue, unsupervised learning methods, which learn useful visual representations without labeled data, have gained considerable attention in recent years. One promising approach in this domain is Self-Supervised Learning (SSL), where models are trained to solve *pretext tasks* that do not require labeled data, but are still able to extract useful representations that can be used in various downstream tasks such as image classification, segmentation, and object detection. Among the different techniques in SSL we can distinguish one of the main approaches, which is Predictive Learning. This approach is often associated to "context prediction", because it learns representations by predicting spatial relationships between patches extracted from an image.

Our objective is to show the effectiveness of SSL for performing tasks in datasets unrelated to the one on which the model was originally trained on, showing how such approach compares to supervised benchmarks even when using transfer learning from a completely different domain. To do this we use the BEiT model (Bidirectional Encoder representation from Image Transformers), which embodies the predictive learning principles and is pre-trained on ImageNet, and fine-tune it on some medical datasets from the MedMNIST library, examining its efficacy under varying input conditions. We compare the model against supervised benchmarks trained directly on the medical datasets, demonstrating outstanding performance. The code is available at the following GitHub repository.

2. Related Work

2.1. Predictive Learning

Predictive Learning is one of the foundational approaches in SSL, where models are trained to predict missing or occluded information in input data. One seminal work in this domain is "Unsupervised Learning by Context Prediction" [5], which introduced the idea of predicting spatial relationships between image patches to learn useful representations. This concept was expanded in the paper by Noorozi and Favaro [9], where the pretext task involved rearranging shuffled image patches in their original order, similar to solving jigsaw puzzles. These works demonstrated that solving such spatial prediction tasks allows models to capture semantic information and spatial hierarchies, which are essential for downstream visual understanding tasks. More recently, predictive learning has evolved to take advantage of transformer-based architectures. BEiT [2] adapts the masked token prediction task from natural language processing to computer vision. The model is trained to predict visual tokens using a tokenizer to discretize image data, and it learns to reconstruct masked image patches through a discrete variational autoencoder.

Contrastive Learning is another dominant approach in SSL, where architectures such as SimCLR [4] learn representations by contrasting positive pairs (augmented views of the same image) against negative pairs (different images), thus enforcing similarity within the learned feature space. Unlike contrastive approaches that focus on instance-level discrimination, predictive learning techniques like BEiT (Figure 1) aim to capture contextual and semantic information by reconstructing image parts, making them particularly well-suited for dense prediction tasks.

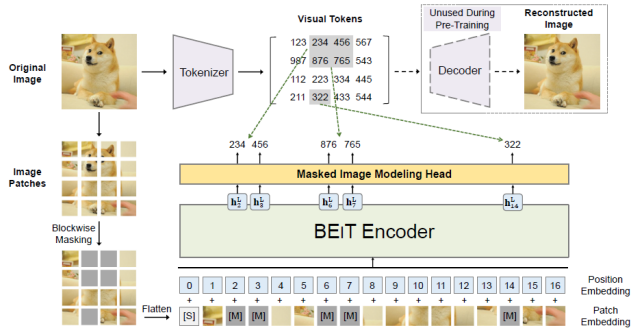

Figure 1: BEiT pre-training

2.2. Self-supervised Transfer Learning

Recent advances in self-supervised transfer learning have demonstrated the efficacy of the relative models in domain adaptation. By leveraging pretext tasks, SSL architectures can learn representations that are both generalizable and robust to domain variations, and studies have shown that self-supervised models pre-trained on large datasets, such as ImageNet, can generalize well to domain-shifted tasks with minimal fine-tuning [6]. This idea has been extensively explored in works such as MoCo [8] and DINO [3], which showcased how self-supervised pre-training can result in superior transfer performance. Similarly, research by Goyal et al. [7] explored the transfer performance of self-supervised models in diverse downstream tasks, providing a comprehensive analysis of their generalization capabilities. Furthermore, SSL approaches have been successfully applied in medical imaging tasks, where labeled data is scarce, achieving performance comparable to supervised models while requiring significantly fewer annotations. For example, Azizi et al.[1] applied contrastive pre-training with domain-relevant augmentations, achieving state-of-the-art results on several benchmarks.

Our work aims at further testing the discoveries outlined above, using a slightly different method to confirm the upsides of taking advantage of this technique.

