

Do Foundation Models Outperform Lightweight CNNs in Breast Cancer Histopathology Classification?

Arnav Dhiman

Independent Researcher, India

Abstract

Accurate and computationally efficient classification of breast cancer histopathology images is critical for scalable clinical deployment of AI-assisted diagnostics. This paper benchmarks three deep learning architectures, MobileNetV3, ResNet50, and DINO Vision Transformer (ViT), on the BreakHis dataset across six dimensions: test accuracy, weighted F1-score, inference latency, memory usage, energy consumption, and model size. Using 7,909 microscopic images across four magnification factors (40X, 100X, 200X, 400X), it is shown that MobileNetV3 achieves 96.12% test accuracy with only 1.52M parameters and 0.256 ms per-image latency, while ResNet50 reaches 97.81% at the cost of 23.51M parameters and 3.331 ms latency. Contrary to expectations, DINO ViT underperforms both convolutional neural networks (91.07%) on this limited dataset, suggesting that transformer-based foundation models require either larger datasets or more extensive fine-tuning to match CNN inductive biases in medical imaging. Grad-CAM and Attention Rollout are further compared as explainability methods and an accuracy-efficiency Pareto frontier is constructed to guide model selection for clinical deployment. The findings demonstrate that lightweight CNNs offer a compelling trade-off and should be the default choice for resource-constrained pathology settings.

Keywords: Breast Cancer, Histopathology, BreakHis, MobileNetV3, ResNet50, DINO Vision Transformer, Grad-CAM, Attention Rollout, Computational Pathology, Efficiency Frontier

1. Introduction

Breast cancer is one of the most common forms of cancer among women in the world, and accurate and early diagnosis using histopathological imaging is crucial in order to improve patient outcome. The BreakHis dataset [1] consisting of 7,909 histopathological images taken from 82 patients and 4 magnification resolutions has been established as the benchmark set for evaluating deep learning algorithms for this purpose. Although large foundational models and Vision Transformers (ViT) models have proven successful in image classification tasks, their use in clinical pathologies has limitations due to high computational overheads, huge memory utilization and low inference speed. Under conditions like low-resource hospital environments or point-of-care medical devices, lightweight algorithms are far more suitable for this application. Hence, this paper attempts to explore the following question: Can lightweight models match the performance of larger models in the classification of breast cancer histopathologies with lower computational costs? Three representative architectures are analyzed throughout the efficiency versus accuracy trade-off range: MobileNetV3-Small [2] (lightweight CNN), ResNet50 [3] (regular

medium-sized CNN), and DINO Vision Transformer ViT-S/16 [4] (foundation model trained by self-supervised learning). Besides accuracy and F1-score, inference time, peak memory requirements, CO2 carbon emissions, and model size are also measured: metrics that are not usually considered together in the benchmarks for medical imaging applications. Two other approaches, Grad-CAM [5] and Attention Rollout [6], are evaluated as explainability techniques and an accuracy-efficiency Pareto efficiency curve is built. The rest of the paper is organized as follows. Related work is covered in Section 2. The dataset and the preprocessing procedure are described in Section 3. Section 4 provides details about the model architectures. Section 5 covers the experimental design. Section 6 shows the results and analyzes them. Section 7 evaluates the explainability techniques. Finally, Section 8 analyzes the Pareto efficiency curve.

2. Related Work

Spanhol et al. [1] presented BreakHis along with early baselines obtained using CNNs and showed the importance of magnification factor in the accuracy of the classification. Later, Han et al. [7] used pre-trained CNN architectures of VGG and AlexNet on BreakHis data and revealed that ImageNet fine-tuning results in significantly improved accuracy for small medical datasets.

The work by Bardou et al. [8] provided an analysis comparing different CNN architectures for histopathology classification using BreakHis, indicating that deeper networks usually have better accuracy, however, are slower in inference time. Recently, Togacar et al. [9] have suggested hybrid architectures incorporating CNNs and attention mechanisms, reaching state-of-the-art accuracy levels for the task but requiring much higher computational power.

