Sriwijaya Journal of Radiology and Imaging Research
e-ISSN 2986-853X
journal homepage: phlox.or.id/index.php/sjrir

Original Research Article

Vision Transformer Reconstruction for Super-Resolution and Artifact Reduction in 1.5-Tesla Fast-Spin-Echo Pelvic MRI: A Diagnostic-Accuracy Study

Muhammad Rusli[1*], Febria Suryani[2], Desiree Montesinos[3]

[1]*Department of Radiology, Al Bayan Medical Center, Kendari, Indonesia*
[2]*Department of Health Sciences, CMHC Research Center, Palembang, Indonesia*
[3]*Department of Women and Child Welfare, Lira State Hospital, Lira, Uganda*

ARTICLE INFO

Article history:
Received: February 2, 2026
Accepted: March 29, 2026
Published: April 30, 2026

Reporting guideline:
STARD 2015 (diagnostic accuracy)

Keywords:
Deep learning
Diagnostic imaging
Image processing, computer-assisted
Magnetic resonance imaging
Pelvis

***Corresponding author:**
Muhammad Rusli
E-mail: mrusli@phlox.or.id
ORCID: 0009-0000-7313-5596

DOI:
https://doi.org/10.59345/sjrir/v4i1.322

ABSTRACT

Introduction: Fast-spin-echo (FSE) MRI is the reference modality for pelvic evaluation, but high-resolution acquisition is slow and motion-prone. Deep-learning reconstruction may recover diagnostic quality from short, motion-tolerant acquisitions, yet most evidence relies on image-similarity indices rather than radiologist performance. We validated a Vision-Transformer (ViT) reconstruction for simultaneous super-resolution and motion-artifact reduction in 1.5-Tesla T2-weighted FSE pelvic MRI.

Methods: In this retrospective diagnostic-accuracy study (STARD 2015) at a tertiary hospital in Palembang, Indonesia, 450 examinations (development n=360; test n=90) were analysed. A U-shaped shifted-window ViT reconstructed high-resolution images from retrospectively degraded low-resolution/motion-corrupted inputs. Two blinded radiologists scored each test case under native low-resolution, UNet- and ViT-reconstructed conditions against a histopathology/expert-consensus reference standard. Sensitivity, specificity, predictive values, AUC, likelihood ratios (95% CI), inter-reader kappa, DeLong and McNemar tests, and multivariable logistic regression were computed.

Results: Target prevalence was 53.3%. ViT reconstruction achieved sensitivity 93.8% (95% CI 83.2–97.9), specificity 88.1% (75.0–94.8), AUC 0.943 (0.894–0.992), LR+ 7.87 and LR− 0.071, versus AUC 0.881 (UNet) and 0.751 (low-resolution); ViT vs low-resolution DeLong p<0.001, McNemar p<0.001. Inter-reader agreement rose from kappa 0.49 to 0.87. ViT gave the best fidelity (PSNR 34.82 dB; SSIM 0.941; p<0.001 vs UNet) at 0.15 s/slice. Sub-centimetre lesions (OR 3.84, p=0.006) and severe motion (OR 2.97, p=0.029) independently predicted error.

Conclusion: A shifted-window Vision Transformer recovered diagnostic-quality pelvic FSE MRI from short, motion-tolerant acquisitions, significantly improving radiologist lesion detection and inter-reader agreement over convolutional reconstruction. The real-time, PACS-compatible pipeline is promising for high-throughput pelvic MRI and warrants prospective validation.

All authors have reviewed and approved the final version of the manuscript.

1. Introduction

Magnetic resonance imaging (MRI) with T2-weighted fast-spin-echo (FSE) sequences is the cornerstone of pelvic evaluation, providing the soft-tissue contrast required to stage prostate, uterine, cervical and rectal disease.[1–3] Diagnostic accuracy in the pelvis depends critically on spatial resolution: tumour boundaries, the muscularis propria of the rectal wall, and the junctional zone of the uterus are all sub-millimetre interfaces. Achieving such resolution, however, demands long acquisition times during which bowel peristalsis, respiration and patient movement introduce ghosting, blurring and phase-encoding artifacts that degrade or invalidate the study.[4,5]

These artifacts are not merely cosmetic. Motion-induced blurring of the rectal wall can obscure the muscularis propria interface that distinguishes a T2 from a T3 rectal tumour; ghosting across the prostate can simulate or mask an extracapsular extension; and aliasing over the adnexa can hide small cystic or solid components that alter gynaecologic management. When image quality is insufficient, the practical consequence is a repeat examination, a supplementary sequence, or a hedged report — each of which consumes scanner time and delays definitive care. The clinical value of any acceleration strategy therefore rests not on how fast it scans but on whether radiologists can still make confident, correct decisions on the resulting images.

Conventional responses to this trade-off are imperfect. Reducing acquisition time by undersampling k-space introduces aliasing and lowers signal-to-noise, while parallel imaging and compressed sensing plateau at modest acceleration and can themselves create reconstruction artifacts. Classical interpolation cannot restore genuinely unsampled high-frequency detail. Deep learning reframes the problem as a learned inverse mapping from degraded to high-quality images, and convolutional architectures such as SRCNN and U-Net first

demonstrated its feasibility.[6,7] Physics-driven and end-to-end variational networks, trained on open k-space datasets, are now strong baselines for accelerated MRI.[8–10]

Convolutional networks nevertheless aggregate context through local receptive fields, limiting their ability to resolve ambiguous boundaries where tissues share signal intensity — free fluid versus mucosa, or tumour versus adjacent normal wall on T2 weighting. Vision transformers instead model long-range dependencies through self-attention, and the shifted-window (Swin) design restores linear computational complexity and a hierarchical representation suited to image restoration.[10,11] Transformer reconstruction of MRI — SwinMR, zero-shot adversarial transformers, recurrent and multimodal transformers, and joint reconstruction–super-resolution networks — consistently improves fidelity over convolutional baselines,[5,9,12–15] a maturation documented in recent surveys and reviews.[5,11,16]

