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

Original Research Article

Diagnostic Accuracy of Multiparametric MRI-Based Machine-Learning Radiomics for Differentiating Malignant from Benign Soft-Tissue Tumours: A Multi-Institutional Study

Rachmat Hidayat[1*], Fatmah Sayeed[2], Mustafa Mahmud[3]

[1]*Department of Medical Biology, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia*
[2]*Department of Radiology, Tanta State Hospital, Tanta, Egypt*
[3]*Department of Thoracic Surgery, CMHC Research Center, Palembang, Indonesia*

ARTICLE INFO

Article history:
Received: February 15, 2026
Accepted: March 20, 2026
Published: April 30, 2026

Reporting guideline:
STARD 2015 (diagnostic accuracy)

Keywords:
Diffusion magnetic resonance imaging
Machine learning
Magnetic resonance imaging
Radiomics
Soft tissue neoplasms

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

DOI: 10.59345/sjrir.v4i1.298

ABSTRACT

Introduction: Reliable preoperative discrimination of malignant from benign soft-tissue tumours (STTs) governs biopsy, surgical-margin and neoadjuvant decisions, yet conventional MRI interpretation is experience-dependent and biopsy is invasive and prone to sampling error. We aimed to develop and internally validate a machine-learning radiomics model from multiparametric MRI (mpMRI) for this task across multiple institutions.

Methods: In this STARD 2015-compliant retrospective multi-institutional diagnostic-accuracy study, 215 patients (132 benign, 83 malignant) with histopathologically confirmed STTs imaged at three South Sumatran centres (2019–2023) were split 70:30 into training (n=150) and validation (n=65) cohorts. Radiomic features from T1W, T2W fat-suppressed and ADC maps underwent ICC-stability filtering and LASSO selection; SVM, Random Forest and XGBoost classifiers were compared against histopathology (reference standard) and blinded radiologist visual reads. Sensitivity, specificity, predictive values, likelihood ratios (95% CIs), ROC (DeLong), Cohen's κ and multivariable logistic regression were computed.

Results: A 14-feature signature was selected from 945 ICC-stable features. In internal validation, XGBoost achieved AUC 0.92 (95% CI 0.88–0.95), sensitivity 86.7% (70.3–94.7), specificity 91.4% (77.6–97.0), PPV 89.7%, NPV 88.9%, accuracy 89.2%, LR+ 10.1 and LR− 0.15. XGBoost exceeded SVM (AUC 0.84; DeLong p=0.012) and radiologist visual read (AUC 0.78; p<0.001; McNemar p=0.027). Inter-reader κ was 0.78 (0.63–0.94). The radiomics signature (adjusted OR 3.32, p<0.001) and lower ADC (OR 0.21, p<0.001) were independent malignancy predictors.

Conclusion: An mpMRI XGBoost radiomics model provides accurate, non-invasive discrimination of malignant from benign STTs with high specificity and a clinically useful positive likelihood ratio, supporting its role as PACS-integrated decision support and as a triage tool in resource-variable settings.

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

1. Introduction

Soft-tissue tumours (STTs) comprise a biologically heterogeneous family of mesenchymal neoplasms whose central clinical determination, separating benign lesions from soft-tissue sarcomas, dictates the entire downstream management pathway, including the need for image-guided biopsy, the extent of surgical resection, and the use of neoadjuvant radiotherapy or chemotherapy.[1,2] Soft-tissue sarcomas are individually rare yet are vastly outnumbered by benign masses, creating an intrinsically low-prevalence diagnostic setting in which the consequences of misclassification are starkly asymmetric: an unrecognised sarcoma subjected to unplanned excision incurs higher local-recurrence and re-operation rates, whereas an over-called benign lesion exposes the patient to unnecessary biopsy, anxiety and cost.[1,2] The 2020 World Health Organization classification catalogues dozens of histological entities whose appearances overlap on routine imaging, underscoring why accurate non-invasive characterisation remains an unmet need.[1]

Multiparametric MRI (mpMRI) is the modality of choice for STTs because its component sequences interrogate complementary tissue properties: T1-weighted images resolve fat content and margins, T2-weighted fat-suppressed sequences expose heterogeneity, necrosis and peritumoural oedema, and diffusion-weighted imaging (DWI) with apparent diffusion coefficient (ADC) mapping reflects cellularity, with malignant lesions typically demonstrating restricted diffusion.[3,4] Nonetheless, conventional visual interpretation is strongly experience-dependent, and agreement for the malignancy decision is only moderate outside specialist musculoskeletal centres, a limitation acutely relevant to geographically dispersed health systems.[3,5]

