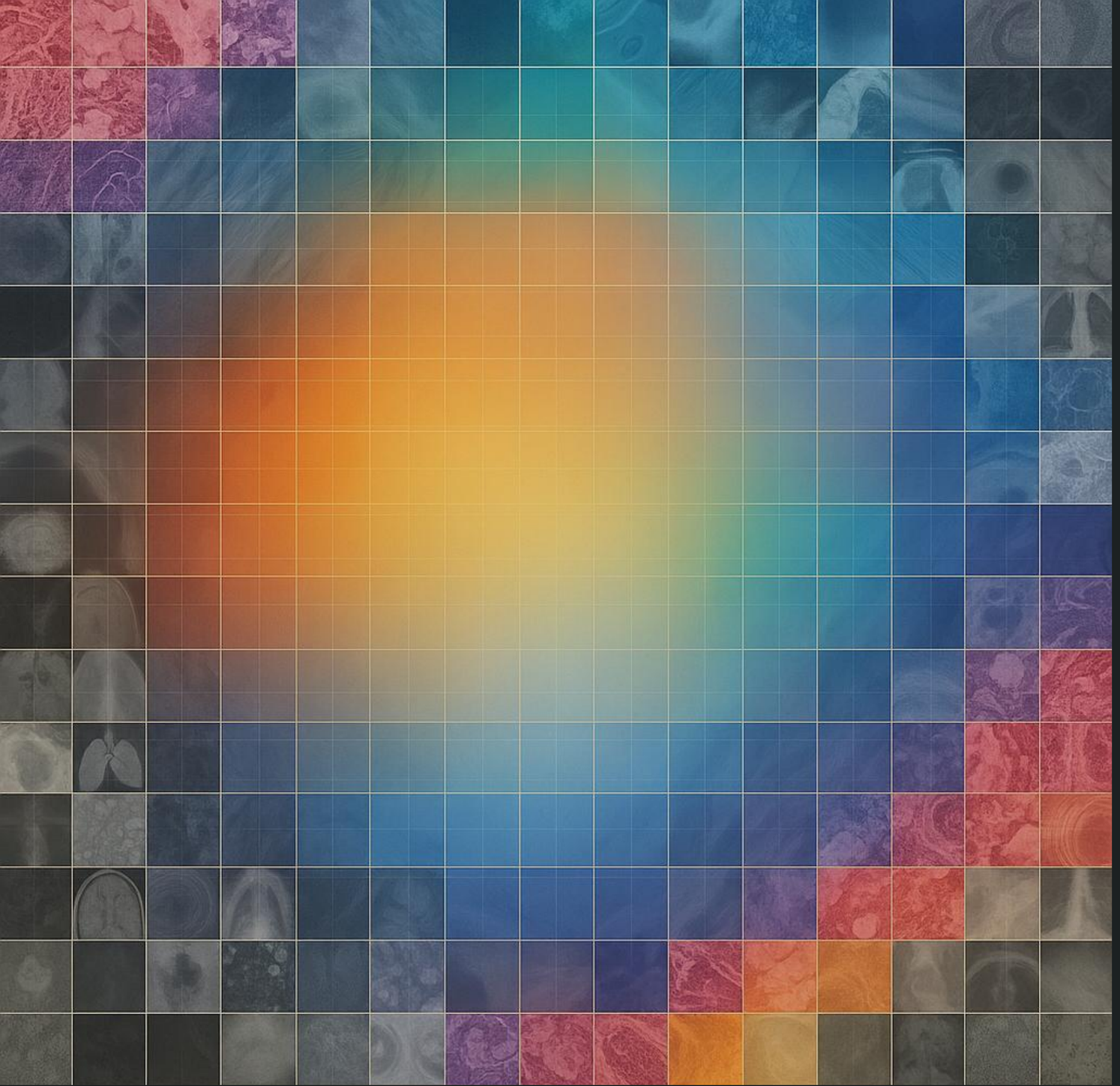
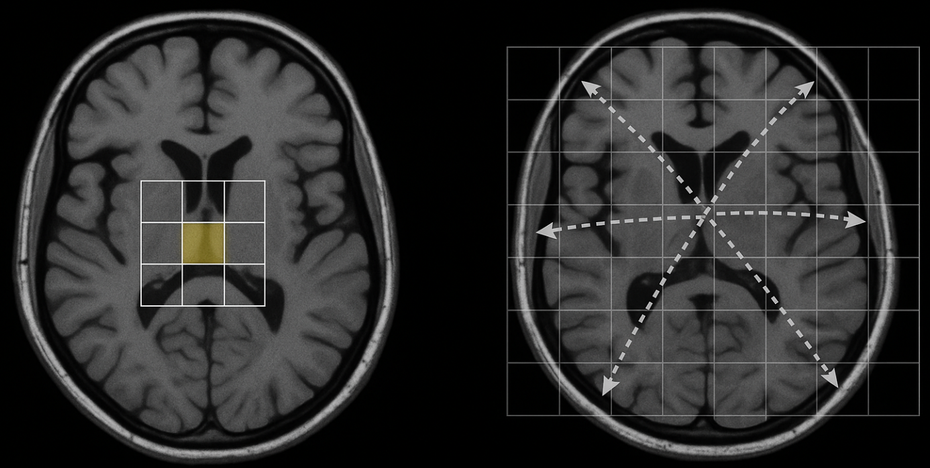


UCD School of Medicine
Scoil an Leighis UCD

Vision Transformers (ViT): A Deep Learning Algorithm for Medical Image Analysis.

Anastasiia Deviataieva

Introduction



CNN receptive field

ViT global self-attention

Data volume explosion	Modern medical scans are high-resolution and volumetric (e.g. a CT scan can be 512×512×300 voxels, ~79 M datapoints), which pose challenges for traditional computer vision models
Local-filter limitation	CNNs (e.g. U-Net) achieve state-of-the-art segmentation yet operate with local receptive fields only
Vision Transformers	ViTs (Dosovitskiy et al., 2020) apply the Transformer architecture (originating from NLP) to images. By using multi-head self-attention → long-range spatial dependencies are captured in one layer
Goal	Exploit ViT to raise accuracy in detection, segmentation, classification and clinical interpretation

Traditional approach: CNNs

CNN Pipeline — Generic Form:

- Input – raw image or multi-channel data array
- Convolution + Activation (e.g., ReLU) – sliding kernels generate local feature maps
- Pooling / Sub-sampling – shrink spatial size while retaining salient responses
- Stacked Convolutional Blocks – wider receptive fields, progressively abstract features
- Flatten or Global Pooling – convert 3-D feature tensor to 1-D feature vector
- Fully Connected layer(s) – integrate all learned features
- Output head (Softmax, Sigmoid, etc.) – produces final class probabilities or regression values

