Optimizing Deep Learning in Radiologic Image Analysis: Expert Eye Tracking-Derived Model Guidance

F Doolan
Zucker School of Medicine at Hofstra/Northwell Residency in Diagnostic Radiology at Norwalk Hospital

PURPOSE

AI decision support tools are increasingly being investigated as radiological adjuncts. Medical imaging prediction deep learning (DL) models employed for image prediction – chiefly the state-of-the-art Vision Transformer (ViT) and the Convolutional Neural Network (CNN), not only require precision outputs, but also must be explainable and transparent.
Concerning potentially critical patient decision making, the governing bodies are clear:

The EU's GDPR (2018) was the first major initiative to introduce significant, legally binding rights regarding automated decision-making, including a **"right to explanation"** (Art. 15, 22, Recital 71), requiring meaningful information be demonstrated about the logic of machine-led decisions that impact data subjects.

California's updated privacy regulations (CCPA/CPRA), the **"Right to Explanation"** is formally part of a new framework governing **Automated Decision-Making Technology (ADMT)**. Finalized in late 2025, these rules grant consumers the right to access "meaningful information" about the logic, inputs, and intended outcomes of automated systems that make **Significant Decisions** about them.

More comprehensive legal frameworks have been developed since, and the message is clear – **if imaging prediction models are to work with radiologists in the future, they must be accurate, robust, and transparent.**

- Eye tracking and gaze (ETAG) data of experts in the training phase of imaging prediction models has strong potential to bridge the explainability gap encountered in model development today.
- ETAG data of human experts identifies features of a radiologic study of importance in reaching the diagnostic decision, providing a blueprint for machine attentional mechanisms, discussed herein.
- A deep learning model employing such data can be trained not only to provide the correct decision (the primary ground truth) but also to ensure that the internal features it uses for that decision strongly correlate with these human-identified areas of interest (secondary ground truth).
- This heightens explainability, transparency, and accuracy.
- Implementation of ETAG data confronts some interrelated challenges, which are also addressed.

Principal deep learning architectures employed today in medical imaging prediction are the Vision Transformer (ViT) and the Convolutional Neural Network (CNN).

CNNs work by mimicking the visual cortex, learning simple patterns (edges, textures) in early layers and combining them into complex features (shapes, objects) in deeper layers, making them highly efficient for image analysis.

The Vision Transformer (ViT) is the latest in deep neural networks being applied to image prediction models.

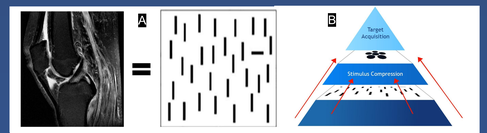

Figure 1: Informational processing in CNNs involves consolidating salient patterns. In A) we see static images can be consolidated to simpler processing patterns, which in B) can be further compressed to make a diagnosis (here, a meniscal root tear).

The Vision Transformer was introduced in 2020 and marked a **significant shift in computer vision**, which had previously been dominated by CNNs. These models treat an image like a sequence of "words" (tokens). Transformer models are already the de facto status quo in Natural Language Processing (NLP). For example, the popular **ChatGPT** AI chatbot is **a transformer-based language model.**

REFERENCES

SCAN ME

FINDINGS

How Vision Transformers Work :

Preprocessing

Patching: An input image is split into a grid of smaller, fixed-size, non-overlapping patches (e.g., 16x16 pixels). The ViT model represents an input image as a series of image patches, like the series of word embeddings used when using transformers to text, and predicts class labels for the image.

Embedding: Each patch is flattened into a 1D vector and then linearly projected into a consistent embedding dimension, similar to how words are converted into token embeddings in NLP.

PositionalEncoding: Because the Transformer architecture is **order-agnostic**, learnable positional encodings are added to the patch embeddings to retain spatial information.

Classification Token: A learnable token is prepended to the sequence of patch embeddings, which serves as the representation for the entire image for classification.

