DINOv3-MIL: Per-Kidney Multi-Label Tumour and Cyst Detection from Foundation-Model Patch Tokens on KiTS23

Vishalakshi M[1], Sahil Sharma[2*], Pramod Kumar P[1]

[1]Department of Computer Science and Artificial Intelligence, SR University, Warangal, India

[2]School of Computing, Ulster University, Belfast, United Kingdom

*** Correspondence:**
Corresponding Author
s.sharma@ulster.ac.uk

Keywords: foundation models, multiple instance learning, prototype networks, renal CT, interpretable ML

Abstract

Foundation vision models trained on natural images transfer to medical tasks without domain pre-training, but volumetric classification requires aggregating tens of thousands of patch tokens per study, and the aggregator constrains how the resulting model can be interpreted. We compare three aggregators on identical frozen DINOv3 ViT-H/16+ features for renal tumour/cyst detection on KiTS23 (966 kidneys; n=97 test): a CLS-token linear probe, gated attention multiple instance learning (MIL) over 55,296 patch tokens, and a prototype head following ProtoViT. Attention MIL achieves the highest AUROC for tumour (0.74, 95% CI 0.64–0.83) and cyst (0.80, 0.70–0.88), with attention enriched 7.5–9.8$\times$ over chance within annotated lesions. The prototype head does not transfer to cyst detection (AUROC 0.51), exposing an interpretability–performance trade-off at this token scale.

1 Introduction

Renal CT is routine in oncological work-up, and deployed classifiers are typically explained by post-hoc saliency methods whose faithfulness has been challenged (Adebayo et al., 2018). For foundation-model patch features on volumetric CT, the aggregator choice including global pooling, attention, or prototype matching determines whether the resulting classifier is interpretable, whether it can detect small lesions, and whether the two goals trade off. Applying a ViT to 2.5D axial slices of a kidney crop produces $\sim$55K patch tokens per kidney; we ask which aggregator delivers on per-kidney multi-label tumour and cyst detection.

Self-supervised vision foundation models now produce dense patch features that transfer to medical imaging without domain pre-training (DINOv3; Siméoni et al., 2025). Attention-based MIL (Ilse et al., 2018) aggregates patch features for whole-slide and volumetric classification with post-hoc interpretability via attention weights. Prototype-based networks (Chen et al., 2019; Ma et al., 2024) offer intrinsic case-based interpretability (Rudin, 2019). Classification work on KiTS23 has been confined to mass-level binary tumour-vs-cyst given a known lesion (Ortlieb, 2026); the per-kidney multi-label setting addressed here has not been studied.

This work makes two contributions. (i) A per-kidney multi-label tumour/cyst task on KiTS23 including healthy kidneys. (ii) A matched comparison of three aggregators on frozen DINOv3 patch features: linear probe, attention MIL, ProtoViT-style head.

2 Methods

From 489 KiTS23 cases (Heller et al., 2023), kidneys were extracted by connected components, resampled to 1 mm isotropic, intensity-windowed to HU $[-100, 400]$, and normalised, yielding 966 crops of $192\times192\times96$ voxels with multi-label (tumour, cyst) at prevalence 0.59 and 0.37. An 80/10/10 case-level split (seed 23) gave 771/98/97 kidneys. Each kidney's 96 axial slices were converted to 2.5D triplets, resized to 384×384, and passed through frozen DINOv3 ViT-H/16+; we cache 576 patch tokens per slice (bf16) (Figure 1).

Three heads share the cache. The linear probe mean-pools the 96 CLS tokens (lr 1e-3, batch 32, 100 epochs). Gated attention MIL (Ilse et al., 2018) attention-pools 96×576 = 55,296 tokens via tanh-sigmoid streams (hidden 256, dropout 0.2; lr 5e-4, batch 4, 30 epochs with cosine annealing). The ProtoViT head (Ma et al., 2024) uses 10 prototypes (5/class) with cosine-similarity max-pooling, a class-constrained classifier, and clustering/separation losses weighted 0.8/0.08 (Chen et al., 2019), trained in three stages (5+15+5 epochs). All heads use AdamW (weight decay 1e-4) and BCEWithLogitsLoss with positive-class weighting; seed 23; thresholds tuned by Youden's J; 1000-resample bootstrap CIs.

