HERTIE INSTITUTE FOR
AI IN BRAIN HEALTH

# A hybrid Fully Convolutional CNN-Transformer Model for Inherently Interpretable Disease Detection from Retinal Images

*Kerol Djoumessi, Samuel Ofosu Mensah, Philipp Berens*

*Hertie Institute for AI in Brain Health, University of Tübingen, Germany*

# Motivation

HERTIE INSTITUTE FOR AI IN BRAIN HEALTH

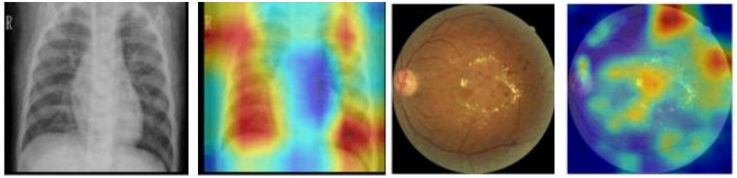
## What is interpretability?

The degree to which a human can understand the cause of a decision *(Miller, 2019)[1].*

## Why interpretable models?

- Learning tools
- Improve user trust
- Discover harmful biases
- Policy/legal considerations

## Why did the model predict this?



**Attribution maps**: highlight relevant regions ~ *Post-hoc explanation*

## Limitations of attribution maps

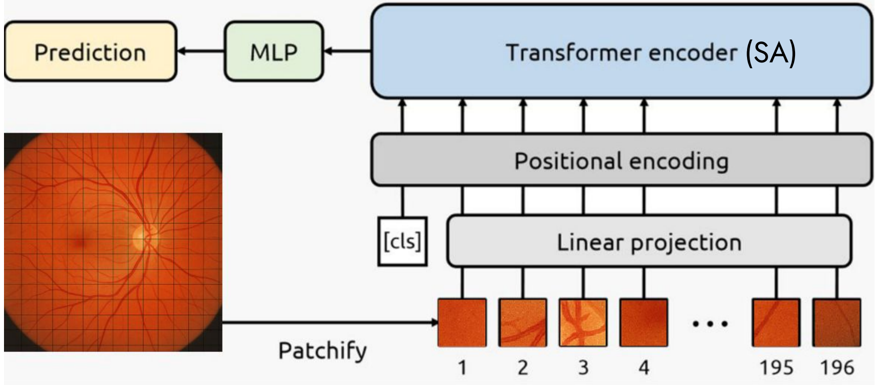
- Coarse-grained explanation
- Not inherently interpretable
- Poor performance on medical data
- Mostly developed for CNN-based models

# Transformers for Image Analysis

HERTIE INSTITUTE FOR AI IN BRAIN HEALTH

## Vision transformers

- Revolutionized computer vision
- Competitive alternative to CNNs
- Self attention long-range dependencies
- Computational intensive [2]