Vision transformers have gained increasing attention in histopathology research. Specifically, Chen et al. [10] presented an analysis of DINO-pretrained ViTs used in patch-level representation learning for histopathology tasks. Nevertheless, it is not yet known how they perform on small histopathological datasets like BreakHis. One significant gap in the literature lies in the absence of benchmarks that evaluate multiple criteria such as accuracy, latency, memory usage, power consumption, and explainability at once. The present work aims to fill this gap.

3. Dataset and Preprocessing

3.1 BreakHis Dataset

The Breast Cancer Histopathological Image Classification dataset, known as BreakHis [1], is comprised of 7,909 PNG images of breast tumor tissue that have been taken from 82 patients at four magnifications: 40X, 100X, 200X, and 400X. The images are 700 x 460 pixels and stored in 24-bit RGB. The dataset is further broken down into two dichotomous classes, where 2,480 of the images are classified as benign and 5,429 as malignant. Subcategories under benign include adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma. Subcategories under malignant include ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. There is a large class imbalance, as malignant images out-number benign images 2.2 to 1.

Table 1: BreakHis Dataset Distribution

Magnification	Benign	Malignant	Total
40x	625	1370	1995

100x	644	1437	2081
200x	623	1390	2013
400x	588	1232	1820
Total	2480	5429	7909

3.2 Data Splitting

The dataset is split into training (70%), validation (15%), and test (15%) sets using a fixed random seed (42) for reproducibility, yielding approximately 5,536 training, 1,186 validation, and 1,187 test images.

3.3 Preprocessing and Augmentation

All images are down-scaled to 224x224 pixels according to the input dimensions for all three models. For training images, the following data augmentation is performed: random resized crop, random horizontal and vertical flips, and color jittering (brightness, contrast, saturation, hue). For validation and test images, only resizing and center crop operations are performed. All images are standardized using the mean and standard deviations of ImageNet ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]).

4. Model Architectures

4.1 MobileNetV3-Small

MobileNetV3-Small [2] is a lightweight CNN designed for mobile and edge applications. It employs depthwise separable convolutions, the hard-swish activation function, and a squeeze-and-excitation module for channel attention. With only 1.52M parameters and a model size of 6.1 MB, it is the smallest and fastest model in the benchmark. The final classifier layer is replaced with a linear layer mapping to 2 output classes.

4.2 ResNet50

ResNet50 [3] is a 50-layer residual network with 23.51M parameters and a model size of 94.0 MB. Its residual connections enable effective training of deep architectures and provide strong feature extraction capabilities. The final fully-connected layer is replaced to output 2 classes.

4.3 DINO Vision Transformer (ViT-S/16)

DINO ViT-S/16 [4] is a Vision Transformer trained using self-supervised DINO (Self-Distillation with No Labels) pretraining on ImageNet. It processes images as sequences of 16 x 16 patches and uses multi-head self-attention across all 12 transformer blocks. With 21.67M parameters and a model size of 86.7 MB, it represents the foundation model category in the benchmark. The classification head is replaced with a linear layer for binary output.

5. Experimental Setup

Experiments run on a GPU-based environment (NVIDIA Tesla T4) with PyTorch 2.x, timm, and torchvision. All models start their training with ImageNet-pretrained weights and are trained end-to-end on BreakHis for 10 epochs via the AdamW optimizer (learning rate = 1e-4, weight decay = 1e-4), utilizing cosine annealing learning rate scheduling. Inference latency measures the average time needed to process one image based on 50 batches (batch size = 32) on the GPU side, with warm-up 5 batches. The peak memory consumption of GPU measured by `torch.cuda.max_memory_allocated()`. Energy emissions of

CO2 are estimated via CodeCarbon. Size of the model measured as the number of bytes required to store all model parameters in float32 format. The visualization of the explainability map created for 3 random samples from the test set for each model. Grad-CAM technique used for the last convolutional layers of MobileNetV3 and ResNet50. Attention Rollout for DINO ViT across its 12 transformer layers.