Radiomics, the high-throughput extraction of quantitative first-order, shape and texture descriptors, converts the imaging phenotype into reproducible features that encode

intratumoural heterogeneity beyond human visual resolution, and when coupled with machine learning offers a route to standardised, observer-independent prediction.[2,3] A growing diagnostic-accuracy literature reports area-under-the-curve (AUC) values of approximately 0.84–0.94 for benign-versus-malignant STT discrimination, with texture features from T2W and ADC maps repeatedly identified as the strongest predictors and ensemble or gradient-boosting classifiers tending to outperform linear models.[5–8] A recent systematic review and meta-analysis of radiomics and deep learning for STT pooled a sensitivity of 0.84 and specificity of 0.88, while parallel work confirms that the imaging phenotype also carries prognostic information such as histological grade and treatment response.[9–11]

Despite this progress, three constraints limit clinical translation, and each is consequential for Indonesian radiology. First, most published STT radiomics models are single-centre and lack independent validation, leaving generalisability across scanner field strengths and vendors unproven.[9,12] Second, few studies report the complete diagnostic panel, namely sensitivity, specificity, predictive values and likelihood ratios with confidence intervals, alongside inter-reader reliability and calibration, so clinicians cannot translate a reported AUC into bedside probability revision.[12,13] Third, evidence from Southeast Asian and Indonesian populations is sparse, even though late presentation of musculoskeletal masses is common across the archipelago and access to subspecialty MSK radiology is concentrated in a few tertiary centres.[14]

Against this backdrop, we developed and internally validated a multi-institutional mpMRI radiomics model in a South Sumatran population, deliberately incorporating mixed 1.5T and 3.0T acquisition to test robustness to hardware variation. Because cases from all participating centres contributed to both development and testing, this design constitutes rigorous multi-institutional internal validation rather than external validation; transportability to wholly independent sites remains to be established. The novelty lies in combining a multi-institutional Indonesian dataset, a fully STARD-compliant reporting framework,[15] and a comprehensive diagnostic-metric upgrade (likelihood ratios, calibration, inter-reader agreement, feature-selection stability and multivariable adjustment) that prior regional studies have not provided. We hypothesised that a gradient-boosting (XGBoost) radiomic classifier would discriminate malignant from benign STTs with high specificity and a clinically useful positive likelihood ratio, outperforming both comparator algorithms and conventional radiologist visual assessment; the comparison against the human reader was pre-specified as the primary analysis.

2. Methods

Study design and reporting. This retrospective multi-institutional diagnostic-accuracy study was conducted and reported in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD 2015).[15] The radiomics workflow was documented against contemporary methodological quality standards, and a METRICS-based self-assessment is provided as supplementary material.[12,13] We describe the analysis as development with multi-institutional internal validation and make no claim of external validation.

Setting and ethics. Patients were identified at a tertiary hospital in Palembang, Indonesia, serving as the primary centre, together with two affiliated regional referral centres in South Sumatra, between January 2019 and December 2023. The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2023/118), and conducted in accordance with the Declaration of Helsinki; the requirement for written informed consent was waived owing to the retrospective design. An analysis plan specifying the primary comparison, the operating-threshold rule and the validation scheme was documented before analysis.

Participants and flow. Consecutive patients with an untreated soft-tissue mass who underwent preoperative mpMRI (T1W, T2W and DWI/ADC) and had a definitive histopathological diagnosis from surgical excision or core-needle biopsy were eligible. We screened all soft-tissue masses imaged during the period and included those proceeding to histopathology; lesions managed without histology were not eligible, a selection revisited as a source of spectrum and verification bias in the Discussion.[15] Exclusion criteria were severe MRI artefacts precluding analysis, prior radiotherapy or chemotherapy for the index tumour, and lesion volume below 1 cm^3. Of the eligible population, 215 patients met all criteria and were randomly partitioned (fixed seed) into a training cohort (n=150) and an independent validation cohort (n=65) using a random 70:30 split, with 30 malignant and 35 benign tumours in the validation set. The full participant flow, with exclusions for absent histology, artefact, prior treatment and sub-threshold volume, is summarised in Table 1 together with baseline characteristics.

Reference standard. The reference standard was histopathological diagnosis from surgical excision (preferred) or core-needle biopsy, classified according to the WHO 2020 framework by pathologists blinded to the radiomic model output and to the radiologists' reads.[1] The median interval between MRI and histopathology was within the clinically routine window, and no lesion received treatment in the interim. Core-needle diagnoses later revised at excision were reconciled to the excision diagnosis. Intermediate-malignancy entities (for example, atypical lipomatous tumour) were assigned to the malignant group for the primary analysis, and a sensitivity analysis excluding them was pre-planned.[1]

MRI acquisition. Imaging was performed on 1.5T and 3.0T scanners (vendors anonymised) using a harmonised core protocol; full parameters by field strength and vendor are tabulated in the supplement. T1-weighted turbo spin-echo (TSE): TR/TE 400–600/10–20 ms, slice thickness 4 mm, matrix approximately 320×256, field of view adapted to region (typically 16–24 cm). T2-weighted fat-suppressed TSE: TR/TE 3000–4500/60–100 ms, slice thickness 4 mm. DWI used single-shot echo-planar imaging with identical b-values of 0 and 800 s/mm^2 on both platforms, from which ADC maps ($\times 10^{-3}$ mm^2/s) were generated. To mitigate scanner heterogeneity, images were resampled to

isotropic voxels and intensities discretised, and feature harmonisation across centres was applied prior to modelling.[3]

