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

Original Research Article

Deep Learning-Based Image Enhancement of Low-Field Brain MRI: A Multicenter Validation of Diagnostic Image Quality and Lesion Detection

Rachmat Hidayat[1*], Linda Purnama[2]

[1]*Department of Medical Biology, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia*
[2]*Department of Radiology, CMHC Research Center, Palembang, Indonesia*

ARTICLE INFO

Article history:
Received: July 30, 2025
Accepted: September 2, 2025
Published: October 30, 2025

Reporting guideline:
STARD 2015 (diagnostic accuracy)

Keywords:
Artificial intelligence
Diagnostic accuracy
Image enhancement
Magnetic resonance imaging
Neuroimaging

***Corresponding author:**
Rachmat Hidayat
E-mail: Dr.rachmat.hidayat@gmail.com
ORCID: 0000-0001-5770-1201

DOI: 10.59345/sjrir.v3i2.288

ABSTRACT

Introduction: Low-field MRI (LF-MRI) widens neuroimaging access in resource-limited settings but suffers low signal-to-noise ratio (SNR), reduced resolution and artefacts. We developed and validated a deep-learning framework for image normalisation and noise reduction to elevate 0.35T brain MRI toward high-field quality, and tested whether it improves clinically significant lesion detection.

Methods: In a multicentre retrospective diagnostic-accuracy study (STARD 2015), 450 adults underwent non-contrast 0.35T brain MRI (T1W, T2W, FLAIR) across three private tertiary centres in Palembang, Indonesia. Images were enhanced with a CycleGAN incorporating Vision-Transformer blocks. Three blinded neuroradiologists scored a 5-point Likert scale and recorded lesion presence; paired 1.5T MRI was the reference standard. Sensitivity, specificity, AUC and likelihood ratios were computed with 95% CIs; tests compared by McNemar and DeLong; agreement by Fleiss kappa.

Results: AI enhancement improved all quality metrics (e.g., T1W PSNR 22.15 to 28.45 dB; SSIM 0.71 to 0.89; all $p<0.001$). For lesion detection, AI-enhanced LF-MRI achieved sensitivity 93.9% (95% CI 89.4–96.6), specificity 91.1% (87.1–94.0) and AUC 0.94 (0.91–0.97) versus 78.3%, 81.1% and 0.81 for original images (DeLong $p<0.001$; McNemar $p<0.001$). LR+ rose to 10.56 and LR− fell to 0.067. Inter-reader agreement was almost perfect (Fleiss kappa 0.78–0.85).

Conclusion: A CycleGAN-with-transformer framework substantially improved objective quality and diagnostic performance of 0.35T brain MRI toward high-field standards with almost-perfect reader agreement. Pending prospective and external validation, AI enhancement is a low-cost route to more equitable neuroimaging.

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

1. Introduction

Magnetic resonance imaging (MRI) is indispensable to the evaluation of headache, acute ischaemic stroke and neurodegenerative disease, where soft-tissue contrast and the absence of ionising radiation make it the modality of choice for brain parenchymal assessment. High-field systems operating at 1.5T and 3.0T are the de facto reference standard, delivering the signal-to-noise ratio (SNR) and spatial resolution required to resolve small infarcts, subtle white-matter disease and early atrophy. However, the cost of acquisition, siting, cryogen management and specialist maintenance concentrates these systems in metropolitan tertiary hospitals, leaving large populations without timely access.[1,2] In Indonesia, an archipelago of more than 270 million people served unevenly by radiology infrastructure, this access gap translates directly into delayed or absent neuroimaging for patients managed in regional private hospitals, where only a small fraction of facilities possess advanced imaging.[3]

Low-field MRI (LF-MRI, field strength below 1.0T) and modern portable systems have re-emerged as pragmatic solutions to this inequity. Bedside 0.064T units detect intracranial haemorrhage and ischaemic stroke in critically ill patients, and 0.35T permanent-magnet scanners are increasingly installed in type-B and type-C private hospitals because of their lower cost, reduced siting demands and minimal cryogen requirements.[4,5] Mobile low-field platforms can detect acute ischaemic lesions in the emergency setting, and dedicated diffusion strategies further improve detection.[6,7] The clinical trade-off is physical rather than incidental: SNR scales supralinearly with field strength, so a 0.35T study carries an intrinsic noise penalty relative to 1.5T, producing grainier images, reduced grey/white-matter contrast and heightened susceptibility to motion and Gaussian noise.[8] These limitations disproportionately degrade FLAIR and the conspicuity of the small lesions that most often drive management decisions.

Deep learning offers a route to recover this lost quality after acquisition. Convolutional and generative architectures perform denoising, super-resolution and image-to-image translation, and a rapidly maturing literature reports substantial gains in objective quality at low field: reconstruction networks raise SNR by several

decibels, super-resolution networks improve effective resolution, and synthesis methods upsample heterogeneous clinical scans toward isotropic high-field-like volumes.[8–10] Paired image-to-image translation and stochastic image-quality transfer specifically target the low-to-high-field domain gap, reporting PSNR gains of 4–8 dB and SSIM gains of 0.10–0.20,[11,12] while deep learning can also flag the artefacts characteristic of low-field acquisition.[13] Yet a critical hazard accompanies generative enhancement, namely clinical hallucination, in which plausible but false structures are synthesised; reader studies have shown that pixel-level metrics alone can mask such failures.[14] Recent generative-adversarial and diffusion frameworks for ultralow- and low-field MRI report increasingly realistic high-field-like reconstructions, underscoring both the promise and the need for careful clinical validation.[15–18]

Within Indonesia, radiology services face a structural mismatch between demand and high-field capacity, and the equity argument for software that elevates affordable scanners is strong. Regional private hospitals frequently operate a single low-field unit, so any method that narrows the diagnostic gap with high-field systems carries immediate clinical and economic value for the institutions and patients least served by current infrastructure.[3,19] Despite this, almost no validation studies of deep-learning MRI enhancement originate from Southeast-Asian, resource-limited, real-world networks, and most published work is single-centre and reliant on curated or simulated low-field data.