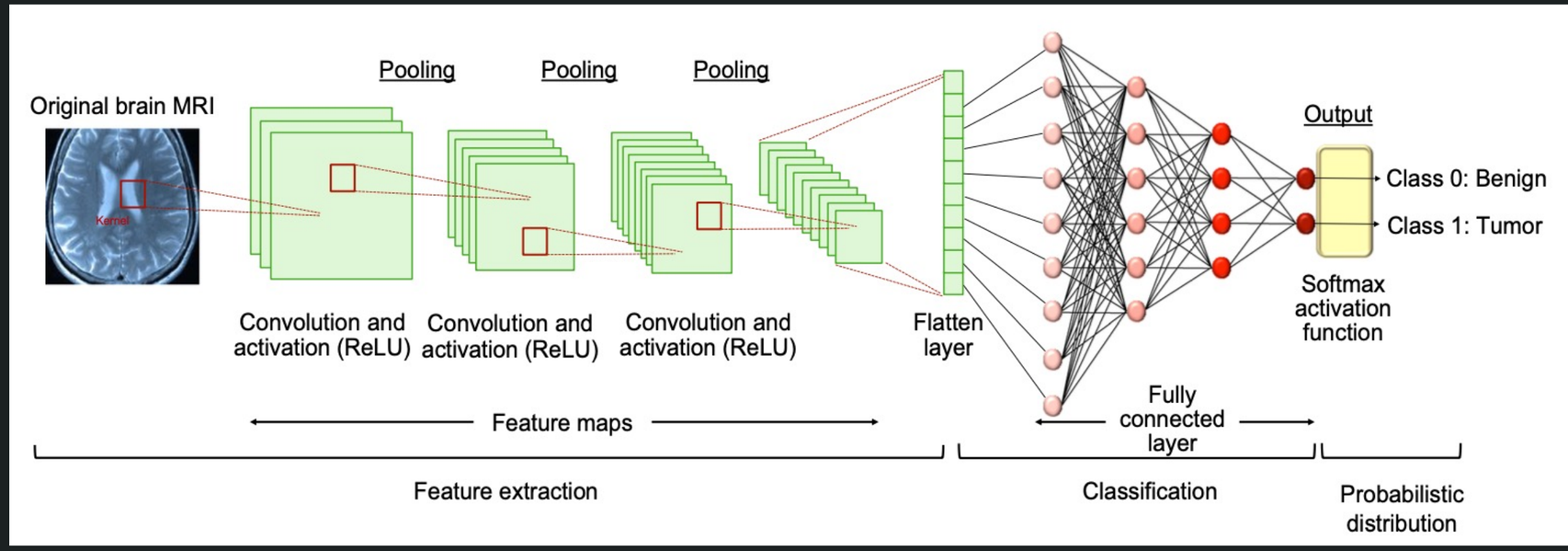


Figure 1. An example of medical image analysis using CNN architecture (brain MRI). (Takahashi et al., 2024)

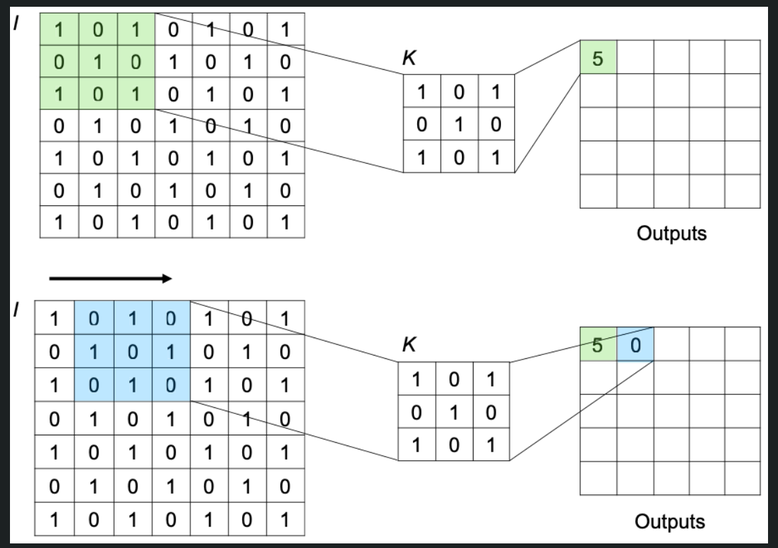


Figure 2. Illustration of the convolution operation. The input matrix I is convolved with the kernel K to produce the output matrix . (Takahashi et al., 2024)

Attention mechanism

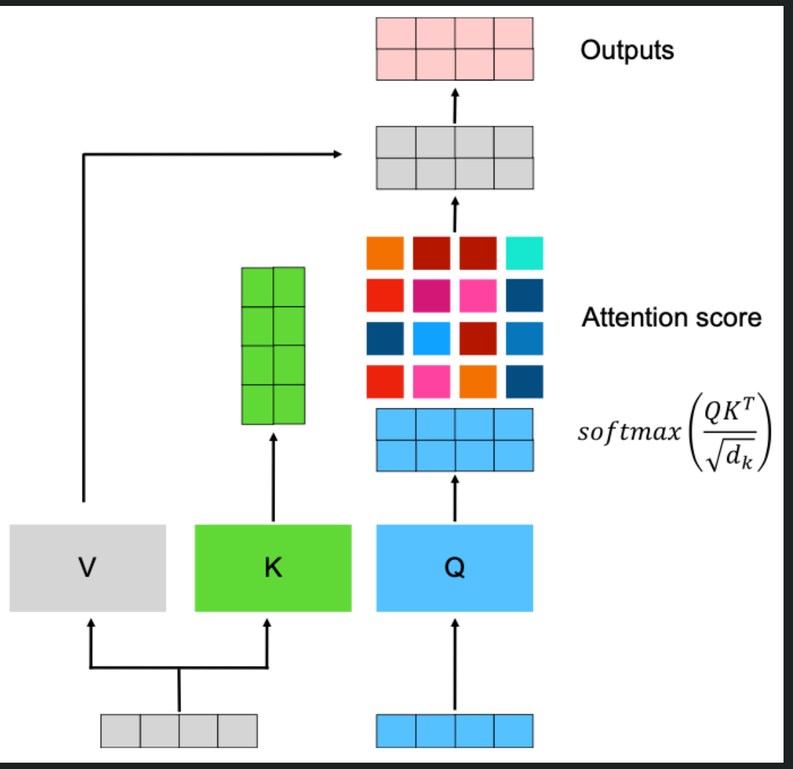


Figure 3. Illustration of the attention operation. The input data are split into three matrices: queries (Q), keys (K), and values (V). These matrices are generated from the input data through learned linear transformations. d – dimension. (Takahashi et al., 2024)