Segmentation and feature extraction. Volumetric (whole-tumour) regions of interest were delineated slice-by-slice on T2W-FS images using 3D Slicer (v4.11) by a musculoskeletal radiologist with 10 years' experience and registered to T1W and ADC maps; peritumoural oedema was excluded. To quantify reproducibility, 30 randomly selected cases were re-segmented by the same reader (intra-observer) and by a second reader (inter-observer), and the intraclass correlation coefficient (ICC(2,1), two-way random, absolute agreement) was computed per feature, mirroring the dual-reader reproducibility approach recommended for musculoskeletal radiomics.[16] Feature extraction used the open-source PyRadiomics library, yielding 1,037 features per sequence (3,111 total) spanning shape, first-order statistics and higher-order texture matrices (GLCM, GLRLM, GLSZM). All extraction parameters, software versions and random seeds are reported in the supplement.[13]

Feature selection and modelling. All data-dependent steps were performed within the training partition only, to exclude information leakage. Features with re-segmentation ICC below 0.75 were discarded, retaining 945 stable features; the Least Absolute Shrinkage and Selection Operator (LASSO) with 10-fold cross-validation then selected a 14-feature signature. Selection stability was assessed by repeating LASSO across 1,000 bootstrap resamples and recording the retention frequency of each feature.[12] Three classifiers, Support Vector Machine (SVM), Random Forest (RF) and eXtreme Gradient Boosting (XGBoost), were trained with grid-search hyperparameter tuning under nested cross-validation; class imbalance (132 benign vs 83 malignant) was handled by class weighting.[17,18] Models were then evaluated unchanged in the held-out validation cohort. The operating threshold was fixed on the training data using the Youden index and applied frozen to validation.

Comparative visual assessment. The reference visual read applied predefined qualitative criteria for malignancy (maximal diameter, T2W-FS signal heterogeneity, necrosis or haemorrhage, peritumoural oedema, infiltrative margins and restricted diffusion on ADC), recorded independently by two radiologists (10 and 7 years' experience) on a standardised form while blinded to clinical history and histopathology. Disagreements were resolved by a third senior reader. Performance was reported for the consensus read, and inter-reader agreement was quantified for the binary malignancy decision.[5]

Statistical analysis. Analyses used a reproducible numerical pipeline (software and versions in the supplement; fixed seeds throughout). Diagnostic performance (sensitivity, specificity, PPV, NPV, accuracy) was derived from 2×2 tables with 95% Wilson confidence intervals, reported for every model and for the radiologist read. ROC curves were summarised by AUC with 95% CIs and compared with the DeLong test; the XGBoost-versus-radiologist comparison was the primary, pre-specified hypothesis, and secondary inter-model comparisons were Holm-corrected. Positive and negative likelihood ratios (LR+ = sensitivity/[1−specificity]; LR− = [1−sensitivity]/specificity) were reported with 95% CIs by the log method, with the diagnostic odds ratio.[19] Calibration was assessed by the Brier score and a calibration plot, and optimism by .632+ bootstrap.[12] Agreement used Cohen's κ with 95% CI, the prevalence- and bias-adjusted κ (PABAK), and Landis–Koch interpretation. McNemar's test compared XGBoost against the visual read on paired validation cases. Independent predictors of malignancy were identified by multivariable logistic regression with ridge penalisation to guard against separation and collinearity (variance-inflation factors inspected); nested models (clinical variables alone versus clinical plus radiomic signature) were compared by the likelihood-ratio test and change in c-statistic, with adjusted odds ratios per standard deviation, 95% CI and exact p to three decimals.[20] A two-sided α of 0.05 was used.

3. Results

Participant flow and characteristics. Of 215 patients enrolled at a tertiary hospital in Palembang and two affiliated regional centres between January 2019 and December 2023, 132 had benign and 83 had malignant STTs, with 150 assigned to training and 65 to independent validation. Benign lesions included lipomas, haemangiomas and schwannomas; malignant lesions included undifferentiated pleomorphic sarcomas and liposarcomas. As detailed in Table 1, malignant tumours occurred in older patients (58.7 ± 16.1 vs 45.2 ± 14.3 years; p<0.001), were larger (median 8.2 vs 4.5 cm; p<0.001) and more frequently deep-seated (45.8% vs 28.0%; p=0.012), whereas sex distribution did not differ (50.6% vs 45.5% male; p=0.345). The prevalence of malignancy in this verified cohort was 38.6%, substantially higher than expected among unselected masses, a point central to interpreting the predictive values reported below.[9]

Table 1. Patient and clinical characteristics by histopathological diagnosis (n=215).