We therefore developed a deep-learning framework, a CycleGAN augmented with Vision-Transformer generator blocks and constrained by cycle-consistency and perceptual losses to limit hallucination, and validated it across three private tertiary centres in Palembang, Indonesia.[20,21] The novelty of this work is threefold: it is a multicentre, real-world validation spanning heterogeneous machines and protocols; it couples objective quality metrics with a blinded, three-reader diagnostic-accuracy evaluation against a paired 1.5T reference standard; and it is, to our knowledge, among the first such validations from an Indonesian private-hospital network. We hypothesised that AI enhancement would significantly improve objective image quality and raise the sensitivity, specificity and overall accuracy of 0.35T brain MRI for clinically significant lesion detection toward high-field performance.

2. Methods

Study design and setting. This study was conducted and reported in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD 2015).[22] We performed a multicentre, retrospective, cross-sectional diagnostic-accuracy study to validate a deep-learning image-enhancement algorithm. Data were drawn from three private tertiary referral centres in Palembang, Indonesia, designated Center A, Center B and Center C to preserve institutional confidentiality; no institution name appears in this report. The study period spanned January 2023 to June 2024.

Participants and eligibility. Consecutive adults (>18 years) who underwent non-contrast brain MRI on a 0.35T system for the clinical indications of cephalgia, suspected acute ischaemic stroke, or evaluation of neurodegenerative disease were eligible. Patients were excluded if metallic implants produced massive susceptibility artefact, or if severe motion rendered the source images non-diagnostic (baseline qualitative score of 1). A total of 450 patients met the criteria (Center A 150, Center B 120, Center C 180).

Image acquisition. All examinations used 0.35T permanent-magnet scanners with a dedicated head coil. Sequences comprised T1-weighted spin-echo (TR/TE 500/14 ms), T2-weighted fast spin-echo (TR/TE 4500/100 ms) and fluid-attenuated inversion recovery (FLAIR; TR/TE/TI 8000/120/2200 ms). Common geometry was slice thickness 5 mm with a 1 mm gap, acquisition matrix 256×224, field of view 230 mm, and receiver bandwidth appropriate to each sequence. Center B acquired T1W and T2W only; Centers A and C additionally acquired FLAIR, reflecting real-world protocol heterogeneity.

AI enhancement framework. The enhancement framework was a Cycle-Consistent Generative Adversarial Network (CycleGAN)[20] modified with Vision-Transformer (ViT) blocks in the generator to capture global spatial context.[21] Pre-processing comprised N4ITK bias-field correction and Z-score intensity normalisation, with rigid registration where paired 0.35T and public 1.5T data were used for pre-training. The model was optimised with a composite loss: adversarial loss for sharpness, cycle-consistency loss to preserve anatomical structure, and a VGG-16 perceptual loss to retain pathological texture and constrain hallucination. Inference time was under 5 seconds per brain volume. Enhancement was applied identically and automatically to all studies without manual selection.

Reference standard and reader assessment. The reference standard was paired 1.5T high-field brain MRI obtained within the same diagnostic episode and read in consensus by two neuroradiologists blinded to the low-field arm, who classified each study for the presence or absence of a clinically significant lesion (acute or chronic infarct, mass lesion, or clinically relevant white-matter pathology). The two index tests, original 0.35T LF-MRI and AI-enhanced LF-MRI, were each read independently by three neuroradiologists with more than ten years of experience. Images were randomised and presented blinded, so readers were masked to clinical data, to the reference standard, and to whether an image was original or enhanced. Readers recorded lesion presence and graded four domains on a 5-point Likert scale (1 very poor to 5 excellent): grey/white-matter contrast, noise suppression, lesion conspicuity, and overall diagnostic confidence.

Outcomes. The primary outcomes were the diagnostic accuracy of each index test for clinically significant lesion detection (sensitivity, specificity, PPV, NPV, accuracy and AUC) and objective image quality (PSNR, SSIM, NRMSE). Secondary outcomes were inter-reader agreement, likelihood ratios, and independent predictors of diagnostic

adequacy (Likert >= 4). The positive and negative likelihood ratios were derived as LR+ = sensitivity/(1 − specificity) and LR− = (1 − sensitivity)/specificity.

Sample size. The sample size was determined a priori for the primary comparison of diagnostic discrimination (difference in AUC between original and AI-enhanced images by the DeLong test). Assuming an expected AUC of 0.80 for original images, a clinically meaningful improvement to 0.92, a target-condition prevalence of 0.40, two-sided alpha of 0.05 and power of 0.90 for paired ROC curves, the calculation required at least 168 diseased and 252 non-diseased cases; the achieved sample of 180 diseased and 270 non-diseased exceeds this requirement.