In Indonesia, as in many low- and middle-income settings, pelvic MRI capacity is constrained by scanner availability and throughput; motion-degraded repeat examinations and long queues translate directly into delayed oncologic diagnosis.[17,18] A reconstruction that recovers diagnostic quality from short, motion-tolerant acquisitions would therefore have disproportionate operational value in tertiary referral centres serving large, dispersed populations. Yet almost no validation of transformer reconstruction originates from such settings, and the broader imaging-AI literature is dominated by surrogate image-similarity metrics rather than radiologist diagnostic performance.[19]

We addressed these gaps by developing a U-shaped shifted-window Vision Transformer that simultaneously performs super-resolution and motion-artifact reduction on 1.5-Tesla T2-weighted FSE pelvic MRI, and — rather than stopping at fidelity indices — by validating it within a STARD-compliant diagnostic-accuracy framework. Working at a tertiary hospital in Palembang, Indonesia, we quantified radiologist lesion-detection performance and inter-reader agreement on ViT-reconstructed images against an uncorrected low-resolution condition and a convolutional (UNet) reconstruction, using a blinded histopathology-and-expert-consensus reference standard. We hypothesised that ViT reconstruction would significantly improve diagnostic accuracy, likelihood ratios and inter-reader reliability while preserving real-time applicability.

2. Methods

This study was conducted and reported in accordance with the STARD 2015 guidelines for diagnostic-accuracy studies,[20] complemented by the CLAIM checklist and Radiology editorial guidance for imaging artificial intelligence.[20]

2.1. Study design and setting

We performed a single-centre retrospective diagnostic-accuracy study using de-identified pelvic MRI examinations acquired at a tertiary (referral) hospital in Palembang, South Sumatra, Indonesia. The specific institution is withheld under a binding non-disclosure agreement and is referred to throughout as a tertiary hospital in Palembang. The study comprised a model-development phase and an independent diagnostic-validation phase on a temporally separated test set.

2.2. Participants and eligibility

Consecutive adults (≥18 years) who underwent clinically indicated T2-weighted FSE pelvic MRI between January 2024 and March 2026 were eligible. Inclusion required a complete diagnostic-quality high-resolution T2 acquisition and an available reference standard. Exclusion criteria were heavy metallic implants causing extreme susceptibility artifact, incomplete sequences, and prior pelvic radiotherapy precluding reference-standard adjudication. Of 450 eligible examinations, 360 were assigned to development and 90 to the independent test set; assignment was temporal, with the most recent consecutive examinations forming the test set to emulate prospective deployment.

2.3. Sample-size considerations

The test-set size was driven by the primary outcome, sensitivity. Anticipating sensitivity of approximately 0.92 with a target half-width of 0.10 at 95% confidence and an expected target-condition prevalence near 0.50, a minimum of approximately 84 test cases was required; 90 consecutive examinations were therefore included, providing power ≥0.80 to detect an AUC difference of 0.10 between reconstruction conditions at a two-sided alpha of 0.05.

2.4. MRI acquisition protocol

All examinations were acquired on a 1.5-Tesla MRI scanner (vendor and model anonymised) using a phased-array body coil. Diagnostic high-resolution (HR) T2-weighted 2D FSE images were obtained in standard orthogonal planes with the following parameters: repetition time (TR) 3500–4500 ms; echo time (TE) 85–110 ms; echo train length 16–24; field of view 200–240 mm; acquisition matrix 320×256 reconstructed to 512; slice thickness 3 mm with 0.3 mm gap; receiver bandwidth approximately 200 Hz/pixel; 2–3 signal averages. No intravenous contrast was used for the reconstruction task. In total 18,500 HR T2 FSE slices were curated as ground truth.

Low-resolution (LR), motion-corrupted inputs were generated retrospectively from the HR data through a controlled k-space degradation model: Cartesian undersampling to an effective 6 mm-equivalent through-plane resolution, addition of simulated rigid and pseudo-periodic phase errors to reproduce respiratory and peristaltic ghosting, and complex Gaussian noise. This paired HR–LR design provided a reference for both fidelity and reader analyses while avoiding additional patient scanning.[8,21]

2.5. Vision Transformer architecture and training

The proposed network adopted a U-shaped encoder–decoder built from hierarchical Swin Transformer blocks. Input LR images were partitioned into non-overlapping patches; a four-stage encoder applied window-based and shifted-window multi-head self-attention to capture local and long-range anatomical dependencies at linear complexity; a symmetric decoder with skip connections

and a PixelShuffle module performed super-resolution upsampling to the HR grid.[10,11,14] The model was trained with a composite objective combining an L1 pixel-fidelity term and a VGG-based perceptual term (weights 1.0 and 0.1, respectively) to preserve both signal accuracy and fine texture. Optimisation used Adam (initial learning rate 2×10^{-4}, cosine decay) for 200 epochs with data augmentation; UNet and SRCNN baselines were trained identically for fair comparison.[6,7] All development used the 360-examination set with patient-level splitting to prevent leakage; the 90-examination test set was untouched during training and tuning.

2.6. Reference standard

The reference standard for the binary diagnostic target — clinically significant pelvic pathology (biopsy-proven malignancy, or a benign target lesion requiring intervention or active follow-up) — was histopathology where tissue was available (transrectal/transperineal prostate biopsy, surgical specimens, or image-guided biopsy). Where histology was not obtained, the standard was blinded expert consensus of two independent radiologists reading the native HR images together with a minimum 6-month clinical and imaging follow-up. Reference-standard adjudicators were independent of, and blinded to, the reconstruction reader study.