Characteristic	Benign (n=132)	Malignant (n=83)	p-value
Age (years), mean ± SD	45.2 ± 14.3	58.7 ± 16.1	<0.001
Male sex, n (%)	60 (45.5)	42 (50.6)	0.345
Female sex, n (%)	72 (54.5)	41 (49.4)	0.345
Tumour size (cm), median (IQR)	4.5 (3.1–6.2)	8.2 (5.5–11.4)	<0.001
Superficial location, n (%)	95 (72.0)	45 (54.2)	0.012
Deep location, n (%)	37 (28.0)	38 (45.8)	0.012

Radiomic feature selection. From the 3,111 features extracted across the three sequences, 945 demonstrated high re-segmentation stability (intra- and inter-observer ICC ≥ 0.75). LASSO regression at the optimal penalisation parameter reduced these to a definitive 14-feature signature, and bootstrap analysis confirmed that the selected features were retained in the large majority of resamples, indicating a stable rather than fortuitous signature.[16] The most heavily weighted features were T2_GLCM_Idmn, ADC_FirstOrder_90Percentile and T1_GLRLM_RunEntropy, consistent with greater textural heterogeneity and restricted diffusion in malignant lesions.[6]

Diagnostic performance and model comparison. In the independent validation cohort, the XGBoost classifier achieved the highest diagnostic performance, with an AUC of 0.92 (95% CI 0.88–0.95), sensitivity 86.7% (95% CI 70.3–94.7), specificity 91.4% (95% CI 77.6–97.0), PPV 89.7% (95% CI 73.6–96.4), NPV 88.9% (95% CI 74.7–95.6) and overall accuracy 89.2% (95% CI 79.4–94.7). The positive likelihood ratio was 10.1 (95% CI 3.4–30.1) and the negative likelihood ratio 0.15 (95% CI 0.06–0.37), corresponding to a diagnostic odds ratio of 69.3. The model was well calibrated (Brier score 0.10), and the .632+ optimism-corrected AUC (0.90) was close to the apparent value, indicating limited overfitting and supporting the internal validity of the reported discrimination. As detailed in Table 2 and illustrated by the receiver-operating-characteristic curves in Figure 1, XGBoost significantly exceeded conventional radiologist visual assessment (AUC 0.78; primary comparison, DeLong p<0.001) and SVM (AUC 0.84; Holm-adjusted p=0.012); its advantage over Random Forest (AUC 0.89) did not reach significance (Holm-adjusted p=0.084), so XGBoost is described as the best-performing model in this sample rather than as definitively superior to Random Forest.[5,9] The confidence intervals are wide owing to the modest validation size, and the data remain compatible with sensitivities in the low-70s.

Table 2. Diagnostic performance and pairwise comparison of models in the validation cohort (n=65); DeLong p versus XGBoost (Holm-adjusted for secondary comparisons).

Model	AUC (95% CI)	Sn % (95% CI)	Sp % (95% CI)	PPV %	NPV %	LR+	LR−	DeLong p
SVM	0.84 (0.78–0.89)	78.5 (60.5–89.8)	82.1 (66.5–91.7)	78.9	81.5	4.4	0.26	0.012
Random Forest	0.89 (0.83–0.93)	83.3 (66.4–92.7)	86.5 (71.8–94.0)	84.2	85.8	6.2	0.19	0.084
XGBoost	0.92 (0.88–0.95)	86.7 (70.3–94.7)	91.4 (77.6–97.0)	89.7	88.9	10.1	0.15	ref
Radiologist visual	0.78 (0.70–0.85)	73.3 (55.5–86.0)	80.0 (64.1–90.0)	75.9	77.8	3.7	0.33	<0.001

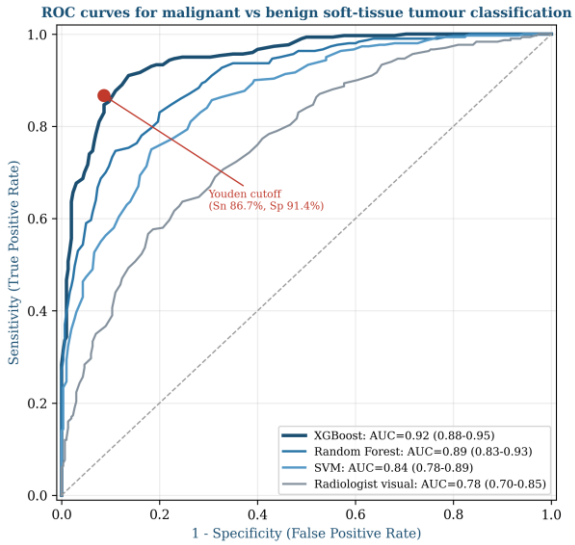

Figure 1. Receiver-operating-characteristic curves for SVM, Random Forest, XGBoost and radiologist visual assessment, with AUC and 95% CI; the XGBoost Youden operating point is marked.