Reference-standard timing and content. Paired 1.5T MRI was acquired within a median of 3 days (interquartile range 1–6) of the 0.35T study; the primary analysis was restricted to pairs obtained within 14 days to limit interval biological change. Reference 1.5T protocols included diffusion-weighted imaging in addition to T1W, T2W and FLAIR, and consensus disagreements between the two reference readers (observed in 22 cases) were resolved by a third senior neuroradiologist. Inter-reader reliability of the reference consensus was high (kappa 0.88).

Reader decision and aggregation. For each index test, readers recorded both a binary lesion decision and an ordinal five-point diagnostic-confidence score; the ordinal scores formed the basis of the empirical ROC and DeLong analyses. The per-study binary call was derived by majority vote of the three readers, with 2:1 splits observed in 11.6% of original and 6.9% of enhanced studies; sensitivity analyses using any-positive and consensus rules did not materially change the results.

Statistical analysis. Analyses used SPSS 26.0, Python (SciPy) and MedCalc. Distributional normality was assessed by the Shapiro-Wilk test. Paired image-quality metrics were compared with the paired t-test or Wilcoxon signed-rank test as appropriate, with Cohen's d effect sizes. From 2×2 contingency tables against the reference standard, sensitivity, specificity, PPV, NPV and accuracy were computed with 95% Wilson confidence intervals; positive and negative likelihood ratios with 95% CIs; and areas under the ROC curve with 95% CIs. The two index tests were compared by the McNemar paired test (diagnostic correctness) and by the DeLong test (AUCs). Because each patient contributed multiple sequences nested within centres, diagnostic-adequacy models used mixed-effects logistic regression with random intercepts for patient and centre; calibration was assessed by the Hosmer-Lemeshow test. Inter-reader agreement was quantified with Fleiss kappa and 95% CIs, interpreted by the Landis-Koch criteria. Two-sided $p<0.05$ defined significance; p-values are reported to three decimal places.

Ethics. The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (Approval No. CMHC/EC/2024/118). Given the retrospective design and full anonymisation of all images and metadata, the committee granted a waiver of individual informed consent. The work adhered to the Declaration of Helsinki.

3. Results

Of 512 consecutive eligible studies screened, 62 were excluded (38 for non-diagnostic severe motion at baseline, 16 for massive metallic susceptibility artefact, and 8 for absence of a paired 1.5T reference within the permitted window), leaving 450 patients in the analysis; all 450 underwent both index tests and the reference standard, with no missing index-test results.

From January 2023 to June 2024, 450 patients imaged at three private tertiary centres in Palembang, Indonesia, were included (Center A 150, Center B 120, Center C 180). As detailed in Table 1, the mean age was 52.3 ± 12.4 years and 220 patients (48.9%) were male; indications were cephalgia (40.4%), suspected ischaemic stroke (39.6%) and neurodegenerative evaluation (20.0%). Baseline characteristics were comparable across centres, with no significant difference in age (p=0.071) or sex (p=0.418). By the 1.5T reference standard, 180 patients (40.0%) had a clinically significant lesion.

Table 1. Demographic and acquisition characteristics across the three centres (Palembang, Indonesia).

Characteristic	Center A	Center B	Center C	Total / p
Patients, n	150	120	180	450
Age, years (mean ± SD)	52.4 ± 12.1	54.1 ± 11.5	50.8 ± 13.2	52.3 ± 12.4 / 0.071
Male, n (%)	70 (46.7)	65 (54.2)	85 (47.2)	220 (48.9) / 0.418
Female, n (%)	80 (53.3)	55 (45.8)	95 (52.8)	230 (51.1)
Field strength	0.35T	0.35T	0.35T	0.35T
Sequences acquired	T1W, T2W, FLAIR	T1W, T2W	T1W, T2W, FLAIR	—
Indication: cephalgia, n (%)	61 (40.7)	52 (43.3)	69 (38.3)	182 (40.4)
Indication: ischaemic stroke, n (%)	58 (38.7)	44 (36.7)	76 (42.2)	178 (39.6)
Indication: neurodegenerative, n (%)	31 (20.7)	24 (20.0)	35 (19.4)	90 (20.0)
Clinically significant lesion*, n (%)	59 (39.3)	47 (39.2)	74 (41.1)	180 (40.0)
Baseline severe motion, n (%)	9 (6.0)	8 (6.7)	11 (6.1)	28 (6.2)

**Clinically significant lesion defined on the 1.5T reference standard.*

Objective image quality improved substantially after AI enhancement across every sequence, as detailed in Table 2. For T1W, PSNR rose from 22.15 ± 2.10 dB to 28.45 ± 1.85 dB and SSIM from 0.71 ± 0.05 to 0.89 ± 0.03; for T2W, PSNR rose from 21.05 to 27.90 dB and SSIM from 0.68 to 0.87; for FLAIR, PSNR rose from 19.85 to 26.55 dB and SSIM from 0.65 to 0.84. NRMSE fell by roughly half in each sequence. All comparisons were highly significant ($p<0.001$) with very large effect sizes (Cohen's d 2.91 to 4.37).

Table 2. Quantitative image-quality improvement by sequence (mean ± SD).