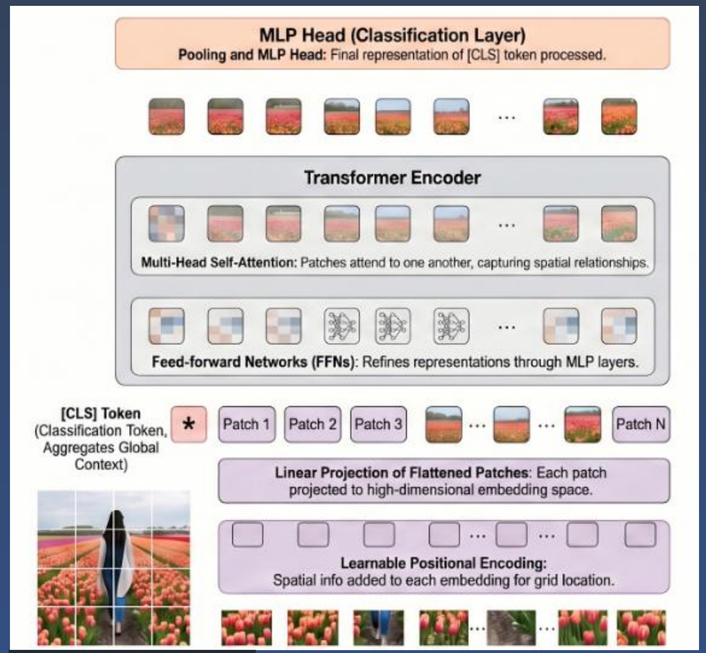

Fig. 2 – Schematic of Vision Transformer Image Processing

Encoder Block:

Self-Attention Weighting: Following patch embedding and positional encoding, the self-attention mechanism allows a ViT model to attend to different regions of the input data, based on their relevance to the task. The self-attention mechanism computes a weighted sum of the input data. The weights are computed based on the similarity between the input features. Self-attention is a computational primitive to quantify pairwise entity interactions that help a network learn hierarchy and alignment present inside input data.

Feed-forward networks (FFNs): A series of linear transformations followed by activation functions. FFN maps the representations produced by the self-attention layers to a lower-dimensional space, more efficient for the prediction task (e.g., image classification).

Pooling Layer: The pooling layer is typically applied after the transformer blocks to reduce the spatial resolution of the feature maps, this helps to reduce the computational complexity and increase the robustness of the model to small translations

Classification Layer: For final prediction in vision transformers. The inputs to the classification layer are the flattened feature maps from after pooling, and the outputs are the class probabilities. Together, the pooling and classification layers make up the last part of a vision transformer's prediction process. They take a picture as input and make a guess about the things that are in it.

! Key Point: While CNNs were built to emulate the biologic properties of differential neurons in the visual system, ViTs substantially diverge from this paradigm, which makes the models **more opaque to post-hoc analysis**.

Deep learning **models learn statistical patterns, while radiologists' gaze tracks clinically meaningful features**. ETAG data quantifies expert attention patterns to be adapted as learning signals for DL models, for example to guide **focus toward clinically significant regions** and **away from artifacts**. This improves accuracy and efficiency (**including energy efficiency**) and **mitigates overfitting. Overfitting is a structural failure where a prediction model "memorizes" specific details and random noise within its training dataset instead of learning the actual clinical signal**. The best known cautionary tale of this is the performance of predictive models in the identification of pneumothorax on chest x-rays.

- **The Shortcut**: Patients with a known pneumothorax often already have a **chest tube** inserted to treat it. A ViT can learn that the presence of a tube = "Positive for Pneumothorax".
- **Result**: model fails in real-world "de novo" cases where the patient has a ptx but no tube has been inserted yet—the moment when the AI would be most useful.

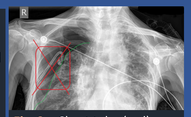

Fig. 3 – Chest tube (red) marked as an incorrect region of interest. The pleural line should be the image patch which leads the ViT to the right diagnosis.

FINDINGS

Eye-tracking data limits overfitting in Vision Transformer (ViT) diagnostic models by acting as a **regularizer,** guiding the model's attention toward relevant features rather than noisy, spurious background correlations, which can consume unnecessary computational time and energy.