3. Datasets

The BEiT model we are going to use has been pre-trained on the widely known ImageNet, which does not contain dedicated medical categories, and consists primarily of everyday objects, animals, scenes, and abstract concepts. While some categories might incidentally contain medically related items (such as syringes, stethoscopes, or medical professionals), it lacks specialized medical imaging data.

3.1. MedMNIST

The datasets for the image classification task come from MedMNIST [10], a large-scale MNIST-like collection of standardized biomedical images, including 12 datasets for 2D and 6 datasets for 3D images, where our focus will be on 2D images. All images from the MedMNIST2D collection are pre-processed into 28x28 with the corresponding classification labels, so that no background knowledge is required for users. Recently, the possibility to have also larger sizes of 64x64, 128x128, and 224x224 was introduced. Covering primary data modalities in biomedical images (e.g., X-Ray, Ultrasound, Electron Microscope), MedMNIST is designed to perform classification with various data scales (from 100 to 100,000) and diverse tasks (binary/multi-class, ordinal regression, and multi-label).

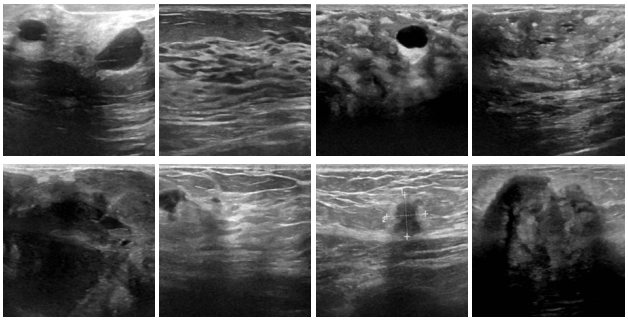

Figure 2: Samples from BreastMNIST, where images on top are classified as normal or benign conditions, while images on the bottom are classified as malign breast cancer.

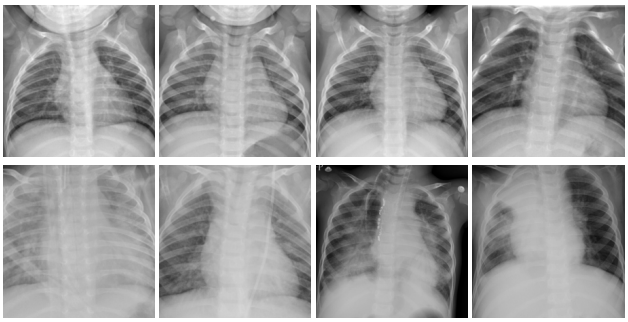

Figure 3: Samples from PneumoniaMNIST, where images on top are classified as a normal condition, while images on the bottom are classified as pneumonia.

For our experiments we selected two specific datasets: PneumoniaMNIST and BreastMNIST, that allow us to explore performance at different data scales, for binary classification. BreastMNIST is based on a dataset of 780 breast ultrasound images categorized into 3 classes: normal, benign, and malignant, where the task is simplified into binary classification by combining normal and benign as positive, classifying them against malignant as negative. The source dataset is split with a ratio of 7:1:2 into training, validation, and test set. PneumoniaMNIST is based on a previous

dataset of 5,856 pediatric chest X-Ray images, where the task is binary classification of pneumonia against normal conditions. The source training set is split with a ratio of 9:1 into training and validation set, and its source validation set is used as the test set. We care to point out that, by looking at the ultrasounds (Figure 2) and the X-rays (Figure 3) one can clearly see how much tricky it can be to perform a diagnosis, reinforcing the previous considerations about annotations. Preprocessing steps consisted in converting the datasets to a format compatible with Hugging Faces's models. Additionally, we performed class balancing to also experiment with a more fair label distribution. Overall, the MedMNIST datasets provide a challenging yet structured benchmark to evaluate the efficacy of self-supervised transfer learning in a domain-shifted task.

4. Method

The idea is to compare the BEiT model, which utilizes predictive learning principles, against the supervised benchmarks provided for the relative datasets. Our methodology involves fine-tuning some of the final layers of the model on the medical imaging datasets, to analyze its classification performance. The key aspect is that we are keeping most of the layers fixed to the parameter values the model learned on ImageNet, therefore applying transfer learning.

Other studies [1] applied contrastive learning in pre-training, we instead focus on predictive learning. Even if in this case pre-training is done on a completely different dateset, not using medical images as inputs, we care to make the point that predictive learning approaches seem to be best suited for medical imaging, considering that augmentations, for example color jittering, could mess with the correct diagnosis and hurt rather than help the model.