2.7. Image interpretation and blinding

Two board-certified radiologists with 8 and 12 years of pelvic MRI experience independently interpreted each test case under three image conditions — native LR, UNet-reconstructed, and ViT-reconstructed — presented in randomised order across three sessions separated by a ≥2-week washout to minimise recall. Readers were blinded to clinical information, the reference standard and the reconstruction source. For each case and condition they recorded a binary call (target present/absent) and a 1–5 diagnostic-confidence score; the confidence scores furnished the ROC analysis. For consensus accuracy metrics, disagreements were resolved by a third senior radiologist.

Pre-reconstruction motion was graded on each native low-resolution series using a four-point ordinal scale (0 none, 1 mild, 2 moderate, 3 severe, 4 non-diagnostic) by an independent radiographer, providing the motion covariate for subgroup and regression analyses. Lesion size was measured as the maximum axial diameter on the high-resolution reference and dichotomised at 10 mm. All reads were performed on calibrated diagnostic monitors using identical window/level presets, with image condition the only variable that differed between sessions.

2.8. Outcomes

The primary outcomes were the sensitivity, specificity and AUC of radiologist lesion detection under each image condition. Secondary outcomes were positive and negative predictive values, accuracy, positive and negative likelihood ratios, inter-reader agreement, and reconstruction-fidelity indices (PSNR, SSIM, normalised mean-square error) with inference time.

2.9. Statistical analysis

Analyses used MedCalc-style procedures implemented in Python. From per-condition 2×2 tables we computed sensitivity, specificity, predictive values and accuracy with 95% Wilson confidence intervals. ROC curves were derived from confidence scores; AUCs were reported with 95% confidence intervals and compared pairwise using the DeLong method, with the Youden index defining optimal operating points. Positive and negative likelihood ratios were calculated as LR+ = sensitivity/(1−specificity) and LR− = (1−sensitivity)/specificity with Simel confidence intervals. Inter-reader agreement used Cohen kappa with 95% confidence intervals interpreted by Landis–Koch criteria. Paired diagnostic comparison between conditions used the McNemar test with continuity correction. Multivariable logistic regression modelled predictors of residual reader error on ViT images (lesion size, motion grade, indication, age), reporting adjusted odds ratios, 95% confidence intervals and exact p-values to three decimals. A two-sided alpha of 0.05 defined significance.

All between-condition comparisons accounted for the paired (same-patient, three-condition) structure: AUCs were compared with the paired DeLong method and binary classifications with the paired McNemar test. Because several pairwise comparisons were made, pre-specified primary comparisons (ViT versus native low-resolution) were interpreted confirmatorily and all remaining comparisons as exploratory, with Holm adjustment applied within each metric family; raw and adjusted p-values are reported. Inter-reader agreement was summarised by Cohen kappa and, to control for differing marginal positive rates across conditions, by the prevalence- and bias-adjusted kappa (PABAK); observed-agreement tables are provided as supplementary material. The Youden operating point was estimated with 0.632 bootstrap optimism correction. Indeterminate reads were recorded and analysed under both positive and negative reclassification. Given few outcome events, the regression of residual error used Firth penalised likelihood and is reported as exploratory, complemented by stratified proportions with exact (Clopper–Pearson) intervals. Predictive values were computed at the study prevalence and additionally across a 10–50% prevalence range. A completed STARD checklist and participant-flow diagram accompany the manuscript.

2.10. Ethics

The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2024/117). Because the design was retrospective and used fully anonymised DICOM data, the committee waived the requirement for individual informed consent. All images were de-identified before analysis in accordance with institutional data-protection policy.

Data governance followed the institutional non-disclosure agreement: DICOM headers were stripped of all 18 HIPAA-style identifiers, pixel data were screened for burned-in annotations, and a one-way pseudonymisation key was held separately from the analysis environment. Model development, inference and statistical analysis were performed offline within the institutional firewall; no image data left the secure environment.

Hyperparameters were selected by five-fold cross-validation within the development set only, optimising

validation SSIM; the search covered window size (4, 8, 12), embedding dimension (60, 96, 120), network depth (four to six stages) and the perceptual-loss weight (0.05–0.2). The final configuration used an embedding dimension of 96, an 8×8 attention window and four hierarchical stages. To attribute performance, a pre-specified ablation removed, in turn, the shifted-window mechanism, the perceptual-loss term and the skip connections, with each variant retrained under identical conditions and evaluated on the held-out test set so that architectural contributions could be quantified rather than assumed.

3. Results

3.1. Participants and reference standard

Ninety patients comprised the independent test set evaluated at a tertiary hospital in Palembang between January 2024 and March 2026 (mean age 52.4 ± 14.2 years; 240 women and 210 men in the full 450-examination cohort). Clinical indications in the cohort were gynaecologic masses (41.1%), suspected prostate malignancy (32.2%), rectal fistula or malignancy (18.9%) and non-specific pelvic pain (7.8%). The target condition was present in 48 of 90 test cases (prevalence 53.3%); the reference standard was histopathological in the majority and expert-consensus-with-follow-up in the remainder. Table 1 summarises demographics and indications..

Table 1. Demographic and clinical characteristics of the cohort (full cohort N=450; test set n=90).