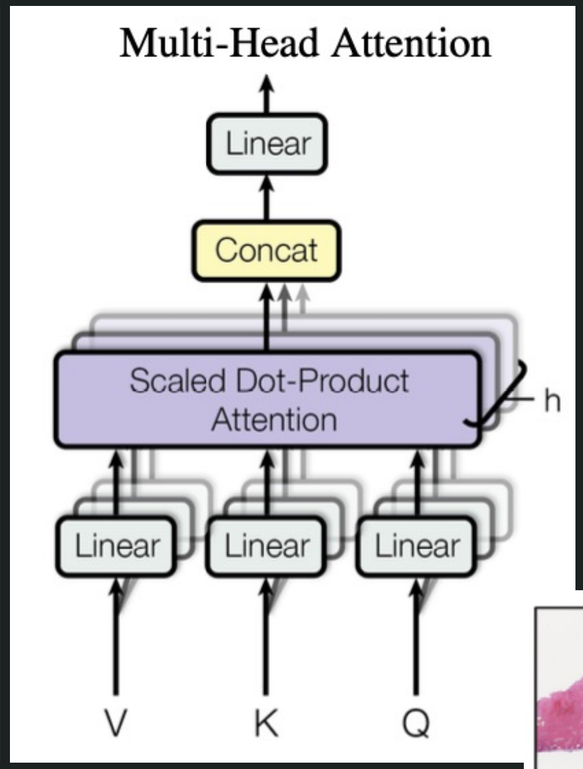


Figure 4. Multi-Head Attention consists of several attention layers running in parallel. (Vaswani et al., 2017)

"Attention is All You Need": In 2017, Vaswani et al. introduced the Transformer architecture in NLP, built on self-attention.

- Scaled dot-product attention turns the similarity between each query and every key into weights that highlight the most relevant value vectors, allowing the network to focus dynamically on different parts of the input.
- Multi-head attention runs the same mechanism in several parallel heads with independent projections and then concatenates their outputs, so the model can capture multiple kinds of relationships across the entire sequence simultaneously.

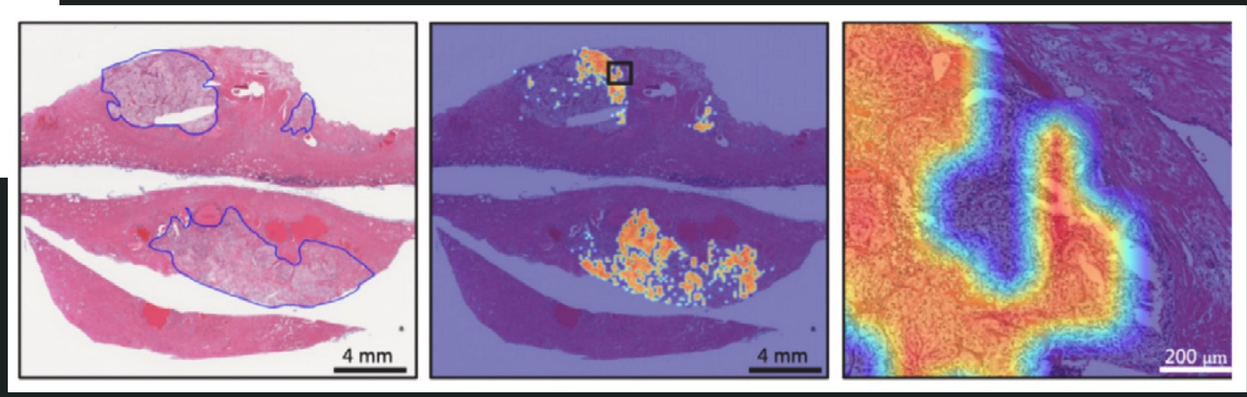


Figure 5. Left: The area within the blue region is the cancer region. Middle: Attention scores from TransMIL are visualized as a heatmap (red for tumor and blue for normal) to interpret the important morphology used for diagnosis. Right: Zoomed-in view of the black square in the middle figure. (Shamshad et al., 2023)

What is Vision Transformer (ViT)?

ViT = Transformer applied to images

Patch tokenization: image is split into fixed-size patches (e.g. 16×16 pixels)

Linear embedding + Position embedding: each patch flattened & projected to a vector; a learned positional encoding is added to each embed, so the model knows each patch's location in the original image

Transformer encoder: the sequence of patch embeddings (plus class token if used) is fed into a transformer encoder – typically a stack of multi-head self-attention layers and feed-forward layers (with layer normalization and skip connections)

Output: For image classification, the Transformer's output class token is used to predict the class label. For other tasks like segmentation -- the output embeddings of all patches (which now contain rich contextual information) and reshape or decode them into an output image mask.

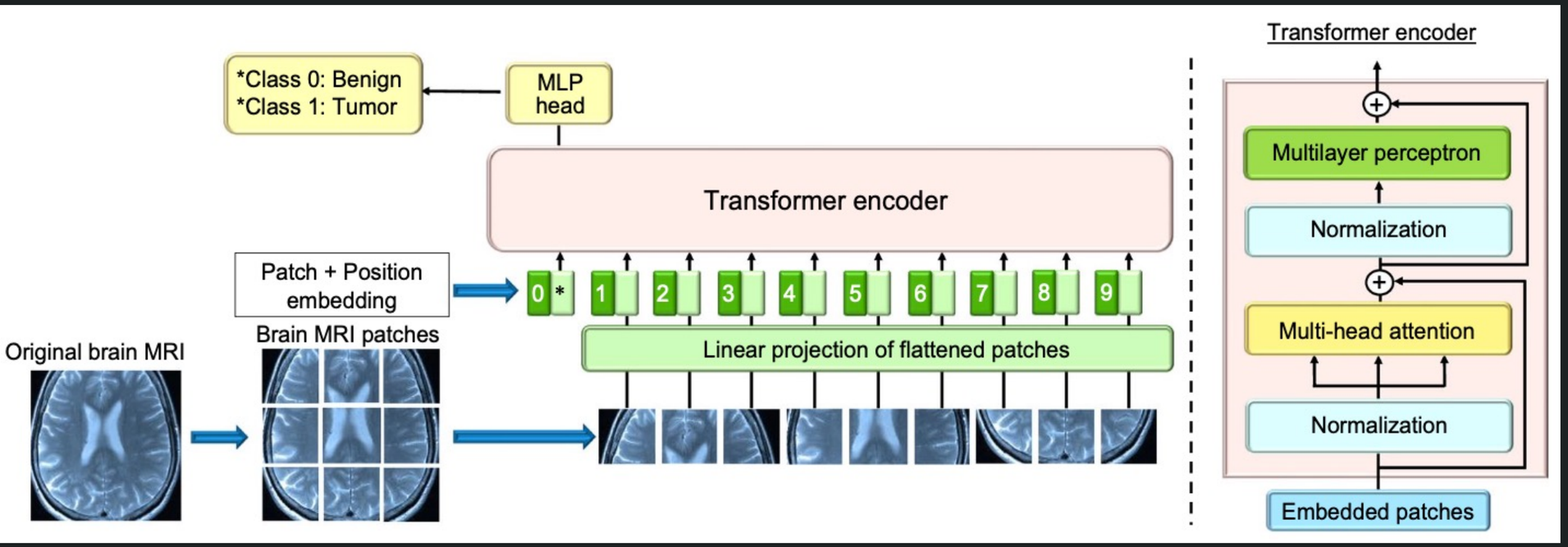


Figure 6. Vision Transformer architecture. An example of image analysis (brain MRI) (Takahashi et al., 2024)