## Hybrid CNN-Transformers

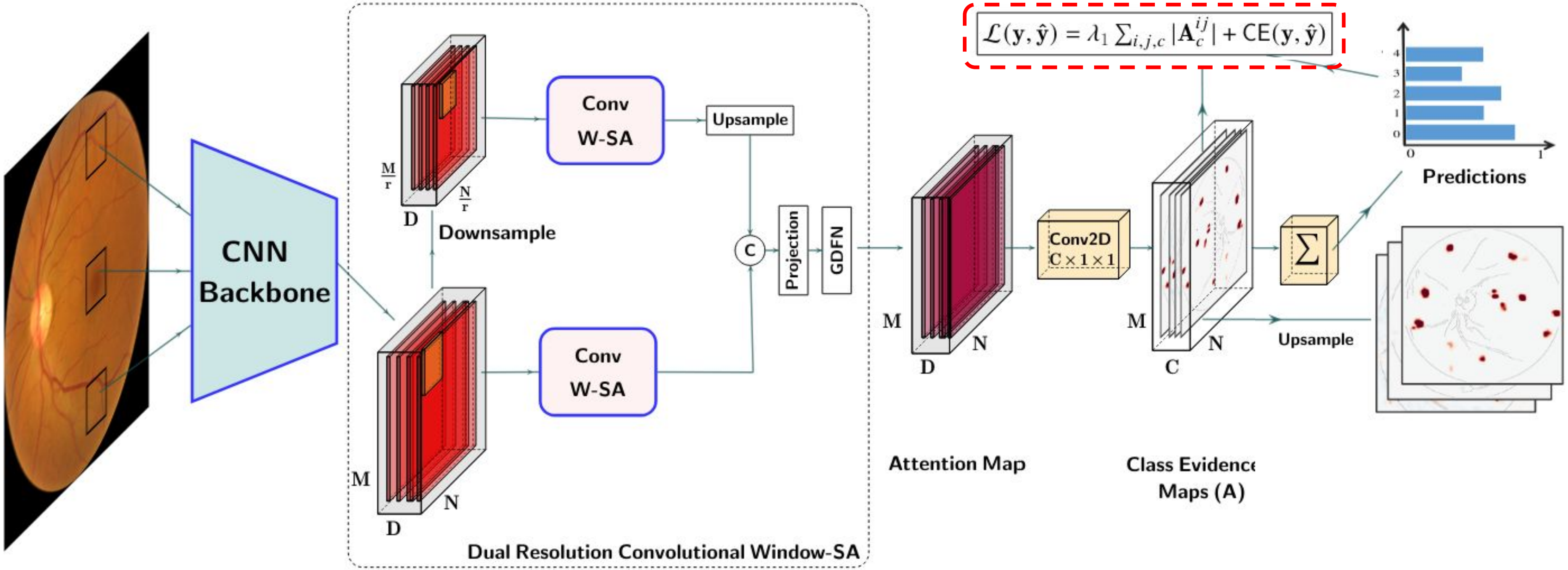
- Balance computational cost
- CNNs + Transformer module
- Locality + global dependencies
- Interpretability remains challenging [2]

## Explaining hybrid CNN-transformers

- Post-hoc for CNNs: GradCAM [3], LRP [4], ....
- Specialized techniques tailored for ViTs
  - Attention maps [5], Rollout [6], ...
- Post-hoc vs self-explainable methods
  - Prototypes-based for ViTs [7]

# Contribution: Self-explainable CNN-Transformer

HERTIE INSTITUTE FOR AI IN BRAIN HEALTH




***W-SA**: Window Self Attention*

***DRSA**: Ilyas et al., MICCAI (2024);* ***Sparse Activations:** Djoumessi et al., MIDL (2023)*

# Application: Retinal disease detection

HERTIE INSTITUTE FOR AI IN BRAIN HEALTH

## About DR and AMD

- Microvascular abnormalities
  - DR: Microeurysms, Hard Exudates
  - AMD: Drusen
- Can lead to vision loss
- Well-defined grading systems

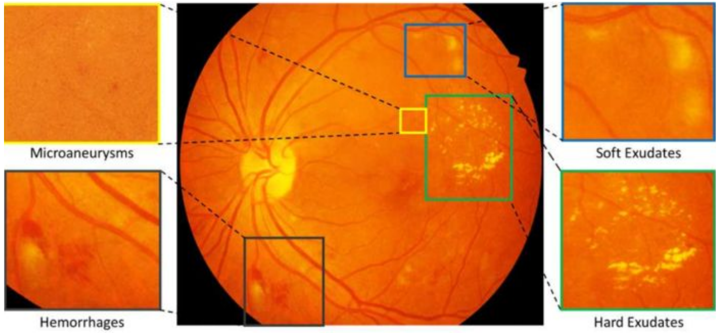
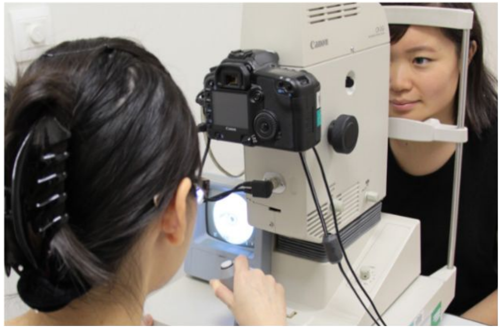