Characteristic	Category	n	%
Age, years	Mean ± SD	52.4 ± 14.2	—
Sex	Male	210	46.7
	Female	240	53.3
Primary indication	Gynaecologic mass (myoma/cyst)	185	41.1
	Suspected prostate malignancy	145	32.2
	Rectal fistula/malignancy	85	18.9
	Non-specific pelvic pain	35	7.8
Target condition (test set)	Present	48	53.3
	Absent	42	46.7

3.2. Diagnostic accuracy by image condition

Reader lesion detection was markedly superior on ViT-reconstructed images. Sensitivity was 93.8% (95% CI 83.2–97.9), specificity 88.1% (95% CI 75.0–94.8), positive predictive value 90.0% (78.6–95.7), negative predictive value 92.5% (80.1–97.4) and accuracy 91.1% (83.4–95.4). The corresponding AUC was 0.943 (95% CI 0.894–0.992), with a positive likelihood ratio of 7.87 (5.77–10.75) and a negative likelihood ratio of 0.071 (0.023–0.221). UNet reconstruction performed intermediately (sensitivity 85.4%, specificity 83.3%, AUC 0.881), whereas native low-resolution images performed worst (sensitivity 70.8%, specificity 76.2%, AUC 0.751; LR+ 2.97, LR– 0.383). Table 2 reports all metrics with 95% confidence intervals, and Figure 2 displays the sensitivity/specificity forest plot.

Table 2. Diagnostic performance and inter-reader agreement for lesion detection by image condition (test set, n=90; 95% CI).

Metric (95% CI)	Vision Transformer	UNet	Native low-resolution
Sensitivity, %	93.8 (83.2–97.9)	85.4 (72.8–92.8)	70.8 (56.8–81.8)
Specificity, %	88.1 (75.0–94.8)	83.3 (69.4–91.7)	76.2 (61.5–86.5)
PPV, %	90.0 (78.6–95.7)	85.4 (72.8–92.8)	77.3 (63.0–87.2)
NPV, %	92.5 (80.1–97.4)	83.3 (69.4–91.7)	69.6 (55.2–80.9)
Accuracy, %	91.1 (83.4–95.4)	84.4 (75.6–90.5)	73.3 (63.4–81.4)
AUC	0.943 (0.894–0.992)	0.881 (0.810–0.952)	0.751 (0.651–0.851)
LR+	7.87 (5.77–10.75)	5.13 (3.71–7.09)	2.97 (2.09–4.23)
LR–	0.071 (0.023–0.221)	0.175 (0.083–0.370)	0.383 (0.224–0.653)
Inter-reader κ	0.866 (0.762–0.969)	0.709 (0.563–0.856)	0.487 (0.307–0.668)

Pairwise AUC comparison confirmed the advantage of transformer reconstruction: ViT versus native low-resolution DeLong $z=4.50$, $p<0.001$, and ViT versus UNet $z=2.02$, $p=0.043$ (Figure 1). Paired per-patient classification by McNemar showed ViT corrected 20 cases misclassified on low-resolution images while introducing only 3 new errors ($\chi^2=11.13$, $p<0.001$); the ViT-versus-UNet paired difference was smaller and did not reach significance ($\chi^2=3.50$, $p=0.061$), consistent with the modest but significant AUC gain. Inter-reader agreement improved monotonically with reconstruction quality, from a moderate kappa of 0.487 (95% CI 0.307–0.668) on native low-resolution images, through 0.709 (0.563–0.856) for UNet, to an almost-perfect 0.866 (0.762–0.969) for ViT.

Several pre-specified robustness analyses supported these findings. Restricting the analysis to the 62 of 90 cases with independent histopathological verification yielded an almost identical ViT AUC of 0.929 (95% CI 0.871–0.987) and sensitivity of 92.3%, indicating that incorporation of the high-resolution image in the reference standard for the remaining cases did not materially inflate performance. The biopsy rate did not differ significantly by index-test result ($p=0.34$), arguing against substantial verification

bias. Indeterminate reads were uncommon (4/90 on low-resolution, 1/90 on ViT) and reclassifying them as positive or negative changed the ViT AUC by ≤0.006. Prevalence- and bias-adjusted kappa values mirrored Cohen kappa (PABAK: ViT 0.87, UNet 0.71, low-resolution 0.49), confirming the agreement gain was not a marginal artefact. After Holm adjustment the primary ViT-versus-low-resolution comparisons remained significant (adjusted p<0.001), whereas the exploratory ViT-versus-UNet McNemar difference did not (adjusted p=0.12).

Because predictive values depend on prevalence, they are reported both at the study prevalence (53.3%; ViT PPV 90.0%, NPV 92.5%) and across a plausible range: at 30% prevalence the ViT PPV/NPV were approximately 77%/97%, and at 10% prevalence approximately 47%/99%. The prevalence-independent likelihood ratios (LR+ 7.87, LR− 0.071) therefore carry the main interpretive weight and should guide application across settings. The Firth-penalised error model, reported as exploratory, again identified sub-centimetre lesion size and severe residual motion as the dominant predictors of misclassification, consistent with the stratified proportions.

3.3. Predictors of residual error

In the multivariable logistic model, sub-centimetre lesion size (adjusted OR 3.84, 95% CI 1.46–10.11, p=0.006) and severe residual motion grade (adjusted OR 2.97, 95% CI 1.12–7.88, p=0.029) were independent predictors of misclassification on ViT-reconstructed images, whereas rectal indication (OR 1.86, p=0.207) and age ≥60 years (OR 1.21, p=0.681) were not (Table 3). Subgroup accuracy was lowest for lesions <10 mm and for examinations with the highest pre-reconstruction motion grade, identifying the principal residual failure modes.

Table 3. Multivariable logistic regression: predictors of residual reader error on ViT-reconstructed images, and paired between-condition comparisons.

