MICCAI2025
Daejeon
REPUBLIC OF KOREA

Evaluating the Explainability of Vision Transformers in Medical Imaging

Leili Barekatain and Ben Glocker

Department of Computing,
Imperial College London, United Kingdom

IMPERIAL

GitHub

Introduction

- Understanding model decisions is crucial in medical imaging.
- **Vision Transformers (ViTs)**: state-of-the-art performance.
- **Challenge**:
  - ViT attention mechanisms are complex → explainability unclear.
  - Not all explainability methods (e.g., attention-based or feature attribution approaches) are always effective.
- **Goal**: Systematically evaluate and compare the explainability of different ViTs in medical imaging.

Methods and Architectures

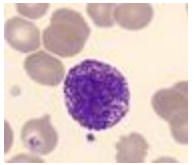
- **Models**:
  - **ViT**: Standard transformer.
  - **DeiT**: Data-efficient, distillation-based ViT.
  - **DINO**: Self-supervised ViT via teacher–student training.
  - **Swin Transformer**: Hierarchical, shifted-window attention.

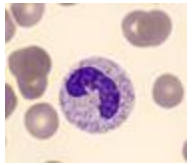
- **Explainability Methods**:
  - **Gradient Attention Rollout**: Aggregates weighted attention across layers.
  - **Grad-CAM**: Highlights class-specific image regions most responsible for the prediction.

Tasks and Datasets

- **Peripheral Blood Cell (PBC) Dataset** [1]:

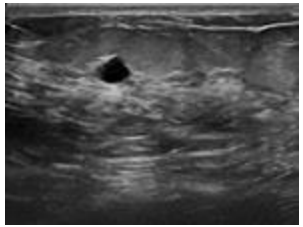
  Images of **eight** blood cell categories, including Basophil, Eosinophil, Erythroblast, Immature Granulocyte, Lymphocyte, Monocyte, Neutrophil, and Platelet.

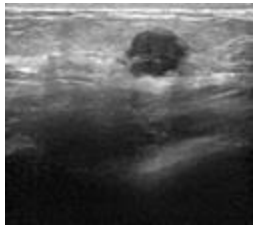




- **Breast Ultrasound Images Dataset** [2]:

  Images of **three** classes, including normal, benign, and malignant.





[1] Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in brief 30, 105474 (2020)
[2] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief 28, 104863 (2020)

IMPERIAL

Results and Analysis

Performance Results

- High accuracy alone is not enough — explainability must also be assessed.

Model	Accuracy (%)	F1-score (%)
ViT	98.68	98.73
DeiT	98.05	97.92
DINO	96.97	97.16
Swin	98.58	98.59

Table 1: Performance Results on PBC

Model	Accuracy (%)	F1-score (%)
ViT	87.18	85.66
DeiT	79.49	75.48
DINO	80.77	77.23
Swin	89.74	88.44

Table 2: Performance Results on Breast Ultrasound

IMPERIAL

Explainability Results — Quantitative

- **Insertion/Deletion:** gradually add or remove important pixels from the heatmap to see how the target class probability changes.
- **DINO + Grad-CAM** gives the best scores for both datasets.

	Grad-CAM				Rollout			
	ViT	DeiT	DINO	Swin	ViT	DeiT	DINO	Swin
Deletion(↓)	0.60	0.38	**0.27**	0.82	0.42	0.52	0.36	0.60
Insertion(↑)	0.44	0.60	**0.75**	0.52	0.44	0.45	0.45	0.56

Table 3: Deletion and Insertion AUC across models for PBC dataset

	Grad-CAM				Rollout			
	ViT	DeiT	DINO	Swin	ViT	DeiT	DINO	Swin
Deletion(↓)	0.63	0.61	**0.51**	0.65	0.61	0.63	0.60	0.56
Insertion(↑)	0.65	0.67	**0.72**	0.69	0.61	0.62	0.62	0.50

Table 4: Deletion and Insertion AUC across models for Breast Ultrasound dataset

IMPERIAL

Explainability Results — Quantitative

- Grad-CAM outperforms Gradient Attention Rollout, with higher AUC in insertion and lower AUC in deletion.
- **Grad-CAM** has better localization of critical visual features.