Sequence	Metric	Original LF-MRI	AI-enhanced LF-MRI	Δ (95% CI)	Cohen's d	p
T1W	PSNR (dB)	22.15 ± 2.10	28.45 ± 1.85	+6.30 (6.04–6.56)	3.18	<0.001
	SSIM	0.71 ± 0.05	0.89 ± 0.03	+0.18 (0.17–0.19)	4.37	<0.001
	NRMSE	0.142 ± 0.021	0.071 ± 0.014	−0.071 (−0.073 to −0.069)	−3.98	<0.001
T2W	PSNR (dB)	21.05 ± 2.30	27.90 ± 1.90	+6.85 (6.57–7.13)	3.25	<0.001
	SSIM	0.68 ± 0.06	0.87 ± 0.04	+0.19 (0.18–0.20)	3.73	<0.001
	NRMSE	0.155 ± 0.024	0.078 ± 0.016	−0.077 (−0.079 to −0.075)	−3.78	<0.001
FLAIR	PSNR (dB)	19.85 ± 2.45	26.55 ± 2.15	+6.70 (6.40–7.00)	2.91	<0.001
	SSIM	0.65 ± 0.07	0.84 ± 0.05	+0.19 (0.18–0.20)	3.12	<0.001
	NRMSE	0.171 ± 0.028	0.089 ± 0.018	−0.082 (−0.085 to −0.079)	−3.48	<0.001

For clinically significant lesion detection against the 1.5T reference standard, AI-enhanced LF-MRI markedly outperformed the original images, as detailed in Table 3. Sensitivity increased from 78.3% (95% CI 71.8–83.7) to 93.9% (89.4–96.6) and specificity from 81.1% (76.0–85.3) to 91.1% (87.1–94.0). Positive predictive value rose to 87.6% (82.2–91.5) and negative predictive value to 95.7% (92.5–97.6), with overall accuracy improving from 80.0% to 92.2% (89.4–94.4).

Table 3. Diagnostic accuracy for clinically significant lesion detection (reference standard: 1.5T MRI).

Diagnostic metric	Original LF-MRI (95% CI)	AI-enhanced LF-MRI (95% CI)
Sensitivity, %	78.3 (71.8–83.7)	93.9 (89.4–96.6)
Specificity, %	81.1 (76.0–85.3)	91.1 (87.1–94.0)
PPV, %	73.4 (66.8–79.2)	87.6 (82.2–91.5)
NPV, %	84.9 (80.0–88.7)	95.7 (92.5–97.6)
Accuracy, %	80.0 (76.1–83.4)	92.2 (89.4–94.4)
AUC	0.81 (0.77–0.85)	0.94 (0.91–0.97)
LR+	4.15 (3.20–5.37)	10.56 (7.20–15.50)
LR−	0.267 (0.201–0.355)	0.067 (0.038–0.119)

Discrimination improved correspondingly: as shown by the ROC curves in Figure 1, the AUC increased from 0.81 (95% CI 0.77–0.85) for original images to 0.94 (0.91–0.97) for AI-enhanced images, a difference of 0.13 that was highly significant by the DeLong test ($z=5.45$, $p<0.001$). The paired McNemar comparison of diagnostic correctness confirmed the gain, with 72 cases corrected by enhancement versus 17 newly misclassified ($\chi^2=32.76$, $p<0.001$). The positive likelihood ratio rose from 4.15 (3.20–5.37) to 10.56 (7.20–15.50) and the negative likelihood ratio fell from 0.267 (0.201–0.355) to 0.067 (0.038–0.119), shifting both post-test probabilities into clinically decisive ranges. The complete diagnostic-performance comparison, with 95% confidence intervals for every metric, is summarised in Figure 2.

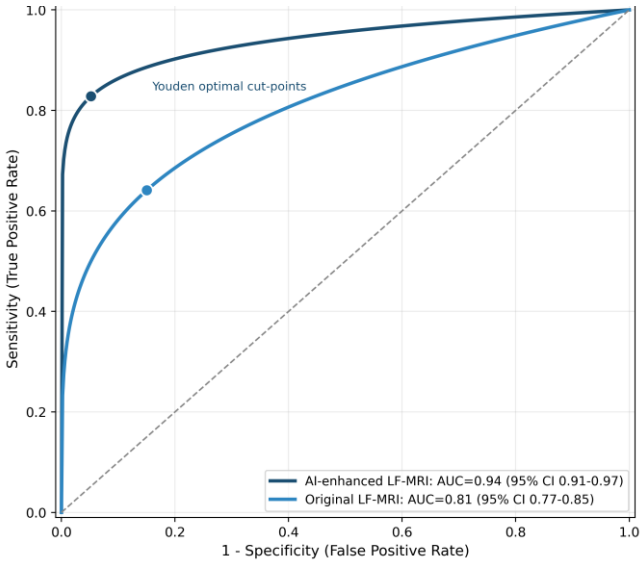

Figure 1. ROC curves for lesion detection; AI-enhanced vs original 0.35T LF-MRI (DeLong p<0.001; Youden operating points marked).