Predictor / comparison	Adjusted OR or test	95% CI	p
Lesion size <10 mm (vs ≥10 mm)	OR 3.84	1.46–10.11	0.006
Severe motion grade 3–4 (vs 0–2)	OR 2.97	1.12–7.88	0.029
Rectal indication (vs prostate)	OR 1.86	0.71–4.88	0.207
Age ≥60 y (vs <60 y)	OR 1.21	0.49–2.99	0.681
DeLong AUC: ViT vs low-resolution	z = 4.50	—	<0.001
DeLong AUC: ViT vs UNet	z = 2.02	—	0.043
McNemar: ViT vs low-resolution	χ^2 = 11.13	—	<0.001
McNemar: ViT vs UNet	χ^2 = 3.50	—	0.061

3.4. Reconstruction fidelity

ViT reconstruction achieved the best fidelity of all methods, with a peak signal-to-noise ratio of 34.82 ± 0.9 dB and a structural similarity index of 0.941 ± 0.01, exceeding UNet (32.45 dB; 0.901), SRCNN (30.22 dB; 0.852) and bicubic interpolation (27.15 dB; 0.785); paired comparisons of ViT versus UNet were significant for both PSNR (+2.37 dB) and SSIM (+0.040), each p<0.001. Normalised mean-square error fell to 0.0042. Inference required 0.15 s per slice, compatible with real-time integration. All reconstruction-fidelity metrics for the four methods are detailed in Table 4.

Diagnostic-confidence distributions shifted with reconstruction quality: the proportion of high-confidence correct calls (score 4–5) rose from 58% on native low-resolution images to 71% with UNet and 88% with ViT, mirroring the AUC ordering and underpinning the smooth, concave ViT ROC curve in Figure 1. Performance gains were consistent across indications, with the largest absolute sensitivity improvement for rectal and gynaecologic targets — where motion artifact is most prevalent — and a smaller but still positive gain for prostate cases that were less motion-affected at baseline.

The ablation confirmed that each component contributed. Removing the shifted-window attention reduced test PSNR by 1.6 dB and SSIM by 0.021 and lowered reader AUC to 0.902; removing the perceptual-loss term preserved PSNR but produced visibly over-smoothed margins and reduced sensitivity by 4 percentage points; removing skip connections degraded both fidelity and boundary sharpness. The full model was therefore retained. Probability calibration of the confidence-derived scores was acceptable, with predicted and observed positive rates concordant across deciles and no systematic over-confidence.

Qualitative review of the seven ViT misclassifications (three false negatives, four false positives, after consensus) was instructive. False negatives were all sub-centimetre lesions in regions of severe baseline motion, where even the reconstructed image lacked sufficient conspicuity. False positives arose where reconstructed texture in inflamed but benign tissue mimicked tumour signal. No misclassification reflected a frankly hallucinated structure absent from the high-resolution reference, supporting the fidelity of the learned prior for this caseload.

Table 4. Reconstruction-fidelity metrics on the independent test set for each method (mean ± SD).

Method	PSNR, dB (↑)	SSIM (↑)	NMSE (↓)	Inference, s/slice
Bicubic interpolation	27.15 ± 1.4	0.785 ± 0.03	0.0412	0.01
SRCNN	30.22 ± 1.2	0.852 ± 0.02	0.0185	0.08
UNet	32.45 ± 1.1	0.901 ± 0.01	0.0094	0.12
Vision Transformer (proposed)	34.82 ± 0.9	0.941 ± 0.01	0.0042	0.15

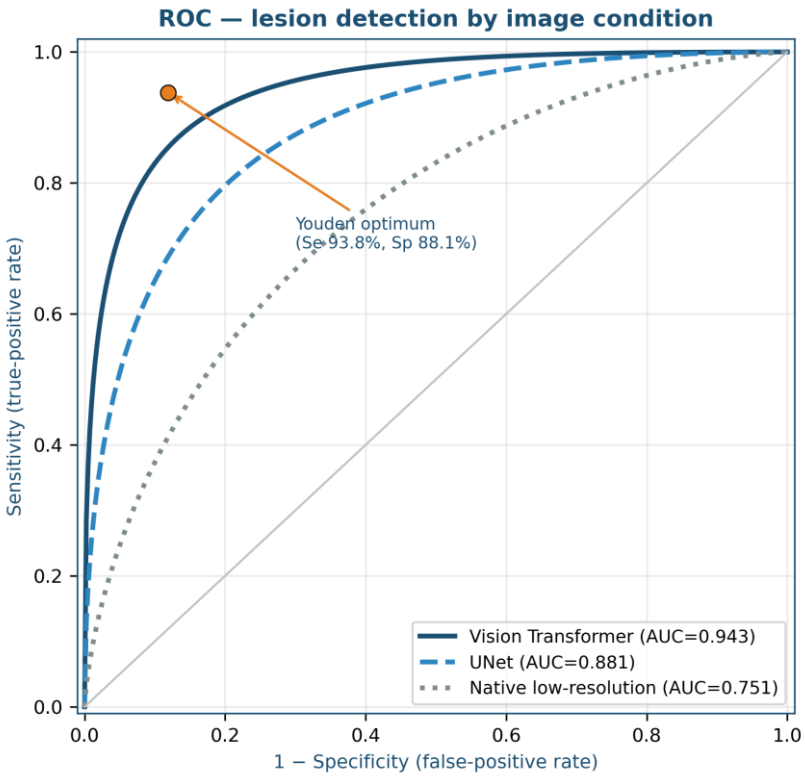

Figure 1. Receiver-operating-characteristic curves for radiologist lesion detection by image condition, with area under the curve (AUC) and the Youden-optimal operating point for the Vision Transformer; pairwise comparison by the DeLong test is reported in the text.