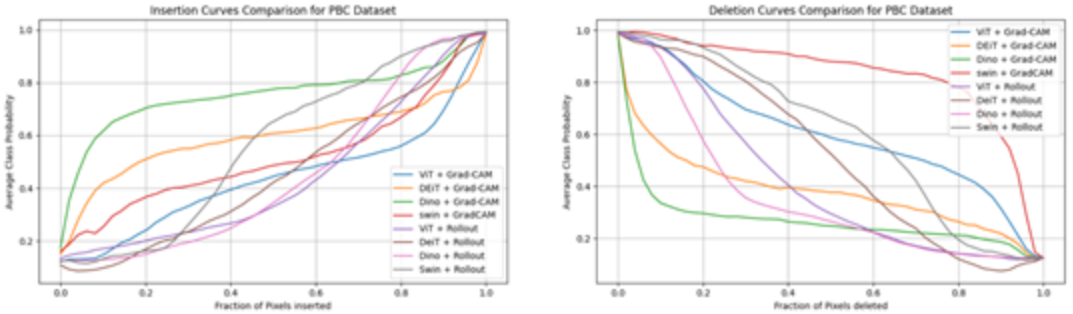


Figure 1: Insertion/Deletion Visualization for PBC Dataset

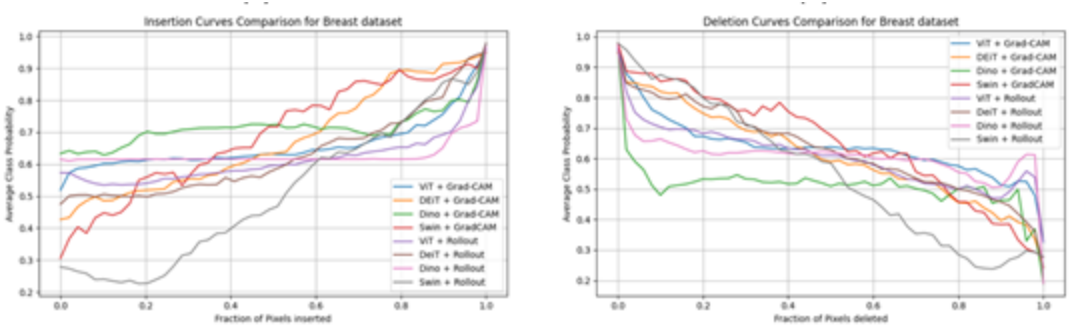


Figure 2: Insertion/Deletion Visualization for Breast Ultrasound Dataset

IMPERIAL

Explainability Results — Qualitative - PBC

- **Grad-CAM** produces more focused, class-specific heatmaps than Gradient Attention Rollout, with **DINO** showing the clearest localization.

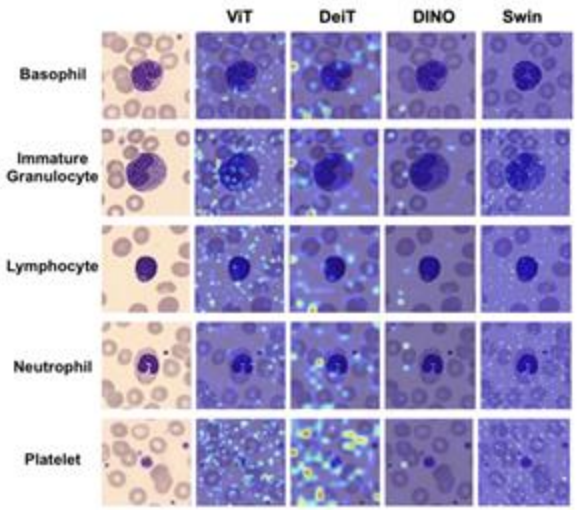


Figure 3: Comparison of Gradient Attention Rollout heatmaps for five blood cell classes from the PBC dataset