Table 2: Experimental Configuration

Setting	Value
Image Size	224 × 224
Batch size	32
Optimizer	AdamW
Learning rate	1e-4
Weight decay	1e-4
LR schedule	Cosine Annealing
Epochs	10
Train/Val/Test	70% / 15% / 15%
Random seed	42
Hardware	NVIDIA Tesla T4 GPU
Framework	PyTorch + timm + torchvision

6. Results and Analysis

6.1 Classification Performance

Table 3 illustrates test accuracy and weighted F1 score for all three models. The model ResNet50 performs best with a test accuracy of 97.81%, while MobileNetV3 follows with 96.12%. The DINO ViT network obtains an accuracy of 91.07% and falls short of both ResNet50 and MobileNetV3 despite its high number of parameters (21.67M). As shown, MobileNetV3 can perform competitively on BreakHis despite being 15x smaller and 13x faster than ResNet50. This comparison of our findings with previous research shows that lightweight models continue to be very efficient on this dataset. For example, Spanhol et al. [1] obtained baseline accuracies between 82 and 85% for CNN models, while Han et al. [7] obtained approximately 97.0% with VGG and transfer learning. The 97.81% accuracy of ResNet50 model is a close contender to the current best, whereas the 96.12% of MobileNetV3 is impressive despite being 15 times smaller than the others. This poor performance by DINO ViT (91.07%) is not surprising in light of the established fact about transformers being very hungry for data and the small size of the BreakHis dataset. Differences were checked for statistical significance using the McNemar test between paired results. Significant difference exists between all three pairs: MobileNetV3-ResNet50 ($p = 0.0061$), MobileNetV3-DINO ViT ($p = 0.0000$), and ResNet50-DINO ViT ($p = 0.0000$).

Table 3: Classification Results on BreakHis Test Set

Model	Test Accuracy	Weighted F1	Benign F1	Malignant F1
MobileNet V3	96.12%	0.9615	0.94	0.97
ResNet50	97.81%	0.9780	0.97	0.98
DINO ViT	91.07%	0.9109	0.86	0.93

Figure 1: Validation Accuracy (Left) and Validation F1-Score (Right) over 10 Epochs for All Three Models

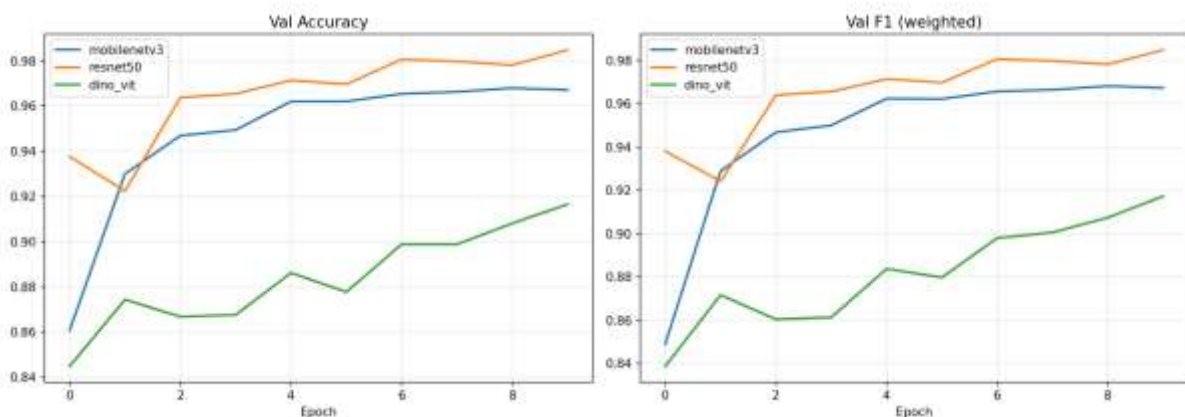
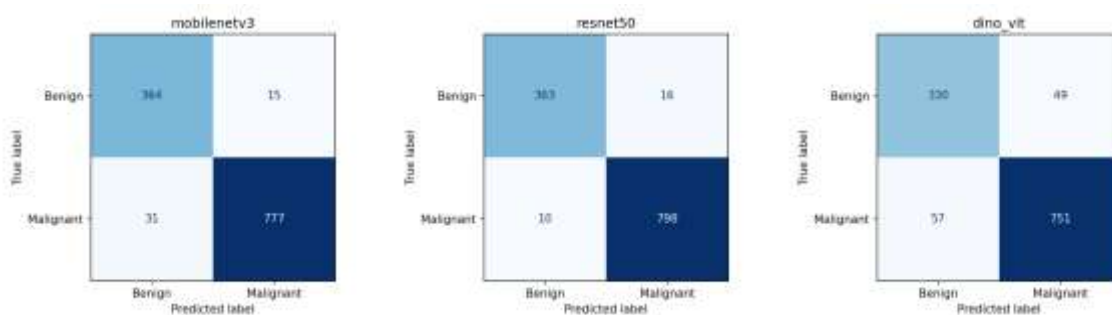