3 Results

Table 1 reports test performance. Gated attention MIL achieves the highest AUROC for both labels (tumour 0.74, cyst 0.80), with 95% confidence intervals well above chance. The linear probe approaches MIL on tumour (0.67) but does not transfer to cyst (0.55, CI lower bound 0.42); mean-pooling discards the small-volume signal that spatial attention recovers. The ProtoViT head matches the linear probe on tumour (0.71) but did not exceed chance on cyst (0.51, CI crossing 0.5), despite operating on identical features.

Table 1. Test-set AUROC with 95% bootstrap CIs (n=97 kidneys); F1 at validation-tuned thresholds.

Method	Tumour AUROC	Cyst AUROC	Tumour F1	Cyst F1
Linear probe (CLS)	0.67 [0.56, 0.77]	0.55 [0.42, 0.66]	0.73	0.26
ProtoViT head	0.71 [0.61, 0.81]	0.51 [0.39, 0.62]	0.54	0.52
Gated attention MIL	0.74 [0.64, 0.83]	0.80 [0.70, 0.88]	0.70	0.66

Figure 2 shows attention overlays on three example kidneys. For the tumour case, attention concentrates inside the radiologist-annotated tumour contour (93.3% of total attention mass inside lesion cells versus 12.4% expected by chance, 7.5× enrichment). For the cyst case, 12.4% of attention lies inside cyst cells against 1.3% expected by chance (9.8× enrichment). For the healthy kidney, attention is distributed across slices with no peak exceeding 8% of total mass with no false concentration on the kidney.

4 Discussion

On identical frozen DINOv3 features, gated attention MIL outperforms both mean-pooling and ProtoViT-style prototype learning on per-kidney tumour and cyst detection. The largest gap is on cysts, where attention MIL gains 0.25 AUROC over mean-pooling and 0.29 over the prototype head; we attribute

this to spatial reasoning over a large patch budget (55,296 tokens per kidney), which neither global pooling nor max-pool-then-prototype efficiently exploits for small-volume lesions. We note that attention maps localise to radiologist-annotated regions but should not be read as causal explanations of the classifier's decision (Adebayo et al., 2018). Limitations include the small held-out set (n=97), a single split, validation-set threshold tuning, and early stopping of both trainable heads. Whether intrinsic prototype interpretability can be recovered at this token scale via relaxed classifier constraints or alternative aggregation remains open.

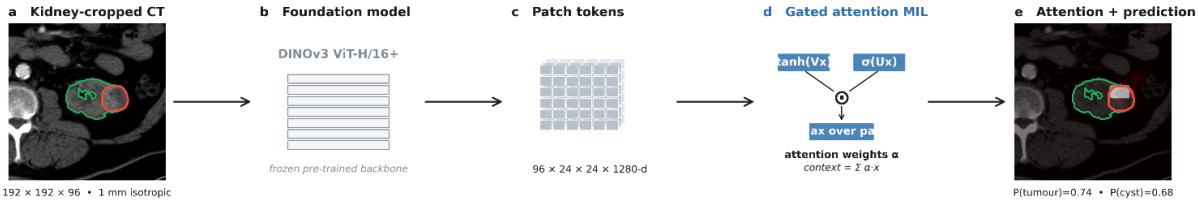

Figure 1. Pipeline. (a) Kidney-cropped CT (192×192×96 voxels, 1 mm isotropic) with kidney (green) and tumour (red) annotations. (b) Frozen DINOv3 ViT-H/16+ encodes 2.5D axial triplets. (c) Cached patch tokens of size 96×24×24×1280-d. (d) Gated attention MIL: tanh and sigmoid projections produce attention weights via softmax over 55,296 tokens. (e) Attention-pooled context drives prediction.