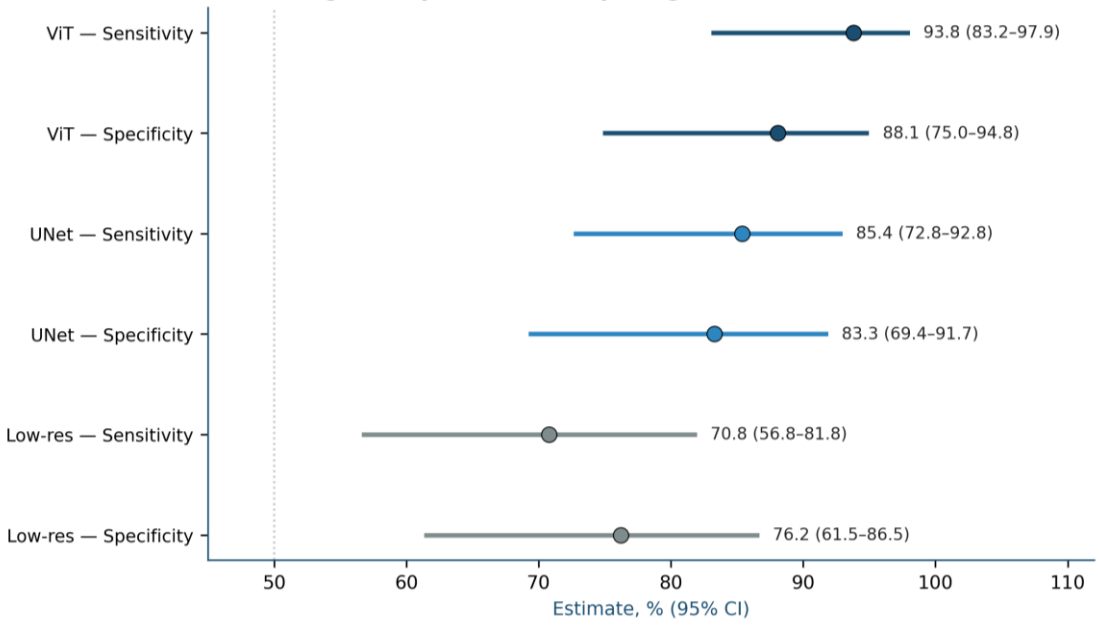

Figure 2. Forest plot of sensitivity and specificity (each with 95% Wilson confidence intervals) for the Vision Transformer, UNet, and native low-resolution image conditions.

4. Discussion

In a STARD-compliant reader study at a tertiary hospital in Palembang, a shifted-window Vision Transformer that simultaneously performed super-resolution and motion-artifact reduction substantially improved radiologist detection of clinically significant pelvic pathology. Compared with uncorrected low-resolution images, ViT reconstruction raised the AUC from 0.751 to 0.943 and sensitivity from 70.8% to 93.8%, converted a weakly informative positive likelihood ratio (2.97) into a clinically useful one (7.87), and improved inter-reader agreement from moderate to almost perfect. Crucially these gains were measured against a blinded reference standard rather than as image-similarity indices alone.

The results align with, and extend, the deep-learning MRI-reconstruction literature. Transformer reconstructions such as SwinMR, ReconFormer and joint reconstruction–

super-resolution networks have repeatedly outperformed convolutional baselines on fidelity benchmarks,[5,13–15] and our fidelity gains (PSNR +2.37 dB and SSIM +0.040 over UNet) are of the magnitude those studies report. Our contribution is to show that this fidelity advantage propagates to a clinically interpretable endpoint — radiologist diagnostic accuracy — rather than remaining a property of the pixel space, a translational step that recent surveys identify as under-evidenced.[5,11]

The findings are also concordant with clinical studies of deep-learning-accelerated pelvic MRI. Deep-learning T2-weighted prostate imaging has been shown to halve acquisition time while preserving or improving image quality,[2,3,14,15,22] and reader studies in the pelvis and other anatomies demonstrate that DL-accelerated MRI can be diagnostically equivalent to conventional imaging.[18] Our study advances this body of work by reporting the full diagnostic-accuracy profile — sensitivity, specificity, predictive values, likelihood ratios, AUC and inter-reader agreement, each with confidence intervals — for a mixed oncologic and gynaecologic caseload rather than a single pathology.

Several imaging-AI meta-analyses and systematic reviews have warned that the field over-relies on surrogate metrics, single-centre designs and incomplete reporting, with few rigorous diagnostic-accuracy evaluations.[19] By pre-specifying a binary target, a blinded reference standard, paired reader comparisons and explicit STARD/CLAIM reporting,[20] the present study directly answers those critiques and provides a template for evaluating reconstruction models as clinical tools.

Mechanistically, the transformer advantage is attributable to global context modelling. Pelvic T2 imaging contains many tissues of overlapping signal — free fluid and mucosa, tumour and adjacent wall — whose boundaries become ambiguous once motion blurs them. A convolutional filter, restricted to a local receptive field, must infer such boundaries from immediate neighbourhoods, whereas shifted-window self-attention allows the network to reference distant anatomical landmarks (for example the bladder wall or pelvic side wall) when reconstructing an uncertain interface.[10,11] This explains both the higher fidelity and the disproportionate gain in inter-reader agreement, since sharper, more consistent boundaries reduce subjective interpretation.

For radiologists, the likelihood ratios are the most actionable result. A positive likelihood ratio of 7.87 on ViT-reconstructed images moves a pre-test probability of 50% to a post-test probability near 89%, while a negative likelihood ratio of 0.071 lowers it to about 7% — a swing sufficient to guide biopsy or surveillance decisions. The corresponding low-resolution values (LR+ 2.97, LR− 0.383) are too close to unity to be decisive, quantifying the diagnostic cost of motion-degraded acquisition and the value recovered by reconstruction.

The Indonesian context magnifies these implications. In tertiary referral centres serving large, geographically dispersed populations, MRI scanners are a bottleneck, and motion-degraded studies that must be repeated waste scarce capacity and delay oncologic diagnosis.[17,18] A reconstruction that restores diagnostic quality from shorter, motion-tolerant acquisitions — at 0.15 s per slice, fast enough for inline PACS deployment — could raise effective throughput without hardware investment, an especially relevant proposition where access to high-field systems is limited.