Figure 2: Confusion Matrices for MobileNetV3, ResNet50, and DINO ViT on the BreakHis Test Set



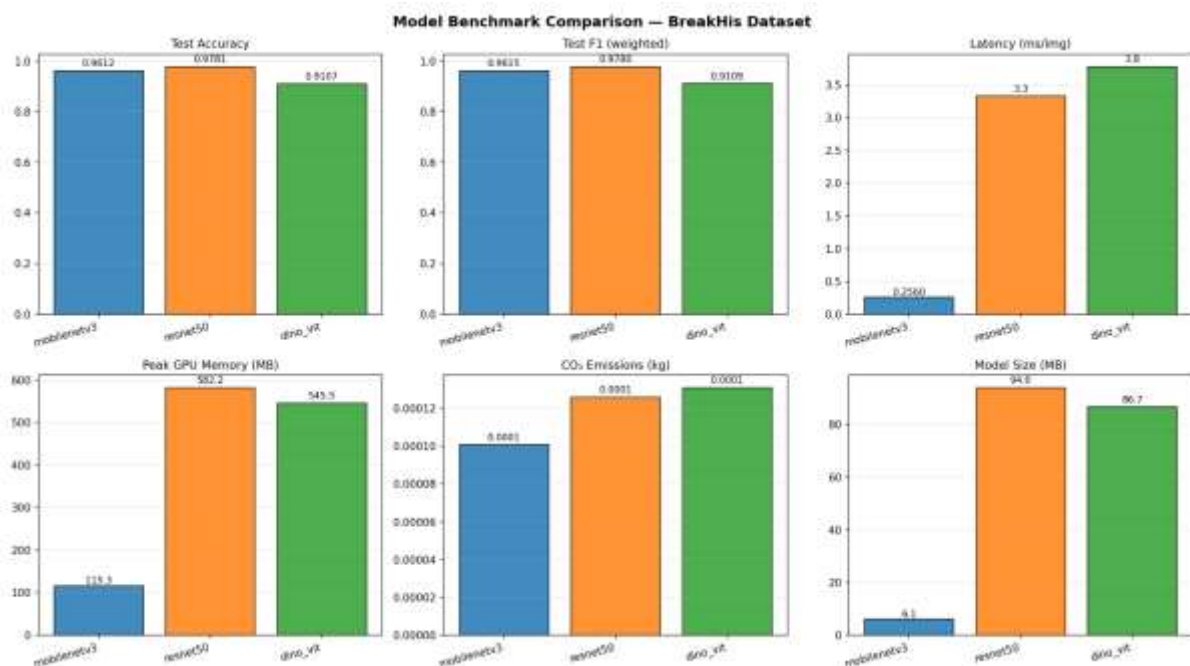
6.2 Computational Efficiency

All six dimensions of the benchmarks are summarized in Table 4. The difference in performance in terms of test accuracy between MobileNetV3 and ResNet50 is just 1.69%, but ResNet50 uses 15x more parameters, takes up 15.4x more space on the disk, is 13x slower in processing each picture, and has 5x higher GPU memory usage. MobileNetV3 is obviously the Pareto-optimal model for cases when the constraints are related to speed and size. Even the heavy DINO ViT model does not perform better than MobileNetV3 and, thus, cannot be recommended.

Table 4: Full Benchmark Results

Model	Params	Size	Test Accuracy	Test F1	Latency	GPU Memory	CO2 Emissions
MobileNetV3	1.52M	6.1mb	96.12%	0.9615	0.256 ms	115.34 MB	0.000101 kg