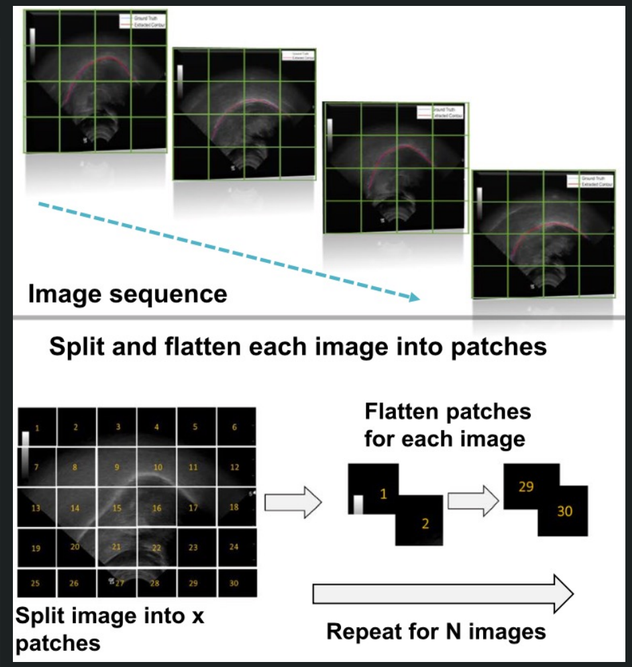


Figure 7. Splitting ultrasound images into patches and flattening them in a linear sequence (Al-hammuri et al., 2023)

ViT vs CNN

Aspect	CNN	Vision Transformer (ViT)	Why It Matters
Core operation	Convolutions + pooling (local receptive fields)	Patch embedding + multi-head self-attention (global)	Drives feature extraction & fusion style
Inductive bias	Strong (translation invariance) → data-efficient	Weak → needs large datasets / pre-training	Impacts training strategy (pre-train vs. scratch)
Context range	Limited; long-range captured only in deeper layers	Global from first layer	Aids segmentation of structures spanning wide areas (e.g., whole organ CT)
Spatial detail	Inherently preserves fine-grained edges and textures	Excels at global context; may require hierarchical or hybrid designs to maintain boundary precision	Sub-pixel-accurate contours are essential for tasks such as tumour-margin delineation
Data / compute	Trains reliably on medium-sized, annotated datasets; typically lighter in parameters and faster at inference	Requires substantial pre-training data and higher GPU memory; self-attention cost grows with image size, leading to longer training and larger models	Determines feasibility in resource-limited clinical settings and informs hardware budgeting / dataset curation
Typical edge	Precise boundaries, faster inference	Better global reasoning; often SOTA when data abundant	Guides architecture choice per task & resources

Application of ViT in Medical Image Analysis

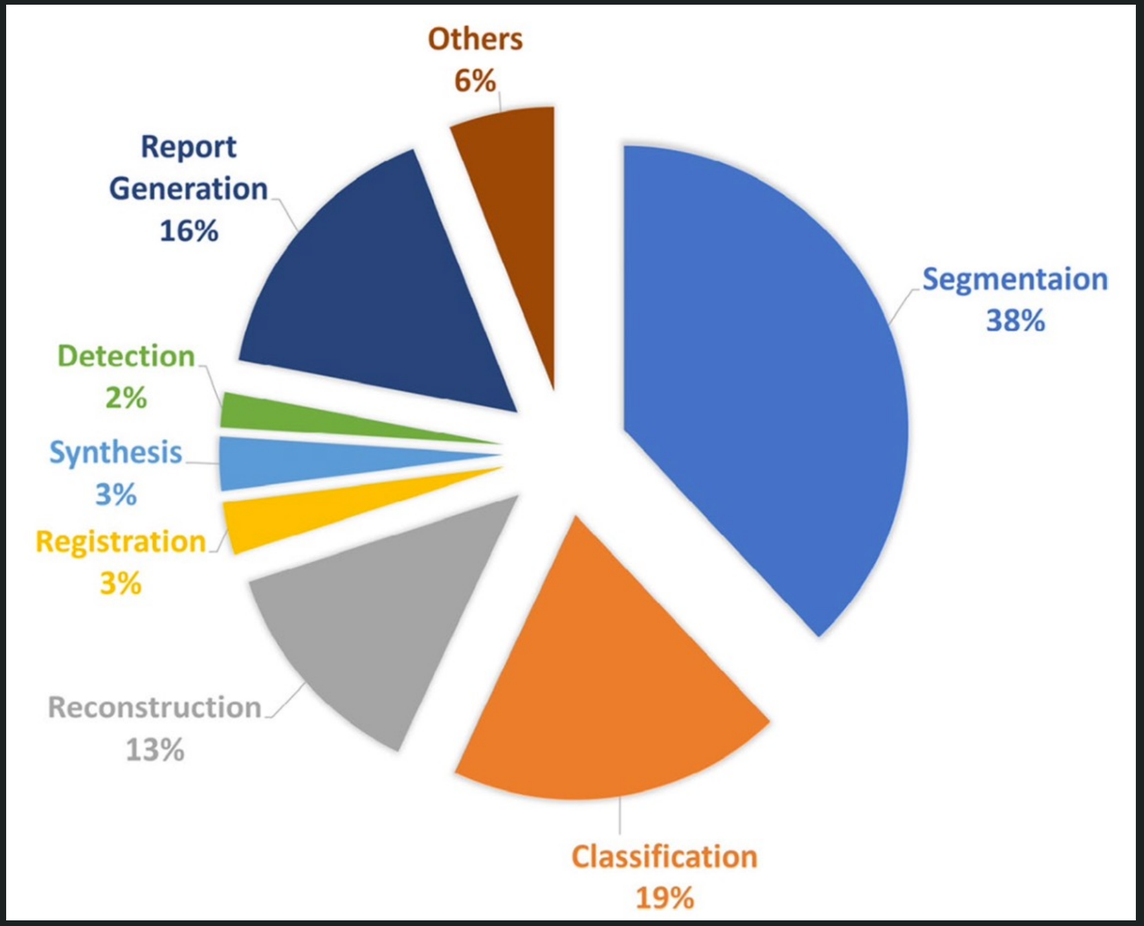


Figure 8. Distribution of medical imaging applications of the ViT (Al-hammuri et al., 2023)