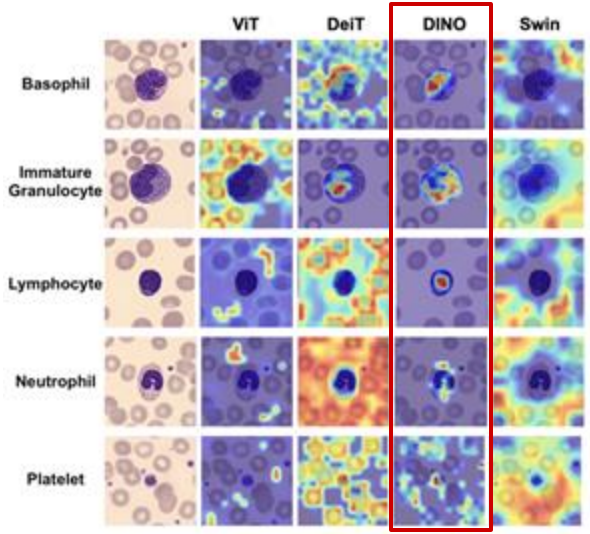


Figure 4: Comparison of Grad-CAM heatmaps for five blood cell classes from the PBC dataset

Explainability Results — Qualitative - Breast

- **Grad-CAM** highlights lesion boundaries (benign) and tumor contours (malignant).
- Gradient Rollout remains scattered and less informative.
- **DINO + Grad-CAM** localizes clinically meaningful regions most consistently.

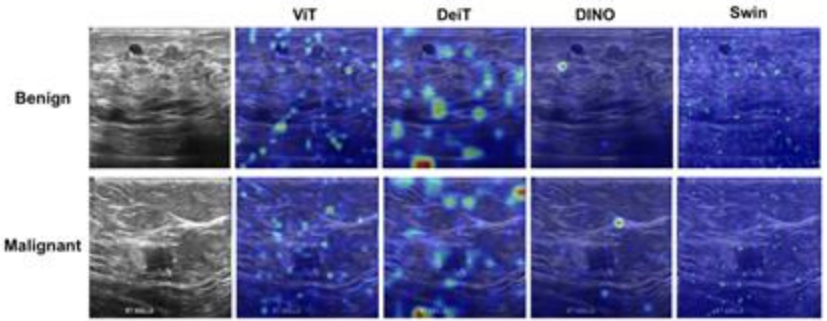


Figure 5: Comparison of Gradient Attention Rollout heatmaps for benign and malignant breast ultrasound images

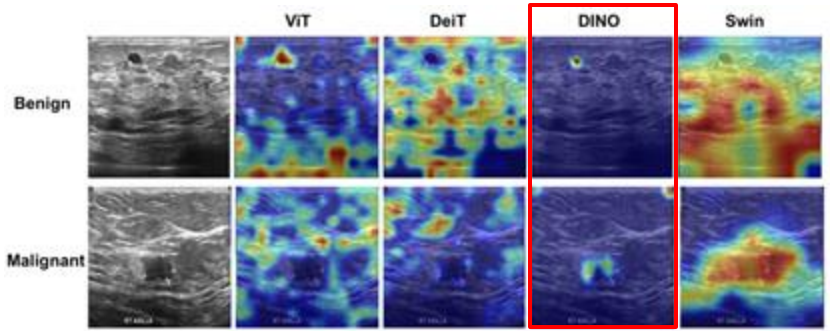


Figure 6: Comparison of Grad-CAM heatmaps for benign and malignant breast ultrasound images

Qualitative Error Analysis

- **Grad-CAM** can reveal the underlying reasons behind model misclassifications.