4.1. Self-supervised Learning with BEiT

The selected model is BEiT-large, which has 24 transformer layers with hidden size of 1024, and 16 attention heads, for a total of 304M parameters. It benefits from pre-training on 14 million 224x224 images by employing Masked Image Modeling (MIM) with 16x16 image patches. so 196 per image. Specifically, during pre-training a portion of image patches are randomly masked and the model is tasked with reconstructing them based on the surrounding context. This forces the network to develop a deep understanding of spatial relationships and semantic content, learning meaningful representations applicable to downstream tasks. Therefore, given an input image $\mathbf{x}$, it is tokenized into patches $\{x_1, x_2, ..., x_N\}$, where each patch corresponds to a vector embedding. A subset of these patches $\mathcal{M}$ is masked, and the model is trained to predict the corresponding visual tokens $\hat{x}_m$ using:

$$\hat{x}_m = \arg\max_{x \in \mathcal{V}} P(x|\mathbf{x}_{\setminus m}) \tag{1}$$

where $\mathcal{V}$ is the visual vocabulary obtained via a discrete variational autoencoder (dVAE), and $\mathbf{x}_{\setminus m}$ represents the visible patches.

A crucial aspect of BEiT's architecture is the use of self-attention mechanisms within the Vision Transformer (ViT) backbone. Self-attention maps are pivotal in capturing long-range dependencies and contextual relationships between different image regions. They provide insights into which specific parts of the image contribute most to the model's decisions, and given the self-supervised approach, these learned representations are not biased by any labels, making them highly transferable to downstream tasks.

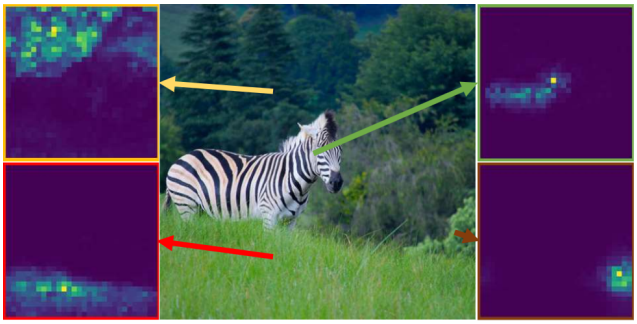

Figure 4: Self-attention maps example

The optimization objective of BEiT is based on the reconstruction of masked patches using dVAE, where each patch corresponds to a visual token, and the model is trained to predict the corresponding masked visual tokens $\hat{x}_m$ using the reconstruction loss:

$$\mathcal{L}_{\text{MIM}} = -\sum_{m \in \mathcal{M}} \log P(\hat{x}_m|\mathbf{x}_{\setminus m}) \tag{2}$$

where $P(\hat{x}_m|\mathbf{x}_{\setminus m})$ is the predicted probability of the correct visual token. The learned representations from BEiT are highly transferable due to their ability to capture hierarchical and contextual features that generalize well, when adapted through fine-tuning, across different domains.

4.2. Fine-tuning

The BEiT model is fine-tuned on the datasets by adjusting the classifier head, the pooler, and the top 5 transformer layers, while keeping all of the others frozen, resulting in approximately 63M trainable parameters, which represent around 20% of the total parameters of the architecture.

This fine-tuning process bridges the gap between the self-supervised pre-training objective and the supervised downstream classification objective, enabling the model to specialize in the new domain while leveraging the general-purpose features learned during pre-training. The transition is therefore from the reconstruction-based objective of MIM to the cross-entropy loss for classification, where this

phase aligns the learned representations to the class labels of the medical datasets. Specifically, in our case, the model is trained to minimize the binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big] \quad (3)$$

where y_i is the ground truth label, $\hat{y}_i$ is the predicted probability for the positive class, and N is the total number of samples. Again, this transition from $\mathcal{L}_{\text{MIM}}$ to $\mathcal{L}_{\text{BCE}}$ is a critical step, as the model must adapt its learned representations to directly predicting class probabilities. The pre-trained features act as a strong foundation, ensuring that the model efficiently learns the decision boundaries in the new domain without overfitting. Eventually, Fine-tuning was performed for 5 epochs with a learning rate of 3×10^{-4}, and with batch size and weight decay tailored for the specific experiment.