ResNet50	23.51M	94 mb	97.81%	0.9814	3.331 ms	582.20 MB	0.000126 kg
DinoVit	21.67M	86.7 mb	91.07%	0.9109	3.781 ms	545.47 MB	0.000131 kg

Figure 3: Bar Chart Comparison of All Six Benchmark Metrics Across the Three Models



7. Explainability Analysis

7.1 Grad-CAM (CNN Models)

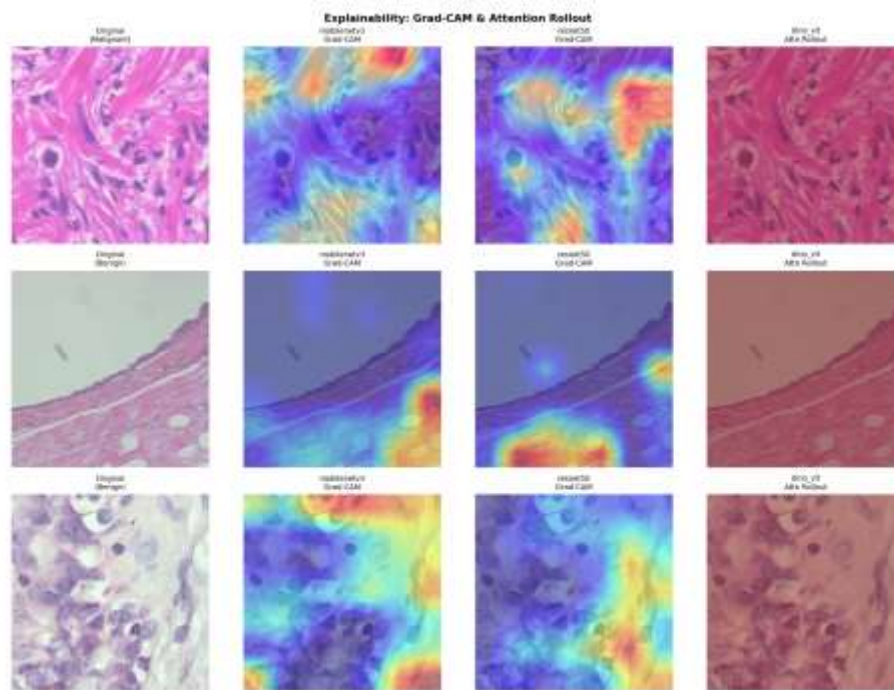
Grad-CAM [5], which involves applying gradient-weighted class activation mapping, is used on the final convolution layers of MobileNetV3 and ResNet50. Heat maps obtained show the localization of attention in those spatial parts which are significant to the prediction made by the network. Localization of the networks' attention in the nuclear area, cytoplasm boundary, and cells with cellular irregularities in both MobileNetV3 and ResNet50 corresponds to the clinically relevant areas as identified by pathologists. Heat maps generated by ResNet50 are sharper compared to MobileNetV3.