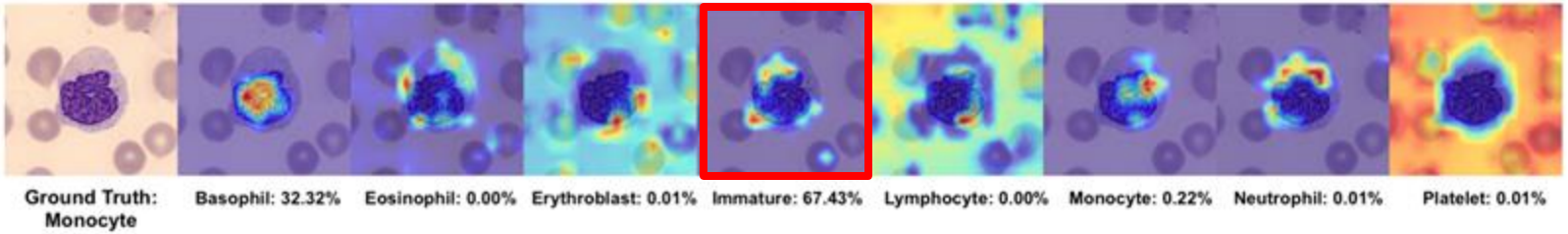


Figure 7: Grad-CAM visualizations of misclassified samples from the PBC dataset using the **DINO-ViT** model

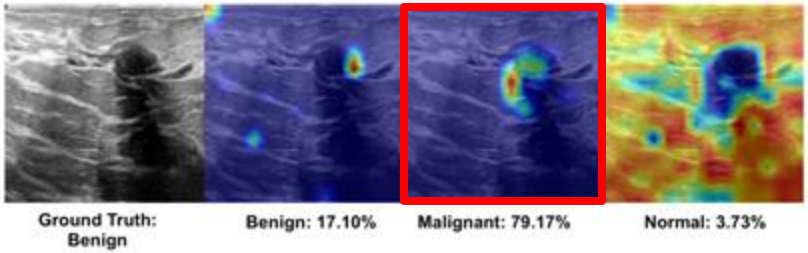


Figure 8: Grad-CAM visualizations of misclassified samples from the Breast Ultrasound dataset using the **DINO-ViT** model.

Conclusion

- Evaluated ViT, DeiT, DINO-ViT, Swin Transformer with Grad-CAM and Gradient Attention Rollout on PBC and Breast Ultrasound datasets.
- All models: High accuracy, but explainability varied.
- Grad-CAM: More localized, class-discriminative than Gradient Attention Rollout.
- DINO + Grad-CAM: Most interpretable setup, even in misclassifications.
- Implication: Model choice in medical imaging should weigh both accuracy and interpretability.

Future Work

- Develop ViT-specific explainability methods with higher faithfulness.
- Explore hybrid techniques (spatial precision + semantic understanding).
- Integrate domain priors/medical constraints to enhance interpretability.

Thank you for your attention!

Contact:
leili.barekatain24@imperial.ac.uk

GitHub

IMPERIAL

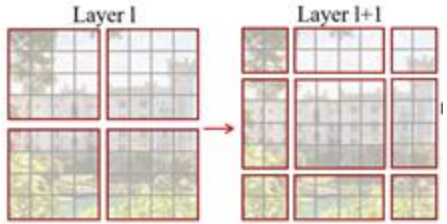
Appendices

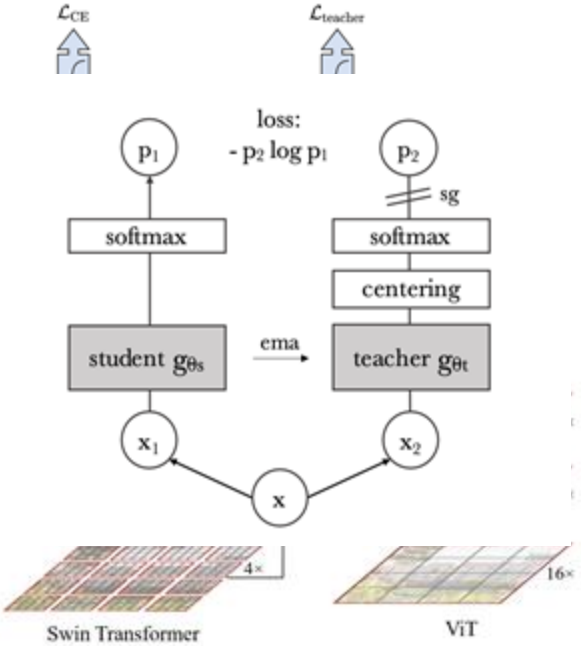
Different Architectures of ViTs

DeiT [2]