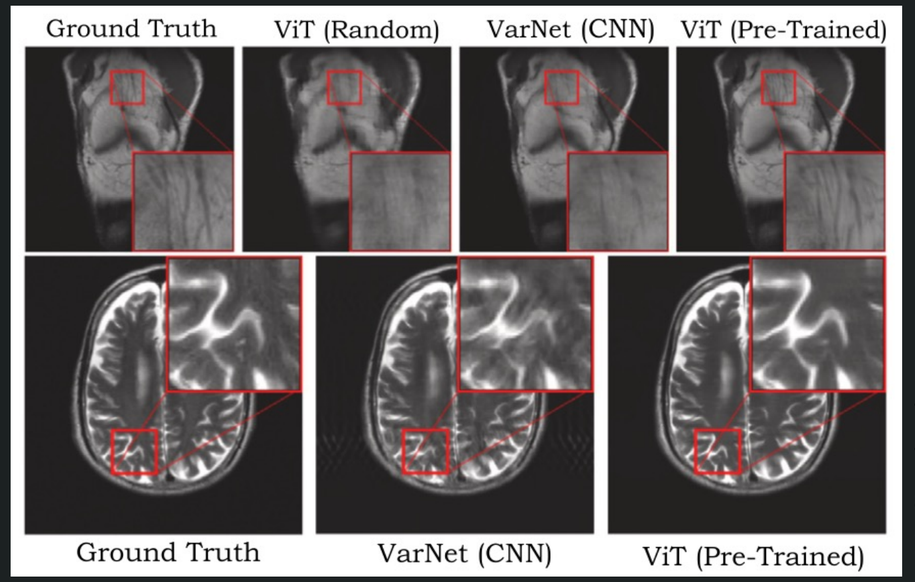


Figure 9. Top row: reconstructions of models trained on 100 images. ViT model pre-trained on ImageNet produce sharp results as compared to recent CNN-based model and randomly initialized ViT. Bottom row: reconstructions of images by models pre-trained on ImageNet and fine-tuned on Knee MRI dataset. Pre-trained ViT models are more robust to anatomical shifts. (Shamshad et al., 2023)

Vits have found broad applications across medical image analysis, ranging from segmentation, classification, reconstruction, registration (aligning images), detection to more complex tasks like image synthesis and even generating textual reports from images.

The medical fields include breast cancer, skin lesions, magnetic resonance imaging brain tumors, lung diseases, retinal and eye analysis, COVID-19, heart diseases, colon cancer, brain disorders, diabetic retinopathy, skin diseases, kidney diseases, lymph node diseases, bone analysis, etc.

Image Segmentation

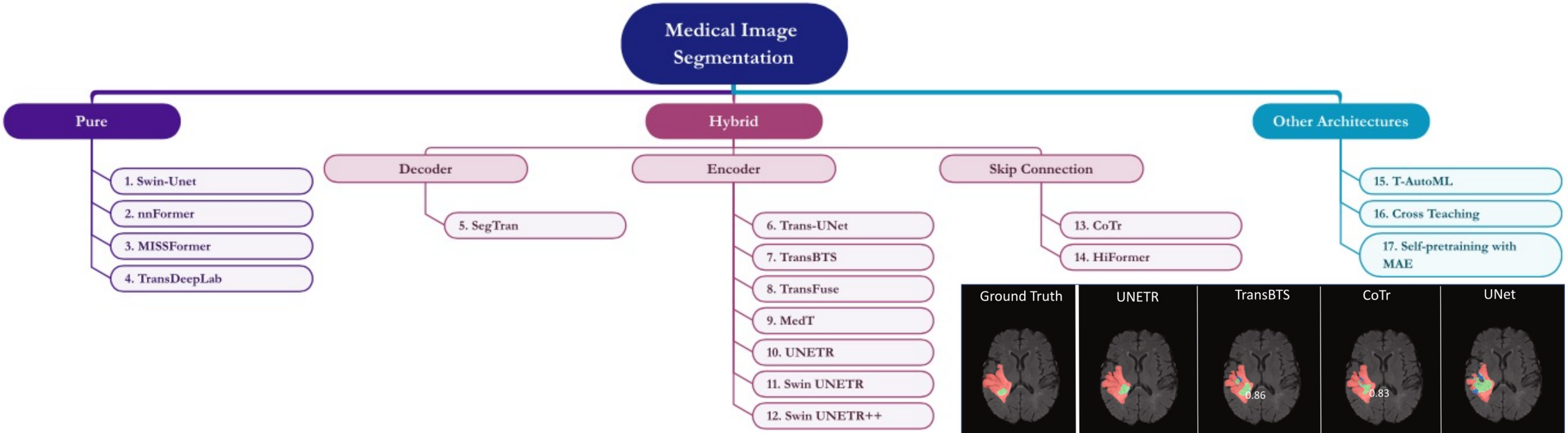


Figure 10. An overview of ViTs in medical image segmentation. Methods are classified into the pure, hybrid, and other architectures according to the positions of the Transformers in the entire architecture. (Azad et al., 2024)

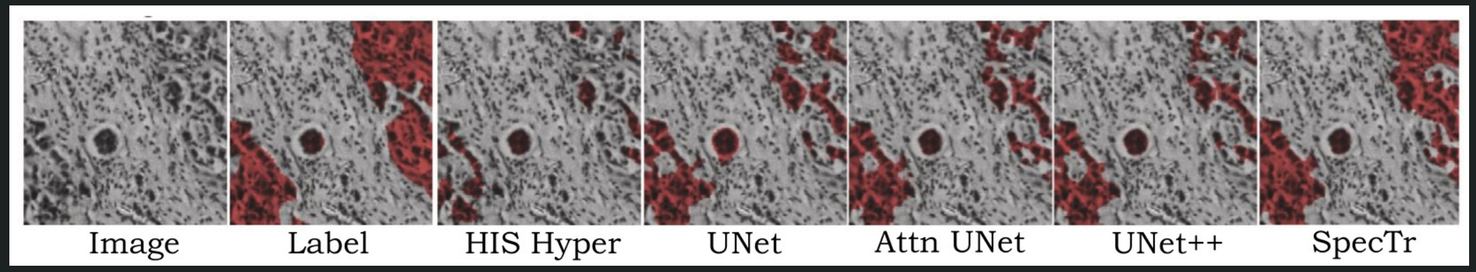


Figure 11. Segmentation results for hyperspectral pathology dataset. From left to right: Input image, Ground truth label, CNN-based (HIS Hyper, UNet, Attn UNet, UNet++), and ViT-based SpecTr. (Shamshad et al., 2023)