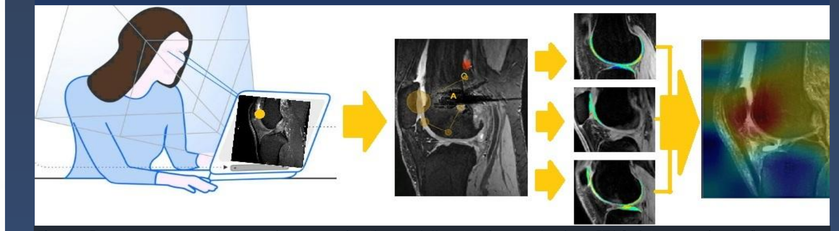

Fig. 4- Expert gaze can be leveraged to supervise the model's attention mechanism. ETAG data comprises a few primary spatial and temporal metrics: **gaze location** (x,y,z coordinates), **dwell time (**time spent on an area), **saccades** (rapid movements between points), **scan paths** (the temporal order of search sequences). T2- weighted sagittal knee MRI.
In this example, a filtered dataset demonstrates the order of fixations (connected by lines , length proportional to milliseconds), beginning with the femoral condyles at the ACL and PCL attachments. The dwell time is represented proportionally by the area of the circles.
A function is used to **transform raw gaze information into a normalized probability distribution that highlights critical image regions for the model**. In this case, the patellar region would receive an additional weighting factor due to the increased dwell time.
The model behaviour is shaped by lean, salient search patterns that afford appropriate analysis to key structures in each plane, and **appropriately neglect regions** (such as artifact **A**, here in the femur from **cruciate ligament repair)** that would otherwise have confounded the model.

Incorporating expert fixations has shown measurable improvements in diagnostic accuracy. For example, gaze-supervised models have improved **F1 scores** by approximately 5.7% and **AUC** by 3.4% in chest X-ray classification tasks compared to image-only models [3].
Present challenges to application of ETAG data to DL models include scalability and data nuances, such as that **dwell time sometimes reflects a pruned scanpath**, explained below:

- The expert's gaze has unique characteristics that can confound implementation in the training stage of a model. One key characteristic, **'Loss of "Negative Evidence,"'** states experts often dismiss healthy areas with a mere peripheral glance, hence, a 'pruned scanpath'. Data smoothing may remove these brief fixations, preventing the AI from learning the exhaustive peripheral gaze patterns needed to comprehensively scan an image.
- Moreover, raw eye-tracking data is noisy, thus large-scale data acquisition is required to achieve adequate power in datasets.

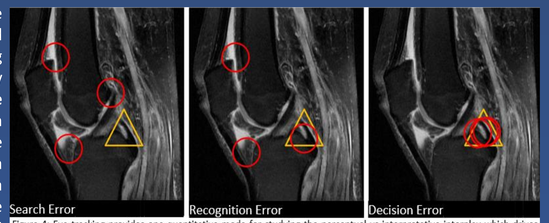

Figure 4: Eye tracking provides one quantitative mode for studying the perceptual vs interpretative interplay which drives radiologists' decisions. Red circles represent fixation and yellow triangle represents the abnormality.
Image interpretation studies have highlighted three ways that a lesion or 'target' might be missed:
(A) In a search error, the target is never fixated.
(B) A recognition error is when the eyes fixate on the target briefly and then move on, with no indication that the reader noted anything of interest.
(C) Multiple and/or long fixations on a target indicate a decision error, if the radiologist still fails to report the finding. This pattern indicates that, implicitly or explicitly, the interpreter knew that this spot deserved attention but then made the incorrect decision and did not flag as abnormal.
This is a T2-weighted sagittal MRI of a medial meniscal root tear.
Search errors represent one reason eye-tracking data can be 'noisy' – it must be filtered to eliminate errors in gaze and collected at scale to minimize sample biases.

CONCLUSION

Integrating expert eye tracking and gaze data during training of radiology DL models can enhance robustness, foster data-efficient learning, build transparent decision-making, and improve accuracy. Moreover, by aligning machine reasoning with clinical knowledge, ETAG data integration transforms 'black box' models into more transparent systems that mirror human expert focus.