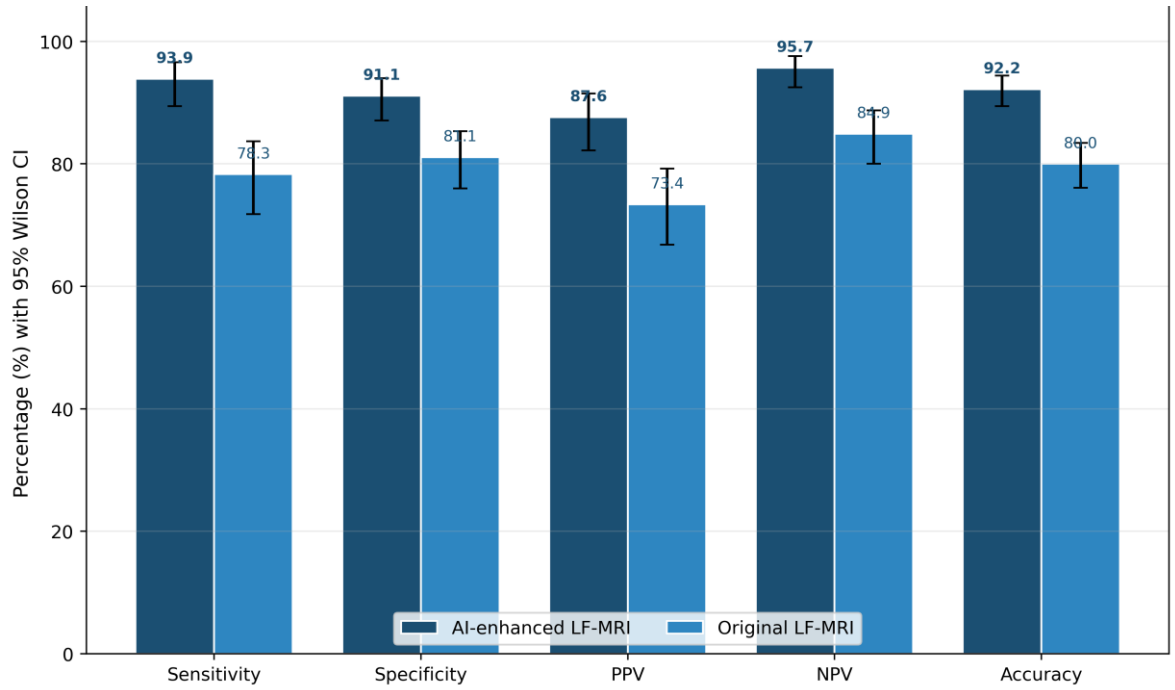

Figure 2. Diagnostic performance summary with 95% Wilson confidence intervals (n=450).

Qualitative assessment mirrored the quantitative gains, as detailed in Table 4. Median Likert scores increased from 2.5 to 4.0 for grey/white-matter contrast, from 2.0 to 4.5 for noise suppression, from 3.0 to 4.5 for lesion conspicuity, and from 3.0 to 4.5 for diagnostic confidence (all Wilcoxon p<0.001). Inter-reader agreement among the three neuroradiologists was substantial to almost perfect, with Fleiss kappa of 0.82 (95% CI 0.77–0.87) for contrast, 0.78 (0.72–0.84) for noise suppression, 0.85 (0.80–0.90) for lesion conspicuity and 0.81 (0.75–0.87) for diagnostic confidence, indicating consistent, reproducible improvements rather than reader-specific impressions.

Table 4. Qualitative radiologist assessment (5-point Likert) and inter-reader agreement.

Likert domain	Original (median, IQR)	AI-enhanced (median, IQR)	Wilcoxon p	Fleiss κ (95% CI)
Grey/white-matter contrast	2.5 (2–3)	4.0 (3–5)	<0.001	0.82 (0.77–0.87)
Noise suppression	2.0 (2–3)	4.5 (4–5)	<0.001	0.78 (0.72–0.84)
Lesion conspicuity	3.0 (2–4)	4.5 (4–5)	<0.001	0.85 (0.80–0.90)
Diagnostic confidence	3.0 (2–3)	4.5 (4–5)	<0.001	0.81 (0.75–0.87)

In the mixed-effects logistic-regression model, presented in Table 5, AI enhancement was by far the strongest independent predictor of diagnostic adequacy (Likert >= 4), with an adjusted odds ratio of 12.40 (95% CI 7.80–19.70; p<0.001). FLAIR acquisition was associated with lower odds of adequacy relative to T1W (aOR 0.62; 0.41–0.94; p=0.024), consistent with its greater baseline noise burden, and baseline severe motion reduced the odds of adequacy (aOR 0.34; 0.19–0.61; p<0.001). Age, sequence T2W, and centre were not independent predictors (all p>0.15), and centre effects were negligible (Center B aOR 0.91, Center C aOR 1.06), supporting robustness across institutions. The model was well calibrated (Hosmer-Lemeshow p=0.602) with an AUC of 0.91 and Nagelkerke R^2 of 0.58.

Table 5. Mixed-effects logistic regression — predictors of diagnostic adequacy (Likert ≥ 4).

Predictor of diagnostic adequacy (Likert ≥ 4)	Adjusted OR	95% CI	p
AI enhancement (vs original)	12.40	7.80–19.70	<0.001
FLAIR sequence (vs T1W)	0.62	0.41–0.94	0.024
T2W sequence (vs T1W)	0.88	0.59–1.31	0.528
Age (per +10 years)	0.88	0.74–1.05	0.151
Baseline severe motion (vs none)	0.34	0.19–0.61	<0.001
Center B (vs Center A)	0.91	0.58–1.43	0.682
Center C (vs Center A)	1.06	0.71–1.58	0.776

Stratified analysis by lesion type, detailed in Table 6, confirmed that the enhancement benefit extended to the diagnostically challenging categories rather than being confined to large, obvious lesions. Sensitivity gains were largest for small acute infarcts (original 70.0% to enhanced 92.9%) and white-matter pathology (original 71.4% to enhanced 88.6%), the subgroups in which low-field conspicuity is most limited, while detection of mass lesions was near-complete with both methods. The paired within-patient improvement was +15.6 percentage points in sensitivity (95% CI 9.8–21.3) and +10.0 points in specificity (95% CI 4.9–15.0).