4.3. Evaluation Metrics

Performance of the model is assessed using standard classification metrics such as:

- **Area Under the Curve (AUC):** The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds. AUC measures the model's ability to distinguish between the positive and negative classes across various classification thresholds, providing a single scalar value summarizing the curve, which ranges from 0.5 (random guessing) to 1.0 (perfect classification). TPR and FPR are defined as:

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4)$$

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \quad (5)$$

- **Accuracy:** Proportion of correctly classified samples.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad (6)$$

4.4. Benchmarks

To evaluate the effectiveness of the BEiT model, we compare its performance against several supervised learning baselines provided as benchmarks for each MedMNIST dataset. These benchmarks are trained directly on the medical datasets without leveraging pre-training on other large-scale image dataset, and they are:

- **ResNet-18**: Widely used CNN architecture known for its deep residual learning framework, which allows training very deep networks by mitigating the vanishing gradient problem.
- **ResNet-50:** Deeper variant of ResNet with 50 layers, offering greater capacity to learn complex features.
- **auto-sklearn:** AutoML framework which applies Bayesian optimization and meta-learning to select the best pipeline for the given task.
- **AutoKeras:** AutoML framework that automatically searches for optimal neural network architectures using neural-architecture-search techniques.
- **Google AutoML Vision:** AutoML platform that leverages Google's proprietary transfer learning and neural-architecture-search techniques to optimize model performance with minimal user input.

For the two ResNet architectures are provided benchkmarks for both 28x28 and 224x224 images.

4.5. Implementation Details

The experiments are carried out on a local machine equipped with a NVIDIA RTX 2080 Super GPU to ensure efficient computation.

5. Experiments

The two training datasets exhibited a notable class imbalance, prompting us to apply an undersampling technique to mitigate the disproportion between classes. Specifically, for the BreastMNIST dataset, we randomly removed instances from the majority class to achieve a more balanced 60/40 distribution. We deliberately avoided a strict 50/50 split balancing to prevent an excessive reduction in the sample size. In contrast, since PneumoniaMNIST dataset contained a larger number of examples, it allowed us to achieve a 50/50 class balance. These balances were assessed only for our target model, to also investigate the impact that a different label distribution could have on performance.

We also need to mention that since the BEiT model was pre-trained on RGB images of size 224×224, the input for fine-tuning must follow the same format. Hugging Face provides a preprocessor that automatically resizes images to the required dimensions and normalizes them across the RGB channels. In our case, all images are converted from single-channel (grayscale) to three-channel (RGB), and those originally sized 28×28 are upscaled to 224×224. Therefore, we must consider these transformations when analyzing the performance across the two different input resolutions we tested.

5.1. BreastMNIST

- BEiT (28) and BEiT (224) are fine-tuned with batch size of 32 and weight decay of 0.01
- BEiT (28) balanced and BEiT (224) balanced are fine-tuned with batch size of 16 and weight decay of 0.1

Further details are in the Fine-tuning subsection (4.2).

Method	AUC	Accuracy
⋆ BEiT (28)	0.910	**0.891**
† BEiT (28) balanced	0.910	0.872
† BEiT (224)	0.904	0.885
† BEiT (224) balanced	0.908	0.878
ResNet-18 (28)	0.901	0.863
ResNet-18 (224)	0.891	0.833
ResNet-50 (28)	0.857	0.812
ResNet-50 (224)	0.866	0.842
auto-sklearn	0.836	0.803
AutoKeras	0.871	0.831
Google AutoML Vision	**0.919**	0.861