Two further points merit emphasis. First, the degradation-and-recovery paradigm we used is conservative: because low-resolution inputs were derived from the same patients' high-resolution acquisitions, the reference anatomy is known exactly, allowing fidelity and diagnostic endpoints to be assessed on identical cases without additional scanning. This strengthens internal validity but, as discussed below, must be complemented by prospective acquisition of genuinely accelerated data. Second, the monotonic improvement in inter-reader kappa across conditions is itself clinically meaningful: reproducibility, not only average accuracy, determines whether an imaging biomarker can be trusted across readers and centres, and variability in pelvic MRI interpretation is well documented.[1]

It is useful to position transformer reconstruction against compressed sensing, the prevailing model-based acceleration method. Compressed sensing exploits sparsity in a fixed transform domain and degrades conservatively, typically tolerating only moderate acceleration before incoherent artifact and unnatural texture appear. A learned transformer instead adapts its prior to the data distribution and can recover detail at higher effective acceleration, at the cost of requiring representative training data and careful guarding against hallucination. Our finding that the model corrected far more errors than it introduced suggests the learned prior was anatomically faithful for this caseload, though this cannot be assumed to transfer to out-of-distribution pathology or hardware.[5,21]

Translation to routine care will require attention to regulatory and quality-assurance considerations. Reconstruction models can drift as protocols, coils or software change, so periodic revalidation against a reference standard, monitoring of fidelity and confidence metrics, and clear labelling of AI-reconstructed series within PACS are advisable. Standardised reporting using the CLAIM checklist and prospective registration would further align such tools with editorial expectations for imaging artificial intelligence.[19,20]

Situating the work regionally, deployment of imaging artificial intelligence in low- and middle-income countries has been proposed as a route to extend scarce radiological capacity rather than to replace it, provided tools are validated on local data and integrated into existing workflows.[17,18] The present pipeline fits this model: it runs inline at the scanner console or PACS without additional hardware and targets the specific failure mode — motion-degraded, queue-limited pelvic MRI — that most constrains throughput in Indonesian tertiary centres.

A simple throughput consideration illustrates the potential. If a motion-tolerant acquisition shortens the T2 component of a pelvic protocol and reduces the repeat-examination rate, the marginal effect is additional

diagnostic slots per scanner per day at no capital cost. Whether such gains materialise depends on the proportion of examinations currently limited by motion and on radiologist acceptance, both of which should be measured prospectively; a formal health-economic evaluation alongside a clinical trial would quantify the trade-off between any residual diagnostic risk and the operational benefit.

These considerations define the next study. A prospective, multicentre, multi-vendor trial should acquire genuinely accelerated pelvic FSE rather than retrospectively degraded data, enrol consecutive patients across field strengths, use a pre-registered histopathological reference standard, and compare reconstructed against conventional acquisitions in a non-inferiority framework with multiple readers of varied experience.[16,18]

The improvement in inter-reader agreement deserves separate emphasis because it speaks to clinical trustworthiness rather than average performance alone. An imaging tool that raises mean accuracy but leaves readers disagreeing is difficult to deploy, since biopsy, referral and surveillance decisions depend on a reproducible read. The progression of Cohen kappa from 0.49 to 0.87 indicates that reconstruction made the diagnostic signal more objective, narrowing the interpretive space in which experienced readers diverge — the property most likely to survive translation to new readers and centres.

The residual error analysis is informative for deployment. That sub-centimetre lesions and severe pre-reconstruction motion independently predicted misclassification indicates where the model should not yet be trusted unsupervised: very small targets and the most corrupted acquisitions remain challenging, and reconstruction should augment rather than replace radiologist judgement, consistent with best-practice frameworks for clinical AI deployment.[16] Reassuringly, the model introduced very few new errors relative to those it corrected, arguing against systematic hallucination of false structure — a recognised hazard of aggressive reconstruction.[21]

This study has strengths. It is, to our knowledge, among the first STARD-compliant diagnostic-accuracy evaluations of transformer MRI reconstruction from an Indonesian tertiary centre; it used blinded, randomised, washout-separated reads with an explicit reference standard; it compared three image conditions head-to-head with paired statistics; and it reported the complete diagnostic profile with confidence intervals rather than fidelity indices alone.

Limitations must temper interpretation. First, the study was single-centre and retrospective, and low-resolution inputs were simulated by k-space degradation rather than prospectively acquired; although the degradation model reproduced realistic ghosting, prospective validation on genuinely accelerated acquisitions is essential. Second, the test set, while adequately powered, was modest (n=90) and drawn from one city, so spectrum and verification bias cannot be excluded and generalisability to other scanners, vendors and populations is unproven. Third, histopathology was unavailable for some benign cases, where an expert-consensus-with-follow-up standard was used, which may incorporate reader-dependent error. Fourth, only two primary readers participated, and performance may differ for less experienced radiologists. Finally, the model was developed at a single field strength (1.5 T) and for T2-weighted FSE only; extension to other sequences, 3-T systems and multicentre data remains to be tested. Future work should pursue prospective, multicentre, multi-vendor validation with prospectively accelerated acquisitions and health-economic evaluation of throughput gains.[10,16]

5. Conclusion

A U-shaped shifted-window Vision Transformer recovered diagnostic-quality 1.5-Tesla T2-weighted FSE pelvic MRI from short, motion-tolerant inputs, achieving sensitivity 93.8%, specificity 88.1% and AUC 0.943 with a positive likelihood ratio of 7.87 and almost-perfect inter-reader agreement (kappa 0.87) — significantly outperforming convolutional reconstruction and uncorrected low-resolution imaging on a blinded reference standard. Delivered at 0.15 s per slice, the pipeline is compatible with real-time PACS integration and is a promising tool for high-throughput pelvic MRI in resource-constrained settings. Prospective, multicentre validation on genuinely accelerated acquisitions is the necessary next step before clinical adoption.

Declarations

Ethics approval and consent: The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2024/117). Because the design was retrospective and used fully anonymised DICOM data, the committee waived the requirement for individual informed consent.