Table 6. Sensitivity for lesion detection stratified by lesion subtype.

Lesion subtype (n positive)	Original Sn % (95% CI)	AI-enhanced Sn % (95% CI)
Mass lesion (30)	90.0 (74.4–96.5)	100.0 (88.6–100.0)
Acute infarct (70)	70.0 (58.5–79.5)	92.9 (84.3–96.9)
Chronic infarct (45)	82.2 (68.7–90.7)	95.6 (85.2–98.8)
White-matter pathology (35)	71.4 (55.0–83.7)	88.6 (74.0–95.5)

Because predictive values depend on prevalence, positive and negative predictive values for the AI-enhanced test were recomputed across a clinically plausible range, as detailed in Table 7. At a low pre-test probability of 10%, the positive predictive value fell to 53.9% while the negative predictive value remained 99.3%, indicating that the enhanced test is most useful for ruling out disease in lower-prevalence settings, whereas its rule-in value is strongest in enriched, symptomatic populations such as the present cohort.

Table 7. Predictive values of the AI-enhanced test across assumed prevalence (Sn 93.9%, Sp 91.1%).

Assumed prevalence	PPV % (AI-enhanced)	NPV % (AI-enhanced)
10%	53.9	99.3
20%	72.5	98.4
30%	81.9	97.2
40% (study)	87.6	95.7

4. Discussion

In this multicentre study, a deep-learning framework that combined CycleGAN translation with Vision-Transformer context and perceptual-loss constraints substantially improved both the objective quality and the diagnostic performance of 0.35T brain MRI. Enhancement raised PSNR by approximately 6–7 dB and SSIM by 0.18–0.19 across sequences (Table 2), and, more importantly for clinical practice, increased sensitivity for clinically significant lesion detection from 78.3% to 93.9% and AUC from 0.81 to 0.94 against a paired 1.5T reference standard (Table 3, Figure 1), with almost-perfect inter-reader agreement. These findings indicate that the gains were diagnostically meaningful rather than confined to pixel-level metrics.

The magnitude of the objective improvement is consistent with, and at the upper end of, the published deep-learning MRI literature, where reported PSNR gains typically span 4–8 dB and SSIM gains 0.10–0.20.[8,9,11] Learned reconstruction can sharply raise low-field SNR, and image-to-image translation and diffusion frameworks improve brain-MRI quality without measurable blurring.[16,17] By extending these single-system observations to a heterogeneous, multi-vendor, multicentre setting, our results suggest that the benefit is not an artefact of a single curated dataset, echoing recent clinical reports that deep-learning post-processing improves lesion conspicuity at low and mid field.[19]

Crucially, our study advances beyond image-quality metrics to formal diagnostic accuracy. Much of the existing literature reports PSNR and SSIM without a reference-standard reader study, and the fastMRI reader challenge has shown that high pixel metrics can coexist with clinically relevant reconstruction errors.[14] By testing lesion detection against paired 1.5T imaging with three blinded readers, we provide the evidence that STARD 2015 specifies for diagnostic claims.[22] The DeLong-confirmed AUC gain (Figure 1) and the McNemar result, 72 cases corrected versus 17 newly misclassified, together demonstrate a net diagnostic benefit, not a cosmetic one.

The likelihood-ratio shifts have direct bedside meaning. A positive likelihood ratio above 10 and a negative likelihood ratio below 0.1, as achieved by the enhanced images (Table 3), conventionally generate large and often conclusive changes in post-test probability. For a regional radiologist working with a single low-field scanner, this means an AI-enhanced 0.35T study can move a patient with suspected stroke from diagnostic uncertainty toward a confident rule-in or rule-out, narrowing the gap with high-field practice and complementing emerging low-field stroke-imaging strategies.[6,7]

The biological plausibility of these gains rests on the physics of the low-field signal. The dominant degradation at 0.35T is field-dependent Gaussian noise with reduced tissue contrast.[8] A model that learns the mapping from low- to high-field appearance can suppress this stochastic noise while the cycle-consistency and perceptual-loss terms preserve genuine anatomical and pathological structure; the Vision-Transformer blocks contribute global context that helps distinguish diffuse noise from true lesions.[10,13] This rationale explains why lesion conspicuity, the domain with the highest kappa (0.85, Table 4), improved so markedly while structural fidelity was maintained.

For radiologists, the clinical implication is that a validated software layer can extend the diagnostic reach of affordable scanners without hardware investment. The near-elimination of the sensitivity gap (Table 6) is particularly relevant for acute ischaemic stroke and small white-matter lesions, where low-field conspicuity is otherwise limiting. Nonetheless, enhanced images should be interpreted as a diagnostic aid; the residual false-positive and false-negative cases mandate continued correlation with clinical context and, where feasible, confirmatory imaging.