DINO [3]

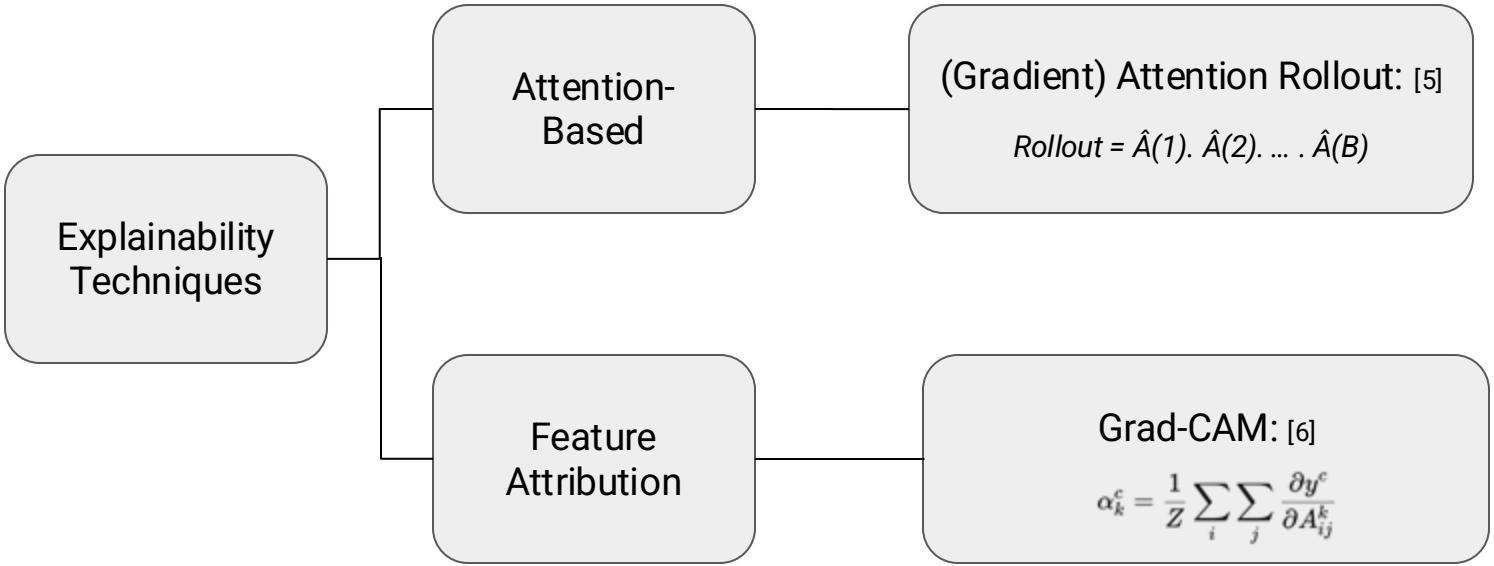
Swin Transformer [4]





[2] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve Jegou. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877, 2021. pages 3, 4
[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294, 2021. pages 4
[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021. pages 5

Explainability Methods



[5] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4190–4197. Association for Computational Linguistics, 2020. pages 6
[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626, 2017. pages 7

References

1. Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020. pages 2
2. Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve Jegou. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877, 2021. pages 3, 4
3. Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294, 2021. pages 4
4. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021. pages 5
5. Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 4190–4197. Association for Computational Linguistics, 2020. pages 6
6. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626, 2017. pages 7
7. Piotr Komorowski, Hubert Baniecki, and Przemyslaw Biecek. Towards evaluating explanations of vision transformers for medical imaging. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3726–3732, 2023. pages 8, 9
8. Arnab Kumar Mondal, Arnab Bhattacharjee, Parag Singla, and AP Prathosh. xvitcos: explainable vision transformer based covid-19 screening using radiography. IEEE Journal of Translational Engineering in Health and Medicine, 10:1 – 10, 2021. pages 9
9. Andrea Acevedo, Anna Merino, Santiago Alferez, Angel Molina, Laura Bold´ u, and Jose Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in brief, 30:105474, 2020. pages 10