Table 1: Results for BreastMNIST

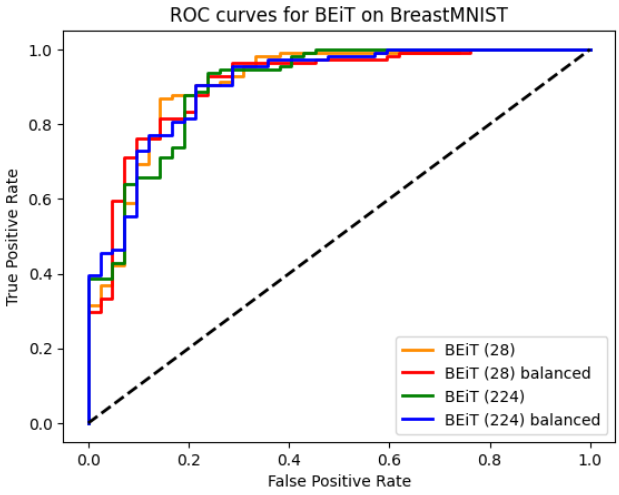

Figure 5: ROC curves for BEiT on BreastMNIST

The results in Table 1 show that our methods achieve outstanding accuracy, surpassing all benchmarks. In particular, the unbalanced 28x28 input version obtains the best result. Regarding the AUC metric, our models perform remarkably well, ranking above all other benchmarks except for the one of Google AutoML Vision. Figure 5 illustrates the ROC curves, confirming the minimal variations in AUC values across the BEiT methods. We notice that the methods trained on the balanced version of the dataset do not achieve better performance, likely due to the limited number of examples available. However, we can see that performance actually benefits from the upscaling, probably because this induces a smoothing effect on the images, reducing noise-like patterns, and therefore helping the model to focus just on meaningful features (see Figure 2).

5.2. PneumoniaMNIST

- All BEiT methods are fine-tuned with batch size of 32 and weight decay of 0.1

Further details are in the Fine-tuning subsection (4.2).

Method	AUC	Accuracy
† BEiT (28)	0.979	0.881
† BEiT (28) balanced	0.974	0.926
† BEiT (224)	0.985	0.912
⋆ BEiT (224) balanced	0.983	0.918
ResNet-18 (28)	0.944	0.854
ResNet-18 (224)	0.956	0.864
ResNet-50 (28)	0.948	0.854
ResNet-50 (224)	0.962	0.884
auto-sklearn	0.942	0.855
AutoKeras	0.947	0.878
Google AutoML Vision	**0.991**	**0.946**

Table 2: Results for PneumoniaMNIST

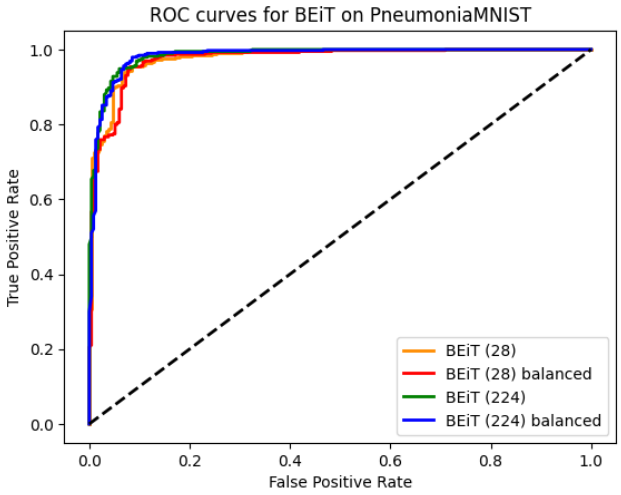

Figure 6: ROC curves for BEiT on PneumoniaMNIST

Table 2 shows that all our BEiT implementations achieve high performance on both metrics. The BEiT (224) balanced method offers the best trade-off between AUC and accuracy, again outperforming all benchmark models except for Google AutoML Vision. From the ROC plots in Figure 6 we can still see where the slight variations come form, meanwhile observing the relative improvement with respect to the experiments on the previous dataset, given by their gathering on the top-left corner. In this case, balancing the training set appears beneficial for accuracy, while it results in a negligible reduction in AUC. On the other hand,

the effect of upscaling the input is less straightforward in this case. In general, it appears to result in a loss of valuable information, likely because fine-grained details in these X-ray images provide important insights (see Figure 3).

6. Conclusion

In this work we have shown how the combination of predictive learning through transformer-based architectures and the strong generalization capabilities of self-supervised approaches, underscores the potential of pre-trained models like BEiT for tackling domain-shifted challenges in specialized applications such as medical imaging. The model outperforms almost every supervised benchmark, by exploiting the power of transfer learning. Future work could focus on exploring additional self-supervised techniques, extending to larger medical datasets, and investigating modern hybrid approaches that combine predictive and contrastive learning paradigms. Furthermore, additional interpretability techniques could be incorporated to better understand model decisions and build trust for clinical applications.

References

[1] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Justin Freyberg, Jamie Deaton, Andy Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, 2021.

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *Advances in Neural Information Processing Systems*, 2021.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[5] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[6] Linus Ericsson, Henry Gouk, and Timothy Hospedales. How well do self-supervised models transfer? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[7] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[9] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[10] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.