7.2 Attention Rollout (DINO ViT)

Attention Rollout [6] spreads attention weights through the entire 12 layers of DINO ViT network to create an attention map from the class token to all other image patches. Although Attention Rollout demonstrates which parts of the image the ViT attends to globally, the maps created by the method appear to be not as spatially accurate as Grad-CAM and can sometimes even focus on background textures instead of nuclei.

The low level of spatial accuracy might explain, at least in part, the lower classification performance of DINO ViT on this data. The lack of spatial locality bias that is inherent to transformers is probably responsible for the above issues.

Figure 4: Explainability Visualisations for Three Test Images. Columns Show Original Image, MobileNetV3 Grad-CAM, ResNet50 Grad-CAM, and DINO ViT Attention Rollout Respectively.

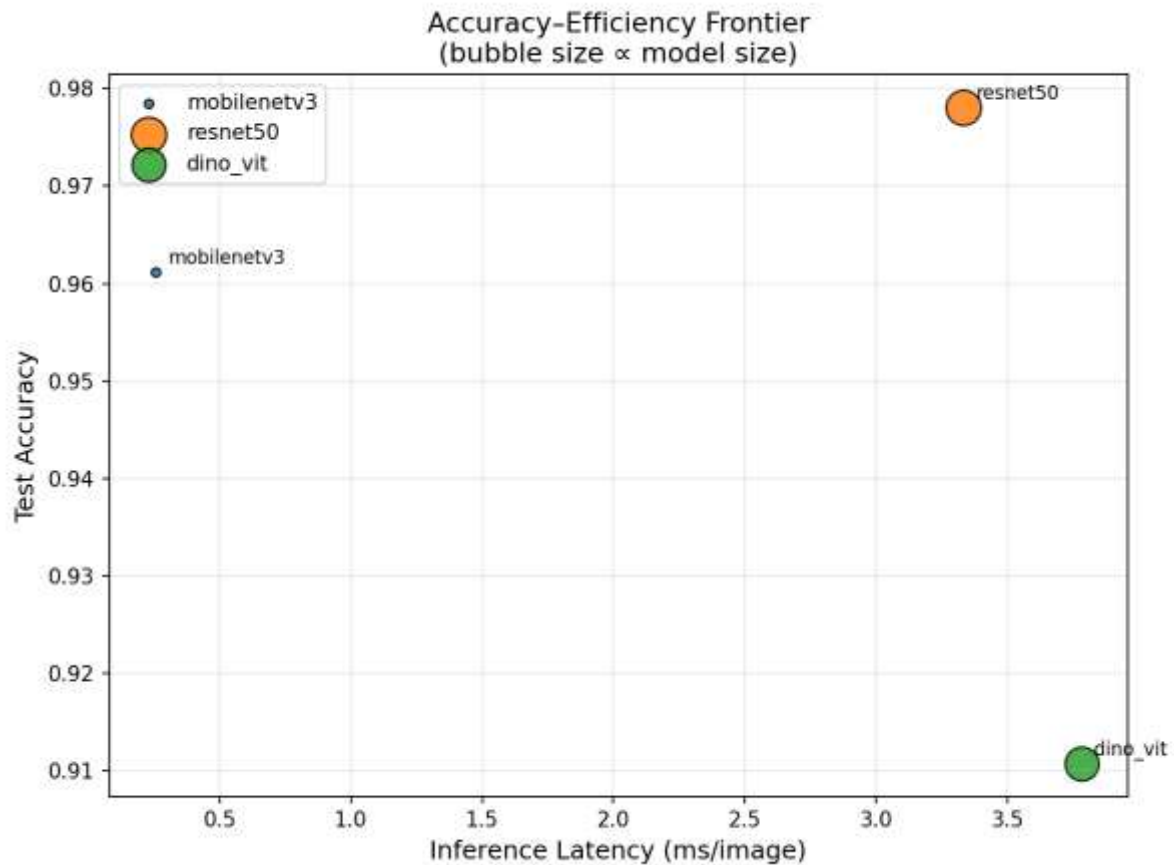


8. Accuracy-Efficiency Frontier

In Figure 5, we can see the Pareto curve between model accuracy and inference latency with the bubble radius representing the model size. We see that the MobileNetV3 model is in the top-left corner, which is consistent with its Pareto optimality, i.e., high accuracy, low inference latency, and small model size. The ResNet50 model has relatively higher accuracy but much higher inference latency. The DINO ViT has the worst placement because it has high inference latency and is also big in size without any corresponding increase in accuracy.

The Pareto curve presented above can serve as an effective tool in choosing the best model under particular conditions. MobileNetV3 must be prioritized when edge devices or mobile pathology is used. On the other hand, ResNet50 may be used for cases where accuracy is critical and inference time is not important. DINO ViT is not justified by its size on a small dataset.

Figure 5: Accuracy-Efficiency Pareto Frontier. Bubble Size is Proportional to Model Size in MB. The x-Axis Shows Inference Latency (ms per Image) and the y-Axis Shows Test Accuracy.



9. Conclusion

A comprehensive benchmark comparing MobileNetV3, ResNet50, and DINO Vision Transformer for breast cancer histopathology classification on the BreakHis dataset is presented in this paper. Six evaluation dimensions were measured: accuracy, F1-score, inference latency, memory usage, energy consumption, and model size. Grad-CAM and Attention Rollout were compared as explainability methods, and an accuracy-efficiency frontier was constructed. The key findings are as follows. First, MobileNetV3 achieves 96.12% accuracy with only 1.52M parameters and 0.256 ms