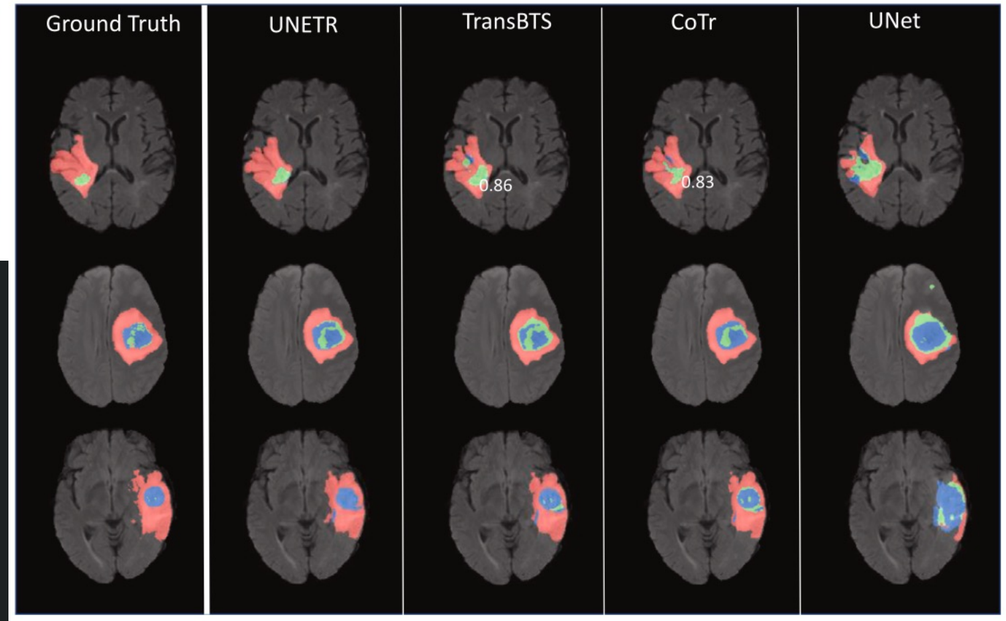


Figure 12. Comparison of visualization of brain tumor segmentation. Note that transformer-based approaches demonstrate better performance in capturing the fine-grained details of brain tumors as compared to CNN-based method (UNet). (Azad et al., 2024)

Image Classification

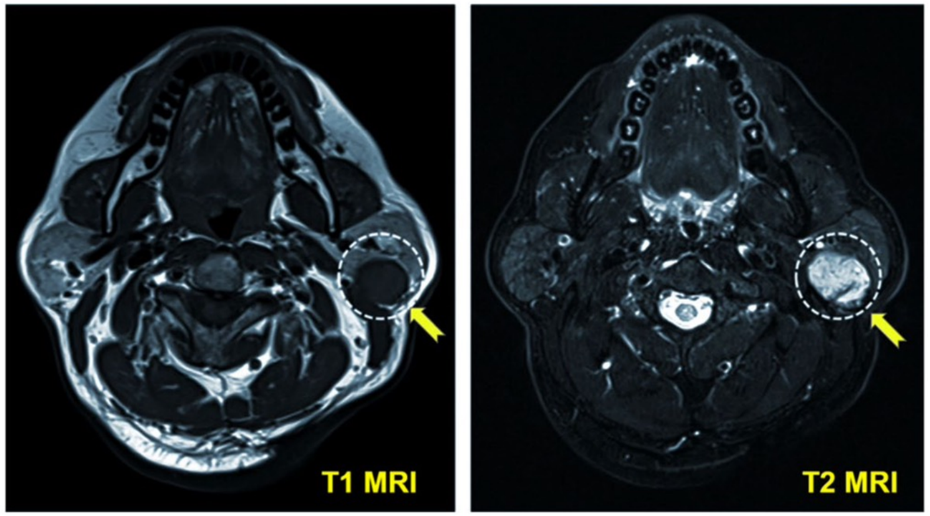


Figure 14. ViT (TransMed) tumor classification in MRI images. The tumor is enclosed by the dashed circle indicated by the yellow arrow. (Al-hammuri et al., 2023)