*Figure reproduced from Powal et al. (2018)*

## Managing DR and AMD

- AI-based retinal image analysis
- Predict disease stage from lesions
- Early diagnosis improves treatment
- Progression: regular follow-ups



# Classification performance

HERTIE INSTITUTE FOR AI IN BRAIN HEALTH

Dataset:

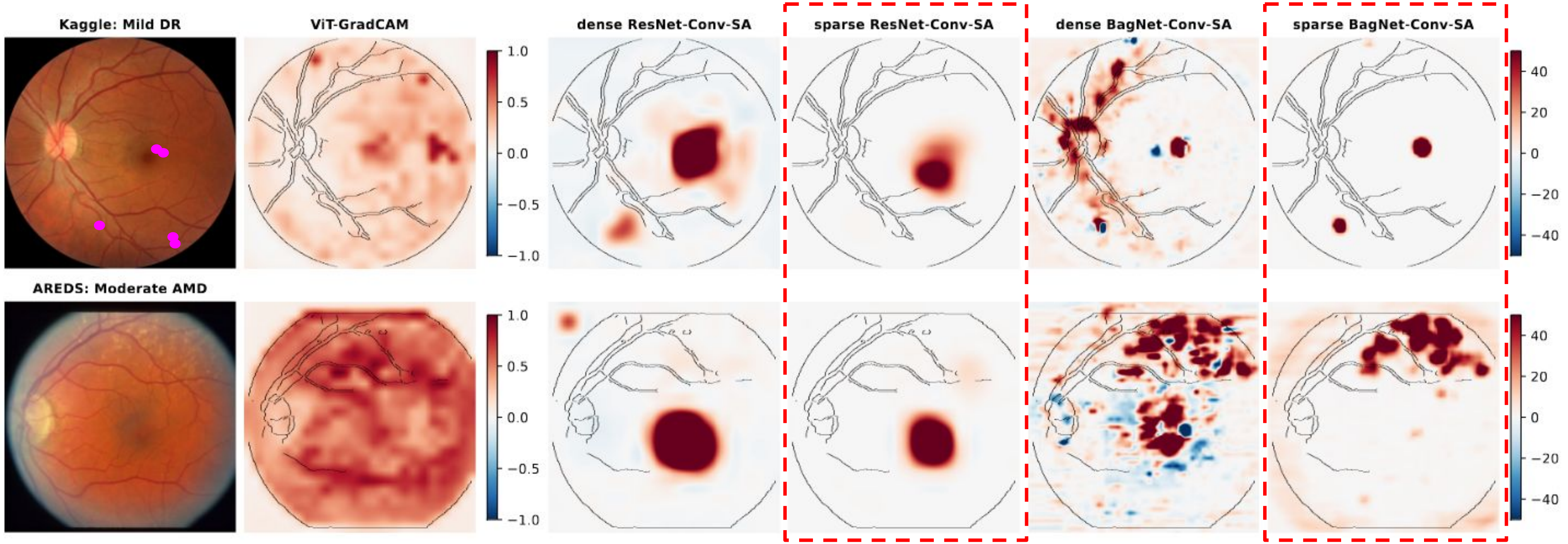
**Kaggle DR detection** (45,923; 5 classes), **AREDS AMD** (34,079; 6 classes)