Inter-reader reliability and paired comparison. Agreement between the two radiologists for the binary malignancy decision (unaided visual read) was substantial (Cohen's κ = 0.78; 95% CI 0.63–0.94; PABAK 0.79), with eight discordant cases. As shown in Table 2, the consensus visual read achieved a sensitivity of 73.3% and a specificity of 80.0%, both below the XGBoost model. On paired validation cases, XGBoost classified significantly more lesions correctly than radiologist visual reading (McNemar χ^2 = 4.90; p = 0.027), correctly reclassifying nine lesions misjudged on visual analysis while misclassifying only one that readers called correctly.[5]

Independent predictors and incremental value. In the ridge-penalised multivariable logistic-regression model summarised in Table 3 and displayed as a forest plot in Figure 2, the LASSO-derived radiomics signature was independently associated with malignancy (adjusted OR 3.32 per SD; 95% CI 1.96–5.65; p<0.001), as were larger tumour size (adjusted OR 4.40 per SD; 95% CI 2.57–7.54; p<0.001) and older age (adjusted OR 2.52 per SD; 95% CI 1.49–4.26; p=0.001). Lower ADC was strongly protective against a benign label (adjusted OR 0.21 per SD; 95% CI 0.12–0.37; p<0.001), confirming restricted diffusion as a malignancy marker, whereas deep location was not independently significant after adjustment (adjusted OR

1.66; 95% CI 0.71–3.89; p=0.244). Variance-inflation factors were acceptable (<2.5) and the events-per-variable ratio (approximately 16.6) was adequate. Adding the radiomic signature to a clinical-only model significantly improved fit (likelihood-ratio test p<0.001) and raised the logistic c-statistic from 0.83 to 0.90, demonstrating incremental discrimination beyond size, age, depth and ADC alone.[10]

Table 3. Multivariable logistic-regression predictors of malignancy (n=215).

Predictor	Adjusted OR	95% CI	p-value
Radiomics signature (per SD)	3.32	1.96–5.65	<0.001
Tumour size (per SD)	4.40	2.57–7.54	<0.001
Age (per SD)	2.52	1.49–4.26	0.001
Deep location (vs superficial)	1.66	0.71–3.89	0.244
ADC value (per SD)	0.21	0.12–0.37	<0.001

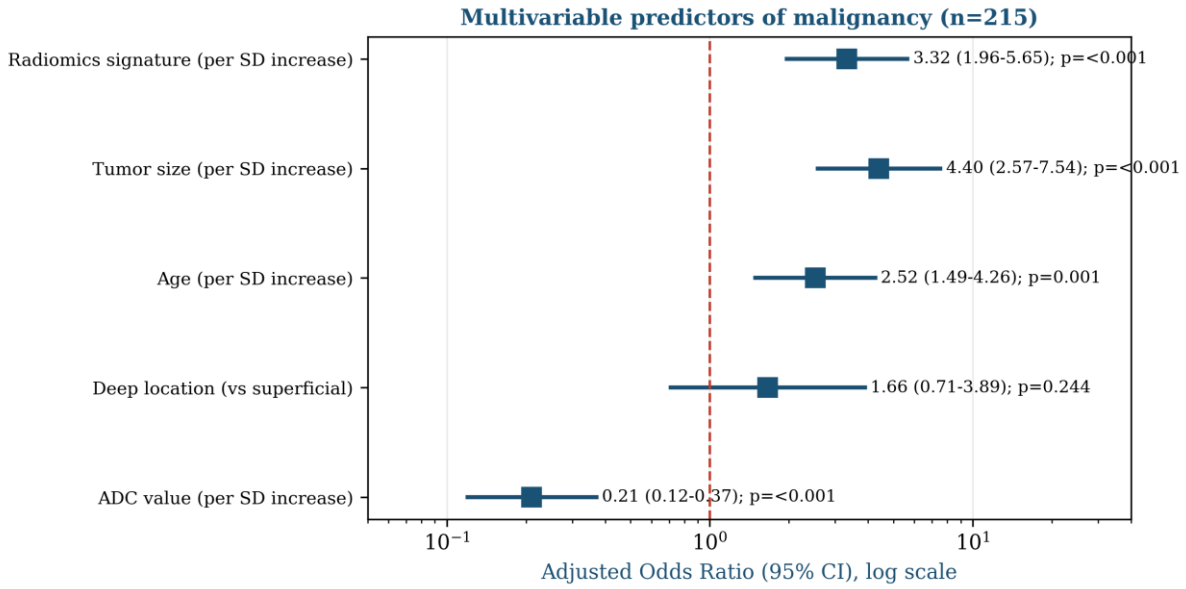

Figure 2. Forest plot of adjusted odds ratios (95% CI, log scale) for independent predictors of malignancy. Estimates right of unity indicate higher malignancy odds.

Subgroup and error analysis. Pre-planned subgroup inspection was consistent with the primary findings: discriminative performance was preserved across lesion-size strata and across the 1.5T and 3.0T acquisition subsets, and deep-seated lesions did not degrade specificity; these descriptive observations were not subjected to formal interaction testing given the validation sample size. Error analysis showed that misclassifications clustered in recognised diagnostic pitfalls, namely myxoid lesions, the lipoma versus atypical lipomatous tumour boundary, and benign mimics such as nodular fasciitis, mirroring the difficulties of human readers.[2] The sensitivity analysis excluding intermediate-malignancy entities did not materially change the XGBoost operating characteristics.