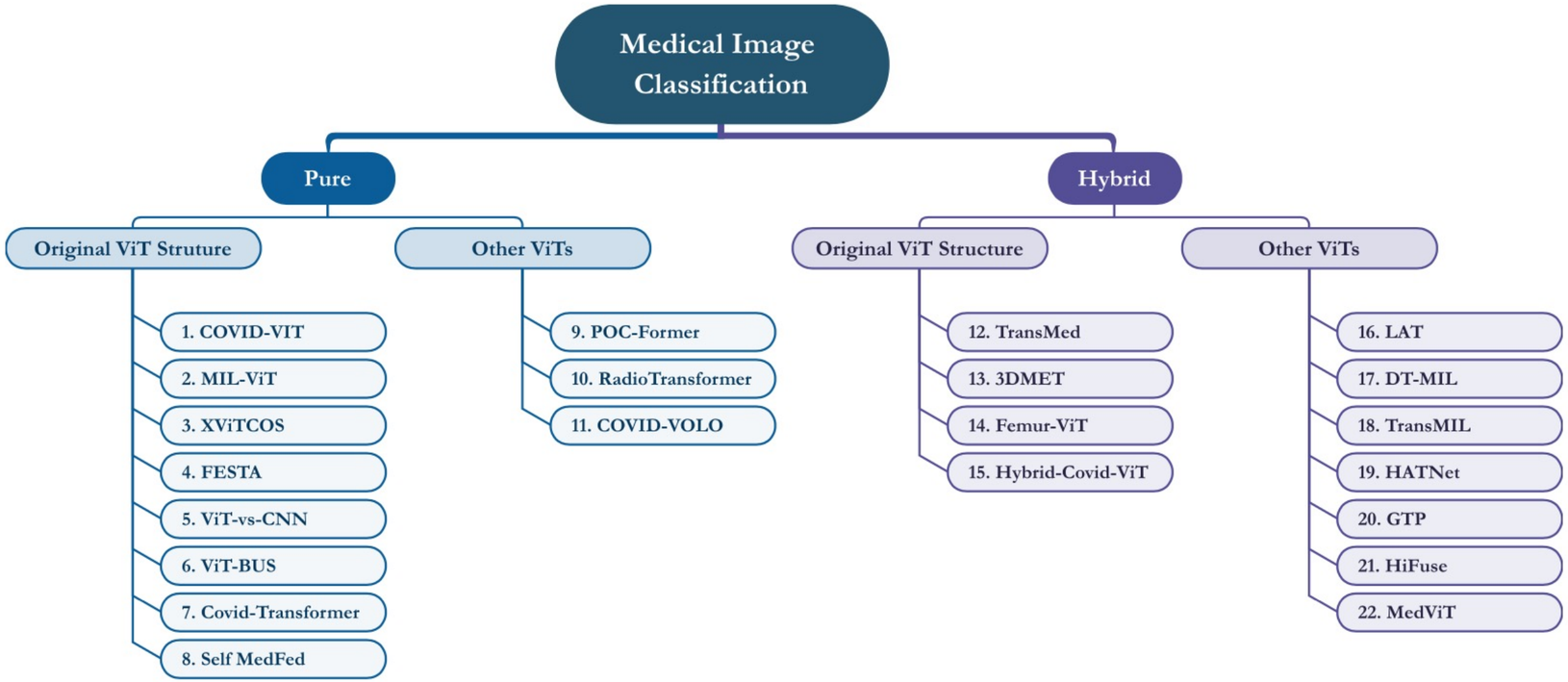


Figure 13. Taxonomy of ViT-based approaches in medical image classification. (Azad et al., 2024)

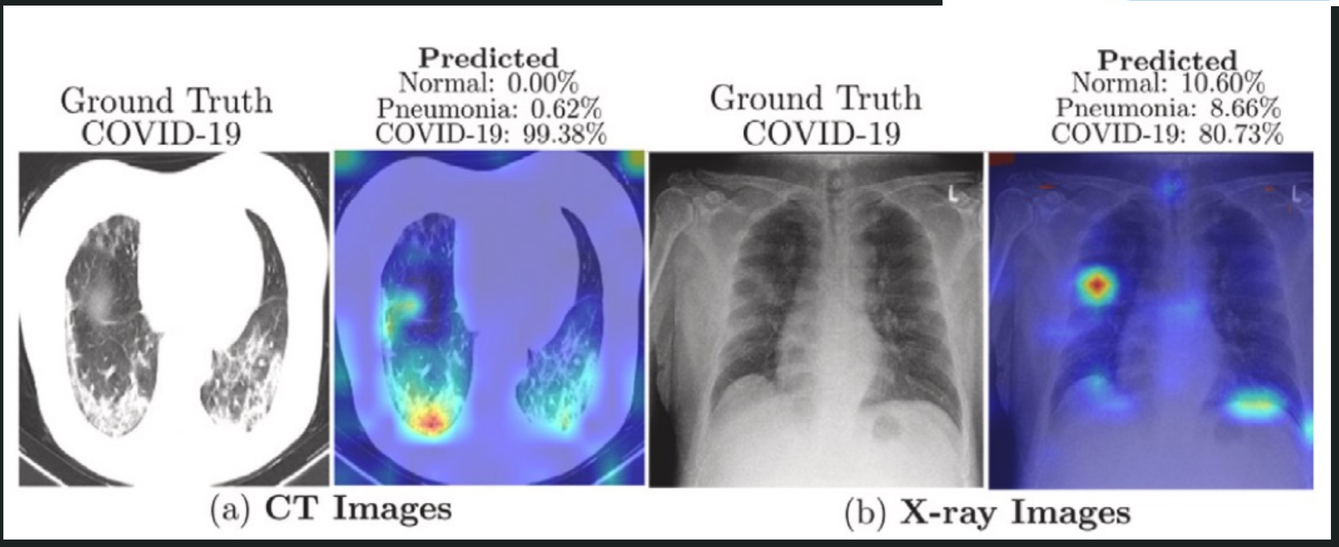


Figure 15. CT (a) and X-ray (b) images along with their ground truth labels (left) and saliency maps (right). (a) xViTCOS-CT localized suspicious lesion regions exhibiting ground glass opacities, consolidation, reticulations in bilateral postero basal lung. (b) thick walled cavity in right middle zone with surrounding consolidation. (Shamshad et al., 2023)