| | AREDS AMD | | Kaggle DR | | Computational Cost | | |
|---|---|---|---|---|---|---|---|
| | **Acc.** | $\kappa$ | **Acc.** | $\kappa$ | **Par.** | **Mem.** | **Time** |
| ViT | $.76 \pm .03$ | $.90 \pm .02$ | $.81 \pm .02$ | $.71 \pm .04$ | 86, 094 | 341 | $09.5 \pm 0.1$ |
| Swin | $.78 \pm 0.2$ | $\mathbf{.92 \pm .02}$ | $.85 \pm .02$ | $.81 \pm .03$ | 86, 883 | 358 | $15.5 \pm 1.1$ |
| ResNet | $.78 \pm .03$ | $.89 \pm .02$ | $.85 \pm .02$ | $.81 \pm .03$ | 23, 518 | 101 | $04.2 \pm 0.5$ |
| BagNet | $.75 \pm .03$ | $.88 \pm .02$ | $.86 \pm .02$ | $.83 \pm .03$ | 16, 271 | 193 | $15.1 \pm 0.1$ |
| ResNet-FCL-SA | $.78 \pm .03$ | $.90 \pm .02$ | $.86 \pm .02$ | $.82 \pm .03$ | 69, 732 | 281 | $06.2 \pm 0.2$ |
| BagNet-FCL-SA | $.77 \pm .03$ | $.89 \pm .02$ | $.85 \pm .02$ | $.83 \pm .03$ | 62, 501 | 306 | $27.3 \pm 0.2$ |
| ResNet-Conv-SA | $.78 \pm .03$ | $.91 \pm .02$ | $.85 \pm .02$ | $.83 \pm .03$ | 69, 735 | 285 | $06.3 \pm 0.6$ |
| BagNet-Conv-SA | $.77 \pm .03$ | $.90 \pm .02$ | $\mathbf{.87 \pm .02}$ | $\mathbf{.84 \pm .02}$ | 62, 913 | 310 | $27.3 \pm 0.3$ |
| sResNet-Conv-SA | $\mathbf{.79 \pm .02}$ | $.90 \pm .02$ | $.85 \pm .02$ | $.80 \pm .03$ | 69, 735 | 285 | $06.3 \pm 0.6$ |
| sBagNet-Conv-SA | $.77 \pm .03$ | $.91 \pm .02$ | $.85 \pm .02$ | $.81 \pm .03$ | 62, 913 | 310 | $27.3 \pm 0.3$ |