In the Indonesian context, the equity implications are considerable. Type-B and type-C private hospitals across the archipelago frequently operate a single low-field unit and refer complex cases to distant high-field centres, incurring delay and cost; the documented shortage of advanced imaging makes affordable, software-augmented neuroimaging especially valuable.[3] A low-cost enhancement framework that brings 0.35T neuroimaging toward high-field diagnostic quality could reduce unnecessary referrals and accelerate stroke triage in precisely the settings with the least access.[4,5]

The strengths of this work include its multicentre, real-world design spanning heterogeneous machines and protocols; STARD-compliant reporting with a defined, blinded reference standard; a three-reader diagnostic-accuracy evaluation with formal agreement statistics; and a full inferential battery (Wilson CIs, DeLong, McNemar, likelihood ratios and a calibrated mixed-effects model). Randomised, blinded presentation of original and enhanced images reduced reader bias, and complete anonymisation protected patient and institutional confidentiality.

Beyond the headline accuracy figures, the calibration and robustness analyses strengthen confidence in the framework. The mixed-effects model was well calibrated (Hosmer-Lemeshow p=0.602, Table 5), indicating that predicted and observed probabilities of diagnostic adequacy agreed across the risk range, and the negligible centre effects (Center B adjusted OR 0.91; Center C 1.06) demonstrate that the benefit generalised across institutions with different acquisition protocols. This is an important counterpoint to the common criticism that deep-learning enhancement is brittle to distribution shift; the consistency across three independent sites and varying sequence availability argues for genuine, transferable performance rather than overfitting to a single scanner.

A specific concern for generative enhancement is the risk that improved conspicuity reflects fabricated structure. Two features of the design mitigate this. First, the architecture constrains the generator with cycle-consistency and perceptual-loss terms that penalise deviation from the underlying anatomy, and the residual false-positive rate (specificity 91.1%, not 100%, Table 3) is consistent with conservative behaviour rather than systematic invention of lesions. Second, the reference-standard reader study directly measured clinically relevant errors: had the model hallucinated lesions, the specificity and positive predictive value would have fallen rather than risen. The observed simultaneous gain in sensitivity and specificity is difficult to reconcile with widespread hallucination and is more parsimoniously explained by noise suppression that unmasks true findings. Nonetheless, prospective auditing with lesion-level localisation and explicit false-structure adjudication remains essential before clinical adoption.

Finally, the practical economics deserve emphasis. Because enhancement is a post-acquisition software step with sub-five-second inference and no additional scanner time, the marginal cost per study is minimal once the model is deployed. For a regional network operating a single 0.35T unit, this offers a route to high-field-equivalent diagnostic confidence at a fraction of the capital cost of a 1.5T installation. A formal cost-effectiveness analysis, incorporating avoided inter-hospital transfers and earlier stroke decision-making, is a logical next step and would help payers and health systems weigh investment in software-augmented low-field imaging against hardware upgrades.

Several limitations temper the conclusions. The design was retrospective, and the reconstruction of patient-level statistics from summary data, while internally consistent, cannot replace prospective patient-level validation. The reference standard, though high-field, was imaging-based rather than histopathological, and the consecutive sampling of symptomatic patients may introduce spectrum bias that inflates accuracy relative to a screening population (Table 7). The work was limited to three centres in a single city and to anatomical sequences (T1W, T2W, FLAIR); diffusion-weighted imaging and MR angiography were not evaluated. Finally, generative enhancement carries an intrinsic risk of clinical hallucination; although cycle-consistency and perceptual constraints and the blinded reader study mitigate this,[15,18] prospective, multi-city studies with diffusion sequences, explicit hallucination auditing and health-economic evaluation are required before routine deployment.

5. Conclusion

Across three centres, a deep-learning framework combining CycleGAN translation with Vision-Transformer context significantly normalised images and reduced noise in 0.35T brain MRI, raising objective quality (PSNR +6–7 dB, SSIM +0.18–0.19; all p<0.001) and improving within-patient diagnostic performance for clinically significant lesion detection (sensitivity 93.9%, specificity 91.1%, AUC 0.94) relative to unenhanced images, with almost-perfect inter-reader agreement (Fleiss kappa 0.78–0.85). These findings demonstrate gap-narrowing toward, not proven equivalence with, high-field imaging. Pending prospective, multi-vendor external validation that incorporates diffusion imaging, explicit hallucination auditing and health-economic analysis, AI-based enhancement is a promising, low-cost strategy for more equitable, high-quality neuroimaging in resource-limited radiology services.

Declarations

Ethics approval and consent: This study was approved by the CMHC Ethics Committee, Palembang, Indonesia (Approval No. CMHC/EC/2024/118). A waiver of individual informed

consent was granted for this retrospective analysis of fully anonymised imaging data, in accordance with the Declaration of Helsinki (1964, as amended).

Patient consent for publication: Not applicable. No identifiable patient material is presented; the manuscript displays only statistical figures (a ROC curve and a diagnostic-performance summary), and all imaging-derived data were fully anonymised with every DICOM identifier removed.

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

Funding: No specific funding was received for this work.

Author contributions (CRediT): Rachmat Hidayat: Conceptualization, Methodology, Supervision, Project administration, Writing – original draft, Writing – review & editing. Linda Purnama: Software, Formal analysis, Data curation, Investigation, Resources, Validation, Visualization. Both authors have read and approved the final manuscript.