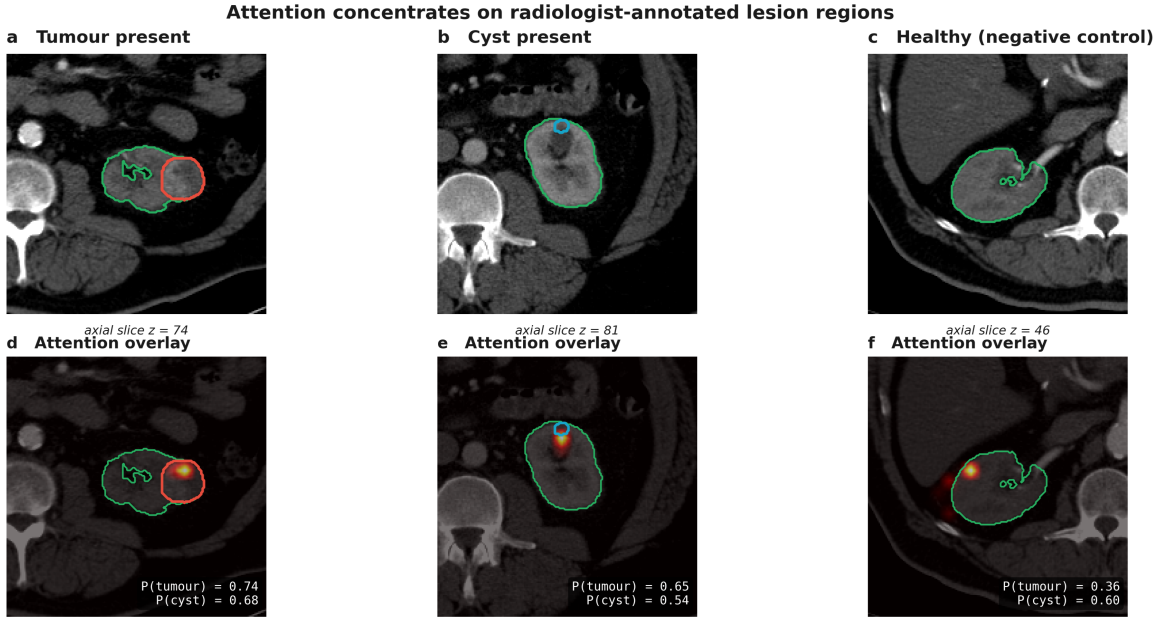

Figure 2. Attention overlays on three test kidneys. Top row (a–c): CT slice with annotations. Bottom row (d–f): attention heatmap overlay. Tumour case (z=74): attention concentrates on tumour (red). Cyst case (z=81, within GT span 81–95): attention concentrates near cyst boundary (cyan). Healthy case (z=46): attention diffuses across kidney slices. Quantitative attention–lesion overlap is computed over the full 3D volume.

5 References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 31, 9505–9515. doi: 10.5555/3327546.3327621

Chen, C., Li, O., Tao, C., Barnett, A., Su, J. K., and Rudin, C. (2019). This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems* (Curran Associates, Inc.), vol. 32, 8930–8941. doi: 10.5555/3454287.3455088

[Dataset] Heller, N., Wood, A., Isensee, F., et al. (2023). The 2023 kidney and kidney tumor segmentation challenge (kits23). https://kits-challenge.org/kits23/

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning* (PMLR), vol. 80 of *Proceedings of Machine Learning Research*, 2127–2136. https://proceedings.mlr.press/v80/ilse18a.html

Ma, C., Donnelly, J., Liu, W., Vosoughi, S., Rudin, C., and Chen, C. (2024). Interpretable image classification with adaptive prototype-based vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 37. doi: 10.5555/3737916.3739228

Ortlieb, O. (2026). Optimizing deep learning for malignant renal tumor vs. cyst classification in ct: rethinking class imbalance and voting strategies. In *Medical Imaging 2026: Clinical and Biomedical Imaging, Proc. SPIE*. vol. 13929, 1392908. doi: 10.1117/12.3089994

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., et al. (2025). DINOv3. *arXiv preprint arXiv:2508.10104* doi: 10.48550/arXiv.2508.10104

6 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Author Contributions

VM conceived the study, implemented the models, conducted the experiments, analysed the results, and wrote the manuscript. SS supervised the implementation, contributed to the narrative, and edited the final manuscript. PKP supervised the overall work. All authors reviewed and approved the final manuscript.

8 Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.