***Sparsity*** *AREDS AMD. ResNet: 1e-04, BagNet: 7e-06*
*Kaggle DR. ResNet: 2e-4, BagNet: 3e-05*

***Parameter (M), Memory (Mb): Time (s)***

# Qualitative heatmap evaluation

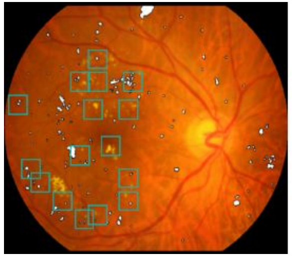


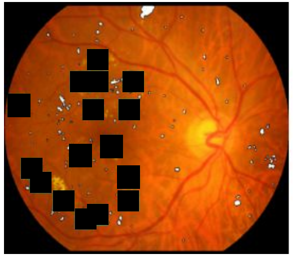
# Quantitative Heatmap Evaluation

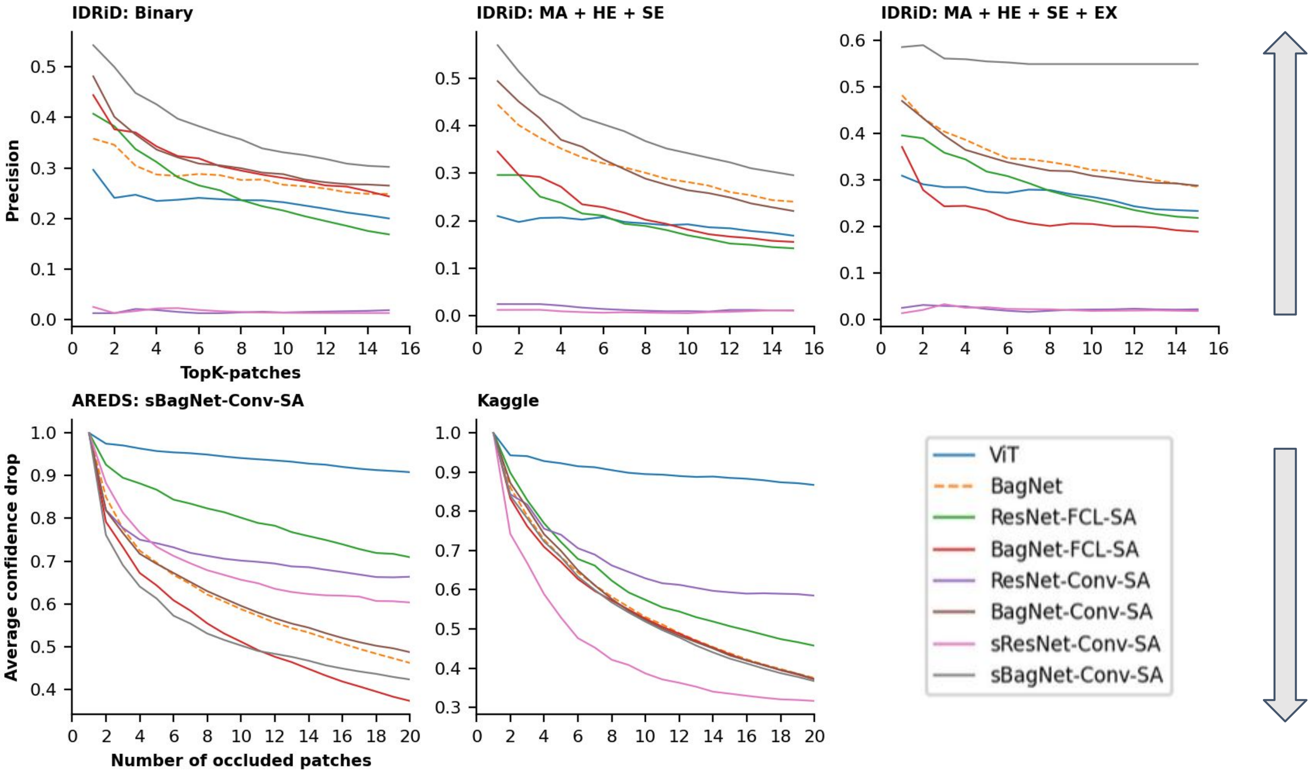
HERTIE INSTITUTE FOR AI IN BRAIN HEALTH

IDRiD dataset: 81 images

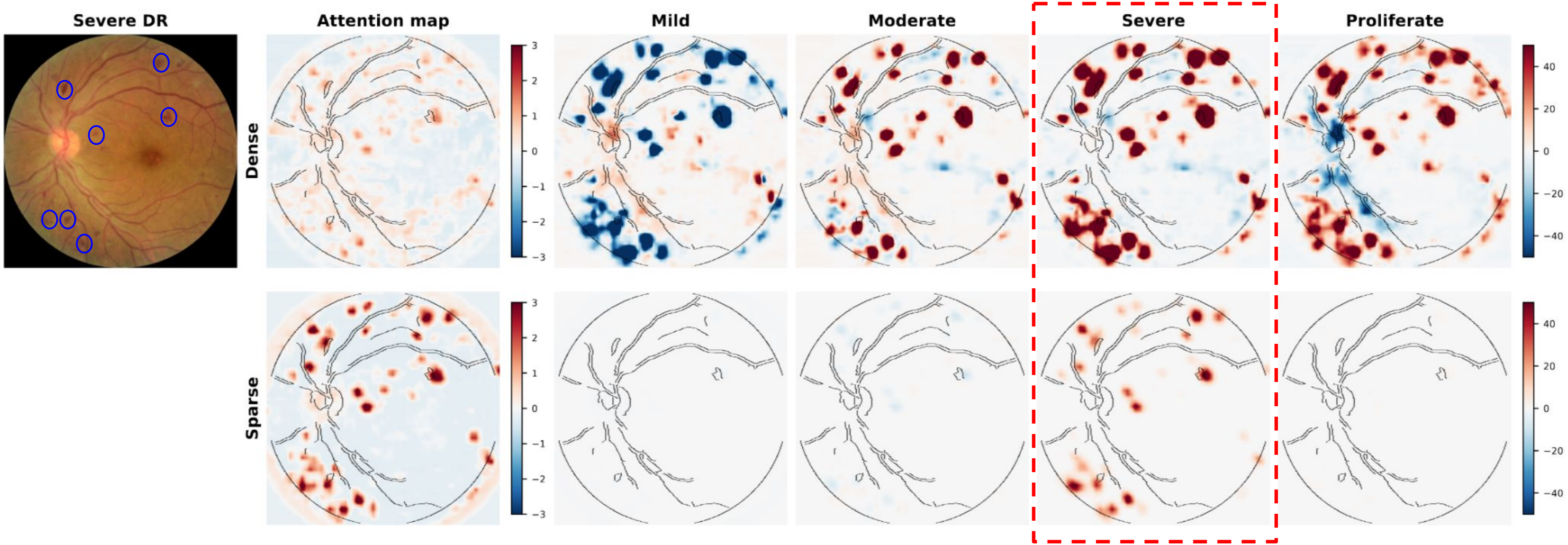
- Lesion annotations







# Interpretability for Multi-class DR Detection



# Conclusion

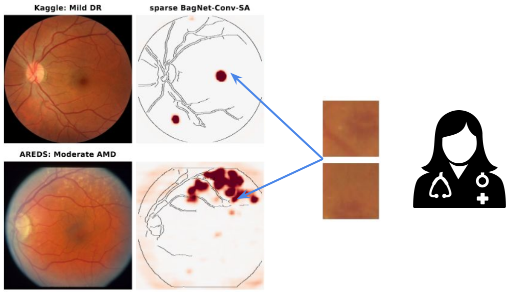
HERTIE INSTITUTE FOR AI IN BRAIN HEALTH

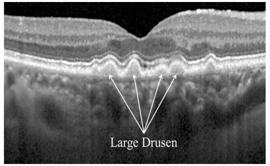
## Summary

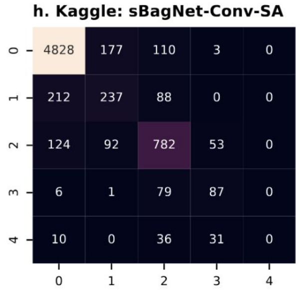
- Effectively combine prior work (DRSA, CvT, sparse explanation)
- Hybrid fully convolutional interpretable CNN-Transformer models
- CNNs: BagNet & ResNet
- Task: retinal disease detection (AMD, DR)
- Qualitative & quantitative metrics to access the interpretability

## Future work

- Clinical user validation
- Generalizability: extend to other modalities
- Improve model performance on data regime (late-stage DR)
- Aggregation techniques & extensive hyperparameters search



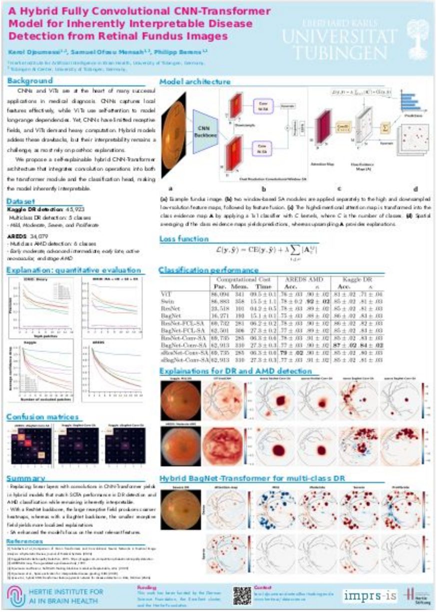



# Acknowledgments

HERTIE INSTITUTE FOR
AI IN BRAIN HEALTH

*Thank you for your attention!*


**Q & A**

# References

HERTIE INSTITUTE FOR AI IN BRAIN HEALTH


*[1] Miller Tim, Explanation in artificial intelligence: Insights from the social sciences (2019)*

*[2] Kim et al., Systematic review of hybrid vision transformer architectures for radiological image analysis (2025)*

*[3] Selvaraju et al., GradCAM: Visual explanations from deep networks via gradient-based localization (2017)*

*[4] Bach et al., On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation (2015)*

*[5] Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020)*

*[6] Abnar et al., Rollout: Quantifying attention flow in transformers (2020)*

*[7] Ashkan et al., Protoformer: Embedding Prototypes for Transformers (2022)*

*[8] Ilyas et al., A Hybrid CNN-Transformer Feature Pyramid Network for Granular Abdominal Aortic Calcification Detection from DXA Images (2024)*

*[9] Wu et al., CvT: Introducing convolutions to vision transformers (2021)*

*[10] Djoumessi et al., Sparse Activations for Interpretable Disease Grading (2023)*

*[11] Djoumessi et al., Soft-CAM: Making black box models self-explainable for high-stakes decisions (2025)*