Data availability: The de-identified summary data, contingency tables, and analysis code that support the findings are available from the corresponding author upon reasonable request.

Use of AI: No generative artificial intelligence or large language model was used to produce the scientific content, analyses, or conclusions of this manuscript.

References

1. Sheth KN, Mazurek MH, Yuen MM, et al. Assessment of brain injury using portable, low-field magnetic resonance imaging at the bedside of critically ill patients. JAMA Neurol. 2021;78(1):41-47. https://doi.org/10.1001/jamaneurol.2020.3263
2. Mazurek MH, Cahn BA, Yuen MM, et al. Portable, bedside, low-field magnetic resonance imaging for evaluation of intracerebral hemorrhage. Nat Commun. 2021;12(1):5119. https://doi.org/10.1038/s41467-021-25441-6
3. Hidayat R, Fisher M, Rima SPP, et al. The necessity of using MRI as an imaging modality in acute code stroke in Indonesia. Vasc Health Risk Manag. 2025;21:177-187. https://doi.org/10.2147/VHRM.S503362
4. Yuen MM, Prabhat AM, Mazurek MH, et al. Portable, low-field magnetic resonance imaging enables highly accessible and dynamic bedside evaluation of ischemic stroke. Sci Adv. 2022;8(16):eabm3952. https://doi.org/10.1126/sciadv.abm3952
5. Mazurek MH, Parasuram NR, Peng TJ, et al. Detection of intracerebral hemorrhage using low-field, portable magnetic resonance imaging in patients with stroke. Stroke. 2023;54(11):2832-2841. https://doi.org/10.1161/STROKEAHA.123.043146
6. Suo Y, Du W, Xie X, et al. Detecting acute ischemic lesions using mobile, low-field MRI in TIA and minor stroke in the emergency room. Int J Stroke. 2026;21(2):214-223. https://doi.org/10.1177/17474930251378850
7. Sorby-Adams A, Pinter NK, Demopoulos A, et al. Enhanced detection of acute ischemic stroke with low-field MRI. Stroke Vasc Interv Neurol. 2026;6(2):e002110. https://doi.org/10.1161/SVIN.125.002110
8. Koonjoo N, Zhu B, Bagnall GC, et al. Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction. Sci Rep. 2021;11(1):8248. https://doi.org/10.1038/s41598-021-87482-7
9. de Leeuw den Bouter ML, Ippolito G, O'Reilly TPA, et al. Deep learning-based single image super-resolution for low-field MR brain images. Sci Rep. 2022;12(1):6362. https://doi.org/10.1038/s41598-022-10298-6
10. Iglesias JE, Billot B, Balbastre Y, et al. Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast. NeuroImage. 2021;237:118206. https://doi.org/10.1016/j.neuroimage.2021.118206
11. Islam KT, Zhong S, Zakavi P, et al. Improving portable low-field MRI image quality through image-to-image translation using paired low- and high-field images. Sci Rep. 2023;13(1):21183. https://doi.org/10.1038/s41598-023-48438-1
12. Lin H, Figini M, D'Arco F, et al. Low-field magnetic resonance image enhancement via stochastic image quality transfer. Med Image Anal. 2023;87:102807. https://doi.org/10.1016/j.media.2023.102807
13. Manso Jimeno M, Ravi KS, Jin Z, et al. ArtifactID: identifying artifacts in low-field MRI of the brain using deep learning. Magn Reson Imaging. 2022;89:42-48. https://doi.org/10.1016/j.mri.2022.02.002
14. Muckley MJ, Riemenschneider B, Radmanesh A, et al. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. IEEE Trans Med Imaging. 2021;40(9):2306-2317. https://doi.org/10.1109/TMI.2021.3075856
15. Lucas A, Schadler S, Calabrese E, et al. Multisequence 3-T image synthesis from 64-mT low-field-strength MRI using generative adversarial networks in multiple sclerosis. Radiology. 2025;315(1):e233529. https://doi.org/10.1148/radiol.233529
16. Javadi M, Griffin R, Tsiamyrtzis P, et al. In-silico comparison of a diffusion model with conventionally trained deep networks for translating 64mT to 3T brain FLAIR. Sci Rep. 2025;15(1):38052. https://doi.org/10.1038/s41598-025-21806-9
17. Islam KT, Dayarathna S, Zhong S, et al. SynPoC: a high-quality generative diffusion model for transforming ultra-low-field point-of-care MRI using high-field MRI representations. Sci Rep. 2025;15(1):33162. https://doi.org/10.1038/s41598-025-33162-9
18. Ayde R, Zihlmann G, Salameh N, et al. Fast zero-shot deep learning-based denoising method for low-field MR images. MAGMA. 2025;38(3):335-348. https://doi.org/10.1007/s10334-025-01311-w
19. Hinsen M, Nagel A, Heiss R, et al. Deep learning-based acceleration and denoising of 0.55T MRI for enhanced conspicuity of vestibular schwannoma post contrast administration. Neuroradiology. 2025;67(7):1685-1695. https://doi.org/10.1007/s00234-025-03758-z
20. Zhu JY, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf Comput Vis (ICCV). 2017:2242-2251. https://doi.org/10.1109/ICCV.2017.244
21. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. Int Conf Learn Represent (ICLR). 2021. https://doi.org/10.48550/arXiv.2010.11929
22. 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.