ViT-based hybrid TransUNet

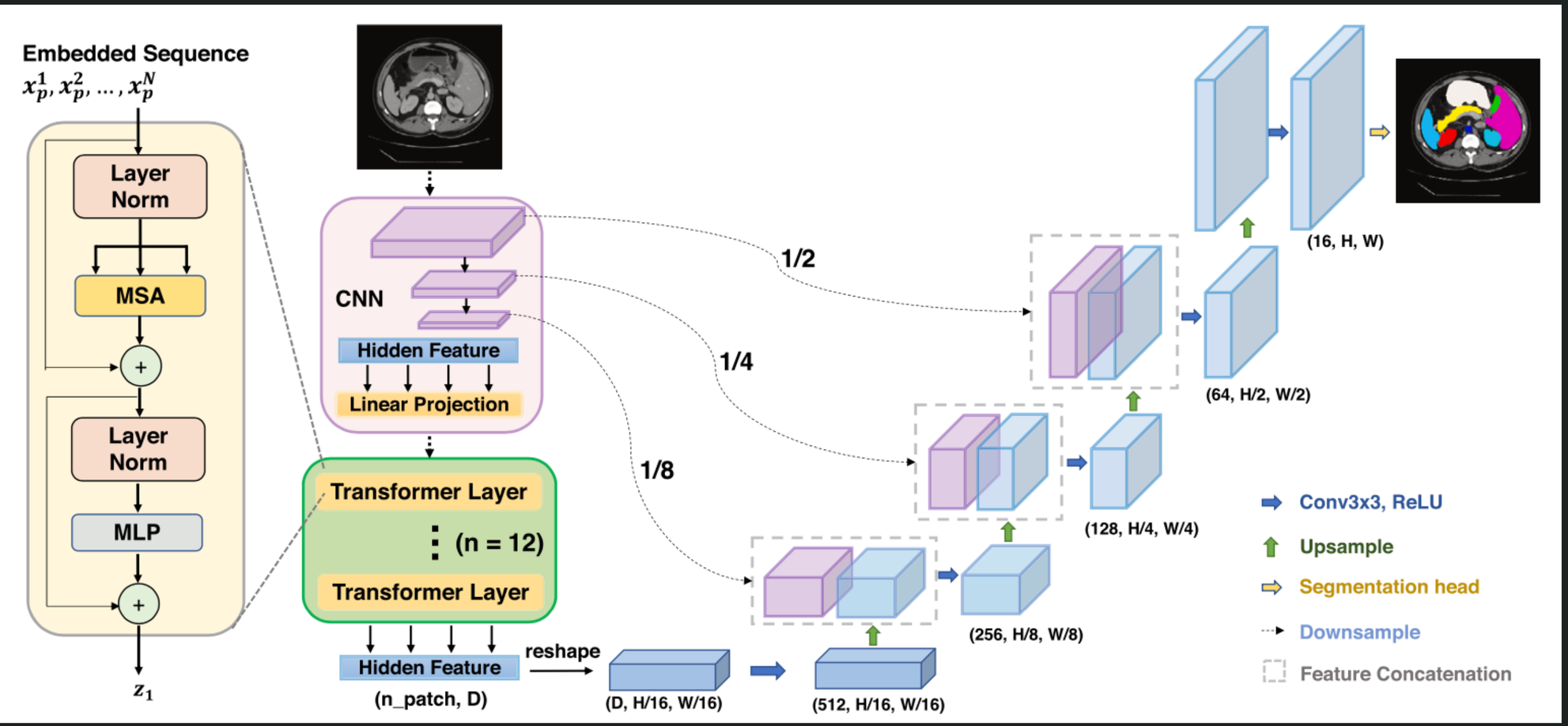


Figure 16. Overview of the framework. (a) schematic of the Transformer layer; (b) architecture of the TransUNet. (Chen et al., 2021)

Hybrid Architecture: TransUNet combines a U-Net (CNN) with a Transformer encoder (ViT) for medical image segmentation
CNN Encoder: Extracts localized, high-resolution feature maps from the input image
Transformer Module: Encodes the CNN features as a sequence of patches, applying self-attention to capture long-range dependencies
Decoder: U-Net style upsampling path merges Transformer outputs with skipped CNN features, yielding precise segmentation
Use Case: Achieved state-of-the-art multi-organ CT segmentation by leveraging both local detail and global context

Advantages and Challenges of ViTs in Medicine

Advantages:

- Global-context awareness
- Transfer-learning flexibility
- Token-based, task-agnostic design
- Hardware-parallel scalability (with efficient attention variants)

Challenges:

- Data Hungry & Training Complexity
- Computational Demands & Memory Cost
- Weak local-detail bias
- Lack of Inductive Bias (Locality)
- Limited interpretability
- Model Size and Overfitting
- Bias and Fairness Concerns

Future Research Directions

- Data-Efficient Training: Investigate semi-supervised, weakly supervised, and self-supervised learning methods to reduce dependency on large labeled datasets.
- Efficient Architectures: Develop lightweight and optimized ViT architectures for deployment in real-world medical applications, particularly in low-resource settings.
- Domain-Specific Adaptations: Explore domain adaptation techniques and create large-scale medical-specific pre-training datasets to bridge the gap between natural and medical image domains.
- Interpretable AI: Integrate explainability techniques to ensure transparency, accountability, and clinical trust in AI-assisted diagnostics.
- Multi-Modal Learning: Combine ViTs with other data sources, such as electronic health records, genomic data, and patient history, to improve diagnostic accuracy and robustness.

References 1

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M. & Hamamoto, R. (2024) 'Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review', Journal of Medical Systems, 48, 84. https://doi.org/10.1007/s10916-024-02105-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017) 'Attention is all you need', Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA.

Al-hammuri, K., Gebali, F., Kanan, A. & Thirumarai Chelvan, I. (2023) 'Vision transformer architecture and applications in digital health: a tutorial and survey', Visual Computing for Industry, Biomedicine, and Art, 6(14). https://doi.org/10.1186/s42492-023-00140-9

Azad, R., Kazerouni, A., Heidari, M., Khodapanah Aghdam, E., Molaei, A., Jia, Y., Jose, A., Roy, R. & Merhof, D. (2024) 'Advances in medical image analysis with vision Transformers: A comprehensive review', Medical Image Analysis, 91, 103000. https://doi.org/10.1016/j.media.2023.103000

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S. & Fu, H. (2023) 'Transformers in medical imaging: A survey', Medical Image Analysis, 88, 102802. https://doi.org/10.1016/j.media.2023.102802

Chen, J., Lu, Y., Lu, Y., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. & Zhou, Y. (2021) 'TransUNet: Transformers make strong encoders for medical image segmentation', arXiv preprint arXiv:2102.04306. Available at: https://arxiv.org/abs/2102.04306

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S. & Fu, H. (2024) 'Transformer for medical image analysis', in Deep Learning for Medical Image Analysis, Elsevier. https://doi.org/10.1016/B978-0-323-85124-4.00012-X

Pan, S., Liu, X., Xie, N. & Chong, Y. (2023) 'EG-TransUNet: a transformer-based U-Net with enhanced and guided models for biomedical image segmentation', BMC Medical Imaging, 23(1). https://doi.org/10.1186/s12880-023-01047-4

Zhang, M., Zhang, Y., Liu, S., Han, Y., Cao, H. & Qiao, B. (2024) 'Dual-attention transformer-based hybrid network for multi-modal medical image segmentation', Scientific Reports, 14, Article number: 25704. https://doi.org/10.1038/s41598-024-25704-9

References 2

Aburass, S., Dorgham, O., Al Shaqsi, J.I., Abu Rumman, M. & Al-Kadi, O. (2025) 'Vision Transformers in Medical Imaging: A Comprehensive Review of Advancements and Applications Across Multiple Diseases', Journal of Imaging Informatics in Medicine. https://doi.org/10.1007/s10278-025-01481-y

Li, J., Yu, H., Chen, C., Ding, M. & Zha, S. (2022) 'Category guided attention network for brain tumor segmentation in MRI', Physics in Medicine & Biology, 67(8), 085014. https://doi.org/10.1088/1361-6560/ac628a

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021) 'An image is worth 16x16 words: Transformers for image recognition at scale', Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), arXiv preprint arXiv:2010.11929. Available at: https://arxiv.org/abs/2010.11929

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A. & Zhou, Y. (2024) 'TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers', Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

Ramadan, H., El Bourakadi, D., Yahyaouy, A. & Tairi, H. (2024) 'Medical image registration in the era of Transformers: A recent review', Informatics in Medicine Unlocked, 49, 101540. https://doi.org/10.1016/j.imu.2024.101540

Zhu, D. & Wang, D. (2023) 'Transformers and their application to medical image processing: A review', Journal of Radiation Research and Applied Sciences, 16(2023), 100680. https://doi.org/10.1016/j.jrras.2023.100680

Raj, R., Mathew, J., Kannath, S.K. & Rajan, J. (2022) 'StrokeViT with AutoML for brain stroke classification', Engineering Applications of Artificial Intelligence, 119, 105772. https://doi.org/10.1016/j.engappai.2022.105772

Oh, S., Kim, N. & Ryu, J. (2024) 'Analyzing to discover origins of CNNs and ViT architectures in medical images', Scientific Reports, 14, Article number: 8755. https://doi.org/10.1038/s41598-024-85769-7

THANK YOU!