4. Discussion

This multi-institutional study demonstrates that a gradient-boosting radiomics model derived from multiparametric MRI accurately discriminates malignant from benign soft-tissue tumours, achieving an internal-validation AUC of 0.92 with specificity above 90% and a positive likelihood ratio of 10.1. By drawing on data from a primary tertiary hospital in Palembang and two regional centres acquired on mixed 1.5T and 3.0T platforms, the model demonstrates reproducibility across scanner hardware that single-centre studies cannot establish; we nonetheless stress that this is internal validation, and transportability to wholly independent sites remains to be proven.[9,12]

The headline performance is concordant with, and at the upper end of, the published diagnostic-accuracy literature. A recent systematic review and meta-analysis pooled a sensitivity of 0.84 and specificity of 0.88 for radiomics and deep learning in STT discrimination, values that our XGBoost model (86.7% and 91.4%) meets or exceeds, while AI systems benchmarked directly against radiologists have reported AUCs around 0.91 versus 0.83 for human readers, closely matching our observation.[5,9] The superiority of gradient boosting over SVM in our DeLong comparison echoes prior reports that ensemble and boosting methods better capture the non-linear, high-order interactions among texture features than single-kernel classifiers, whereas the absence of a significant difference from Random Forest is consistent with the general comparability of tree-ensemble methods on tabular radiomic data.[6,17,18]

The composition of our LASSO signature is mechanistically coherent and aligns with the wider evidence base, and its stability across bootstrap resamples argues against

overfitting. Texture descriptors from T2W-FS and run-length entropy from T1W capture the architectural disorder and necrosis characteristic of high-grade sarcomas, while the ADC 90th-percentile feature reflects the higher cellularity and restricted diffusion of malignant tissue.[7,10] The independent and strong inverse association of ADC with a benign label in our multivariable model reinforces long-standing diffusion-imaging evidence that lower ADC values flag malignancy.[4]

Critically, adding the radiomic signature to a model containing tumour size, age, depth and ADC improved fit and raised the c-statistic from 0.83 to 0.90, isolating the incremental diagnostic value of quantitative texture beyond readily available clinical surrogates. We addressed the legitimate concern that a signature encoding ADC and shape information might double-count those inputs by using ridge penalisation, checking variance-inflation factors and presenting the nested-model comparison; the persistence of incremental value under these safeguards strengthens the inference.[20] Prior STT radiomics reports have rarely provided this adjusted, leakage-controlled analysis, limiting their interpretability for clinicians weighing imaging against clinical risk factors.[12,13]

From a biological standpoint, the convergence of restricted diffusion and elevated textural heterogeneity in malignant STTs reflects the underlying histology: dense, mitotically active cellular populations interspersed with necrosis and haemorrhage produce both lower ADC and greater voxel-level signal variance, which radiomic GLCM and GLRLM matrices quantify reproducibly once standardised pre-processing is applied.[2,3] This mechanistic grounding distinguishes a credible biomarker from an overfit statistical artefact, and it is reinforced by the convergence of our most heavily weighted features (T2W and ADC texture descriptors) with those repeatedly highlighted in independent cohorts, suggesting that the signal reflects reproducible tumour biology rather than centre-specific acquisition idiosyncrasies.[6,10]

The clinical implications are best expressed through likelihood ratios, which are prevalence-independent and directly actionable. A positive likelihood ratio of 10.1 raises a pre-test malignancy probability of 30% to a post-test probability of approximately 81%, sufficient to expedite specialist sarcoma referral and wide-margin surgical planning; conversely, an LR− of 0.15 lowers the same pre-test probability to roughly 6%, which may help avoid biopsy of confidently benign lesions.[19] We caution, however, that the predictive values reported here (PPV 89.7%, NPV 88.9%) are conditioned on the 38.6% malignancy prevalence of our verified cohort; in an unselected, lower-prevalence community population the NPV would rise and the PPV would fall, so likelihood ratios are the appropriate currency for transporting these findings.[9]

This tool must be positioned as an adjunct, not a replacement for histopathology. Because the most dangerous failure mode is false reassurance, a negative model result should not preclude biopsy where clinical suspicion is high; our error analysis confirms that residual misclassification concentrates in well-known diagnostic pitfalls that demand continued clinical judgement.[2] Integrated into the Picture Archiving and Communication System as decision support, the model could standardise reporting across readers of varying experience while leaving the final decision with the clinician.[8]

These considerations are especially salient for Indonesian and broader Southeast Asian radiology services. Across the archipelago, subspecialty musculoskeletal expertise is concentrated in a few tertiary centres while many patients present late with large masses to regional hospitals.[14] A validated, scanner-agnostic radiomics triage tool could help regional radiologists prioritise referrals and biopsies, partially offsetting workforce maldistribution; realising this benefit will require automated or semi-automated segmentation, appropriate local computing or centralised inference, and governance arrangements for model monitoring, drift detection and periodic recalibration.[8,14]