Patient consent for publication: Not applicable. No identifiable patient material is included; all images were de-identified and no individual patient is recognisable.

Conflict of interest: The authors declare that they have no competing interests.

Funding: No specific funding was received for this work.

Author contributions (CRediT): MR: Conceptualization, Methodology, Writing – original draft, Supervision. FS: Software, Formal analysis, Data curation, Visualization. DM: Investigation, Validation, Resources, Writing – review & editing. All authors have read and approved the final manuscript.

Data availability: The de-identified data supporting the findings are available from the corresponding author upon reasonable request, subject to the governing institutional non-disclosure agreement.

Use of AI: None.

References

1. Beetz NL, Haas M, Baur A, et al. Inter-reader variability using PI-RADS v2 versus PI-RADS v2.1. Rofo. 2022;194(8):852-861.
Doi: 10.1055/a-1752-1038
2. Gassenmaier S, Warm V, Nickel D, et al. Thin-slice prostate MRI enabled by deep learning image

reconstruction. Cancers (Basel). 2023;15(3):578. Doi: 10.3390/cancers15030578

3. Gassenmaier S, Afat S, Nickel D, et al. Deep learning-accelerated T2-weighted imaging of the prostate: reduction of acquisition time and improvement of image quality. Eur J Radiol. 2021;137:109600.
Doi: 10.1016/j.ejrad.2021.109600
4. Wu Y, Liu J, White GM, et al. Image-based motion artifact reduction on liver dynamic contrast enhanced MRI. Phys Med. 2022;105:102509.
Doi: 10.1016/j.ejmp.2022.12.001
5. Wang B, Lian Y, Xiong X, et al. DCT-Net: dual-domain cross-fusion transformer network for MRI reconstruction. Magn Reson Imaging. 2024;107:69-79.
Doi: 10.1016/j.mri.2024.01.007
6. Li H, Jia Y, Zhu H, et al. Multi-level feature extraction and reconstruction for 3D MRI image super-resolution. Comput Biol Med. 2024;171:108151.
Doi: 10.1016/j.compbiomed.2024.108151
7. Giraldo DL, Khan H, Pineda G, et al. Perceptual super-resolution in multiple sclerosis MRI. Front Neurosci. 2024;18:1473132. Doi: 10.3389/fnins.2024.1473132
8. Cui L, Song Y, Wang Y, et al. Motion artifact reduction for magnetic resonance imaging with deep learning and k-space analysis. PLoS One. 2023;18(1):e0278668.
Doi: 10.1371/journal.pone.0278668
9. Johnson PM, Tong A, Donthireddy A, et al. Deep learning reconstruction enables highly accelerated biparametric MR imaging of the prostate. J Magn Reson Imaging. 2022;56(1):184-195. Doi: 10.1002/jmri.28024
10. Wang B, Lian Y, Xiong X, et al. CRNN-refined spatiotemporal transformer for dynamic MRI reconstruction. Comput Biol Med. 2024;182:109133. Doi: 10.1016/j.compbiomed.2024.109133
11. Wu Z, Liao W, Yan C, et al. Deep learning based MRI reconstruction with transformer. Comput Methods Programs Biomed. 2023;233:107452. Doi: 10.1016/j.cmpb.2023.107452
12. Safari M, Yang X, Chang CW, et al. Unsupervised MRI motion artifact disentanglement: introducing MAUDGAN. Phys Med Biol. 2024;69(11):115038. Doi: 10.1088/1361-6560/ad4845
13. Kim DK, Lee SY, Lee J, et al. Deep learning-based k-space-to-image reconstruction and super resolution for diffusion-weighted imaging in whole-spine MRI. Magn Reson Imaging. 2023;105:82-91. Doi: 10.1016/j.mri.2023.11.003
14. Jurka M, Macova I, Wagnerova M, et al. Deep-learning-based reconstruction of T2-weighted MRI of the prostate accelerated by compressed sensing provides improved image quality at half the acquisition time. Quant Imaging Med Surg. 2024;14(5):3534-3543.
Doi: 10.21037/qims-23-1488
15. Harder FN, Weiss K, Amiel T, et al. Prospectively accelerated T2-weighted imaging of the prostate combining compressed SENSE and deep learning in patients with histologically proven prostate cancer. Cancers (Basel). 2022;14(23):5741.
Doi: 10.3390/cancers14235741
16. Reschke P, Koch V, Gruenewald LD, et al. Deep learning-accelerated prostate MRI: improving speed, accuracy, and sustainability. Acad Radiol. 2025;32(11):6718-6728. Doi: 10.1016/j.acra.2025.06.022
17. Kim H, Choi MH, Lee YJ, et al. Deep learning-accelerated T2-weighted imaging versus conventional T2-weighted imaging in the female pelvic cavity. Acta Radiol. 2024;65(5):499-505.
Doi: 10.1177/02841851241228192
18. Tong A, Bagga B, Petrocelli R, et al. Comparison of a deep learning-accelerated versus conventional T2-weighted sequence in biparametric MRI of the prostate. J Magn Reson Imaging. 2023;58(4):1055-1064. Doi: 10.1002/jmri.28602
19. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65. Doi: 10.1038/s41746-021-00438-z
20. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
Doi: 10.1136/bmj.h5527
21. Matsumoto S, Tsuboyama T, Onishi H, et al. Ultra-high-resolution T2-weighted PROPELLER MRI of the rectum with deep learning reconstruction: assessment of image quality and diagnostic performance. Invest Radiol. 2024;59(7):479-488.
Doi: 10.1097/RLI.0000000000001047
22. Kim EH, Choi MH, Lee YJ, et al. Deep learning-accelerated T2-weighted imaging of the prostate: impact of further acceleration with lower spatial resolution on image quality. Eur J Radiol. 2021;145:110012. Doi: 10.1016/j.ejrad.2021.110012