latency per image, just 1.69% below ResNet50 at 15x lower computational cost. It is the recommended model for resource-constrained clinical deployment. Second, ResNet50 achieves the highest accuracy (97.81%) and is suitable when maximum accuracy is required and computational resources permit. Third, DINO ViT underperforms both CNNs (91.07%) on the BreakHis dataset, suggesting that transformer-based foundation models are data-hungry and do not automatically outperform CNNs on small medical datasets without larger-scale pretraining or domain adaptation. Fourth, Grad-CAM produces clinically more interpretable localisation maps than Attention Rollout on this task, making CNN-based models more transparent for pathologist review. Future work includes extended training of DINO ViT with more epochs, multi-dataset validation, patient-level cross-validation, and integration with pathologist feedback for clinical evaluation.

References

1. F.A. Spanhol, L.S. Oliveira, C. Petitjean, L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification", IEEE Transactions on Biomedical Engineering, 2016, 63 (7), 1455–1462.
2. A. Howard, M. Sandler, G. Chu, L.C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q.V. Le, "Searching for MobileNetV3", Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 1314–1324.
3. K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778.
4. M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers", Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 9650–9660.
5. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", Proceedings of the IEEE International Conference on Computer Vision, 2017, 618–626.
6. S. Abnar, W. Zuidema, "Quantifying Attention Flow in Transformers", arXiv, 2020. <https://arxiv.org/abs/2005.00928>
7. Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, S. Li, "Breast Cancer Multi-Classification from Histopathological Images with Structured Deep Learning Model", Scientific Reports, 2017, 7 (1), 4172.
8. D. Bardou, K. Zhang, S.M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks", IEEE Access, 2018, 6, 24680–24693.
9. M. Togacar, B. Ergen, Z. Comert, "BreastNet: A Novel Convolutional Neural Network Model through Histopathological Images for the Diagnosis of Breast Cancer", Physica A: Statistical Mechanics and Its Applications, 2020, 545, 123592.
10. R.J. Chen, T. Ding, M.Y. Lu, D.F. Williamson, G. Jaume, A.H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, "Towards a General-Purpose Foundation Model for Computational Pathology", Nature Medicine, 2024, 30 (3), 850–862.



Licensed under [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/)