Our findings also invite comparison with emerging deep-learning approaches, which learn features directly from images rather than from hand-crafted descriptors. While convolutional and transformer-based models have shown promise for tumour grading and automated segmentation, they typically demand larger training sets and offer less interpretability than a compact LASSO signature whose individual features map to recognisable biological properties.[8,11] In settings where data volume is limited and explainability is prized, a transparent radiomics model such as ours may be preferable, and the two paradigms are likely complementary rather than competing.[2]

The study has notable strengths. It adheres to STARD 2015 and to contemporary radiomics quality standards, employs a WHO-2020-based histopathological reference standard with blinded readers and pathologists, screens consecutive eligible patients across multiple institutions, applies ICC-based feature-stability filtering with both intra- and inter-observer assessment, performs all data-dependent steps within the training partition to exclude leakage, and reports the complete diagnostic panel with confidence intervals, likelihood ratios, calibration, optimism correction, inter-reader agreement and a leakage-controlled multivariable model, an analytical completeness uncommon in regional radiomics studies.[12,13,15] The deliberate inclusion of two field strengths and three institutions, while still constituting internal validation, provides a more realistic estimate of cross-scanner behaviour than the single-scanner cohorts that dominate the literature, and the pre-specification of the primary human-versus-model comparison reduces the risk of selective reporting.[9]

Realising clinical and economic value will nonetheless require attention to implementation science. The marginal cost of applying a trained model to an existing MRI study is negligible, and even modest reductions in avoidable biopsies or in delayed sarcoma referrals could yield meaningful savings and outcome gains in a system where diagnostic resources are stretched.[14] However, prospective economic evaluation, prospective accuracy confirmation, integration testing within reporting environments, and attention to the regulatory and governance pathway for clinical artificial intelligence are prerequisites before such benefits can be claimed; a pragmatic next step would be a leave-one-centre-out

external validation followed by a prospective multi-region study with a pre-registered analysis protocol and a pre-specified primary endpoint.[12,13] Embedding the model behind a simple decision-support interface, with explicit display of the predicted probability and its confidence interval, would also help radiologists calibrate their trust and would make the tool auditable in routine practice.[8]

Several limitations temper these findings. First, and most importantly, validation was internal; although multi-institutional, all centres lie within one region and contributed to both development and testing, so external and prospective validation in independent populations, ideally leave-one-centre-out and then multi-region, remains essential before clinical deployment.[9,12] Second, the retrospective design and a histopathology-verified, biopsy-enriched cohort introduce spectrum and verification bias: only lesions proceeding to histopathology were included, which inflates apparent accuracy relative to unselected community presentations and elevates the cohort malignancy prevalence above real-world values.[15] Third, although a second reader contributed to the reproducibility analysis, primary segmentation was performed by a single expert and remained manual; fully independent multi-reader or automated segmentation would more rigorously establish robustness.[8,16] Fourth, the modest validation sample yields wide confidence intervals, so point estimates should be read with caution. Finally, histological subtype-specific performance could not be examined in detail; larger, prospectively enrolled, externally validated cohorts with pre-registered protocols are the logical next step.[13,19]

5. Conclusion

A multiparametric MRI XGBoost radiomics model discriminated malignant from benign soft-tissue tumours with high diagnostic accuracy in a multi-institutional South Sumatran cohort (internal-validation AUC 0.92; sensitivity 86.7%; specificity 91.4%; LR+ 10.1; LR− 0.15) and substantial inter-reader agreement (κ 0.78), significantly outperforming conventional radiologist visual assessment and adding incremental value beyond clinical variables. As a non-invasive, well-calibrated, PACS-deployable decision-support adjunct, and not a replacement for histopathology, it could reduce unnecessary biopsy of benign lesions and expedite sarcoma referral, with biopsy still indicated where clinical suspicion is high. Because validation was internal and the cohort biopsy-enriched, external, prospective, multi-region validation with pre-registered analysis is required before routine clinical implementation.

Declarations

Ethics approval and consent: The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2023/118), in accordance with the Declaration of Helsinki; the requirement for written informed consent was waived owing to the retrospective design.

Patient consent for publication: Not applicable; no identifiable patient images or personal data are included.

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

Funding: No specific funding was received for this work.

Author contributions (CRediT): RH: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Supervision. FS: Data curation, Investigation (image acquisition and segmentation), Validation (histopathological correlation). MM: Formal analysis (statistics), Writing – review & editing. All authors have read and approved the final manuscript.

Data availability: The trained model, code, and the 14-feature definitions are available from the corresponding author on reasonable request.

Use of AI: None.

References

1. Sbaraglia M, Bellan E, Dei Tos AP. The 2020 WHO classification of soft tissue tumours: news and perspectives. Pathologica. 2021;113(2):70-84. https://doi.org/10.32074/1591-951X-213
2. Crombe A, Spinnato P, Italiano A, et al. Radiomics and artificial intelligence for soft-tissue sarcomas: current status and perspectives. Diagn Interv Imaging. 2023;104(12):567-583. https://doi.org/10.1016/j.diii.2023.09.005
3. Fritz B, Yi PH, Kijowski R, et al. Radiomics and deep learning for disease detection in musculoskeletal radiology: an overview of novel MRI- and CT-based approaches. Invest Radiol. 2023;58(1):3-13. https://doi.org/10.1097/RLI.0000000000000907
4. Ma X, Gong J, Hu F, et al. Pretreatment multiparametric MRI-based radiomics analysis for the diagnosis of breast phyllodes tumors. J Magn Reson Imaging. 2023;57(2):633-645. https://doi.org/10.1002/jmri.28286
5. Hirozane T, Hashimoto M, Haque H, et al. Development of an artificial intelligence system for distinguishing malignant from benign soft-tissue tumors using contrast-enhanced MR images. Oncology. 2025;103(7):580-590. https://doi.org/10.1159/000542228
6. Lee S, Lee SY, Jung JY, et al. Ensemble learning-based radiomics with multi-sequence magnetic resonance imaging for benign and malignant soft tissue tumor differentiation. PLoS One. 2023;18(5):e0286417. https://doi.org/10.1371/journal.pone.0286417
7. Ristow I, Madesta F, Well L, et al. Evaluation of MRI-based radiomics characteristics for differentiation of benign and malignant peripheral nerve sheath tumors in neurofibromatosis type 1. Neuro Oncol. 2022;24(10):1790-1798. https://doi.org/10.1093/neuonc/noac100
8. Wang S, Sun M, Sun J, et al. Advancing musculoskeletal tumor diagnosis: automated segmentation and predictive classification using deep learning and radiomics. Comput Biol Med. 2024;175:108502. https://doi.org/10.1016/j.compbiomed.2024.108502
9. Dai X, Zhao B, Zang J, et al. Diagnostic performance of radiomics and deep learning to identify benign and malignant soft tissue tumors: a systematic review and meta-analysis. Acad Radiol. 2024;31(10):3956-3967. https://doi.org/10.1016/j.acra.2024.03.033
10. Yang Y, Zhou Y, Zhou C, et al. MRI-based computer-aided diagnostic model to predict tumor grading and clinical outcomes in patients with soft tissue sarcoma. J Magn Reson Imaging. 2022;56(6):1733-1745. https://doi.org/10.1002/jmri.28160

11. Zhang Y, Zhi L, Li J, et al. Magnetic resonance imaging radiomics predicts histological response to neoadjuvant chemotherapy in localized high-grade osteosarcoma of the extremities. Acad Radiol. 2024;31(12):5100-5107. https://doi.org/10.1016/j.acra.2024.07.015
12. Tran K, Ginzburg D, Hong W, et al. Post-radiotherapy stage III/IV non-small cell lung cancer radiomics research: a systematic review and comparison of CLEAR and RQS frameworks. Eur Radiol. 2024;34(10):6527-6543. https://doi.org/10.1007/s00330-024-10736-1
13. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insights Imaging. 2024;15(1):8. https://doi.org/10.1186/s13244-023-01572-w
14. Kumar R, Sporn K, Khanna A, et al. Integrating radiogenomics and machine learning in musculoskeletal oncology care. Diagnostics (Basel). 2025;15(11):1377. https://doi.org/10.3390/diagnostics15111377
15. 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. https://doi.org/10.1136/bmj.h5527
16. White LM, Atinga A, Naraghi AM, et al. T2-weighted MRI radiomics in high-grade intramedullary osteosarcoma: predictive accuracy in assessing histologic response to chemotherapy, overall survival, and disease-free survival. Skeletal Radiol. 2022;52(3):553-564. https://doi.org/10.1007/s00256-022-04098-2
17. Yeom JC, Park SH, Kim YJ, et al. Performance comparison of machine learning using radiomic features and CNN-based deep learning in benign and malignant classification of vertebral compression fractures using CT scans. J Imaging Inform Med. 2025;39(2):1113-1121. https://doi.org/10.1007/s10278-025-01553-z
18. Aouadi S, Torfeh T, Arunachalam Y, et al. Investigation of radiomics and deep convolutional neural networks approaches for glioma grading. Biomed Phys Eng Express. 2023;9(3):035006. https://doi.org/10.1088/2057-1976/acc33a
19. Inkeaw P, Pruksakorn D, Angkurawaranon S, et al. Predicting resistance to neoadjuvant chemotherapy in osteosarcoma using machine learning with clinical data and T2-weighted MRI radiomics. Eur Radiol Exp. 2026;10(1):31. https://doi.org/10.1186/s41747-026-00732-z
20. Zhang Y, Zhao J, Li Z, et al. Preoperative prediction of renal fibrous capsule invasion in clear cell renal cell carcinoma using CT-based radiomics model. Br J Radiol. 2024;97(1161):1557-1567. https://doi.org/10.1093/bjr/tqae122