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

Original Research Article

Radiomics-Based Machine Learning for Automated Detection and Rupture-Risk Stratification of Cerebral Vascular Malformations: A Retrospective Cohort Study

Hesti Putri[1], Nur Diana[2*], Paula Magna Pablo-Rodriguez[3], Nadia Khoirina[4]

[1]*Department of Informatics Science, CMHC Research Center, Palembang, Indonesia*
[2]*Department of Molecular Biology, CMHC Research Center, Palembang, Indonesia*
[3]*Department of Radiology and Nuclear Medicine, Faculty of Health, Universidade do Estado do Pará, Belém, Brazil*
[4]*Department of Radiology and Nuclear Medicine, CMHC Research Center, Palembang, Indonesia*

ARTICLE INFO

Article history:
Received: August 15, 2025
Accepted: September 20, 2025
Published: October 30, 2025

Reporting guideline:
STARD 2015 (diagnostic accuracy); CLAIM; TRIPOD (prediction model)

Keywords:
Computed tomography angiography
Intracranial aneurysm
Machine learning
Magnetic resonance angiography
Radiomics

***Corresponding author:**
Nur Diana
E-mail: nurdiananurdiana353@gmail.com

DOI: 10.59345/sjrir.v3i2.291

ABSTRACT

Introduction: Cerebral vascular malformations (CVMs) are the leading cause of spontaneous intracranial haemorrhage, yet their detection on CT/MR angiography is operator-dependent and existing machine-learning models derive almost exclusively from Caucasian or East Asian cohorts. We developed and internally validated a population-specific radiomics pipeline for CVM detection and rupture-risk stratification in a Southeast Asian population.

Methods: In this retrospective diagnostic-and-prognostic cohort (STARD 2015; CLAIM) at a tertiary hospital in Palembang, Indonesia (2020–2025), 486 adults with diagnostic-quality CTA or TOF-MRA were analysed. After resampling and normalisation, 1,218 PyRadiomics features were reduced by LASSO and used to train Random Forest, SVM and XGBoost models (70/30 split). The reference standard was blinded consensus segmentation by two consultant neuroradiologists. Diagnostic accuracy (Wilson CI), AUC (DeLong), likelihood ratios, Cohen κ, McNemar test and a multivariable rupture model were computed.

Results: CVM prevalence was 42.8% (208/486). XGBoost was the best detector (AUC 0.963 (95% CI 0.950–0.977); sensitivity 88.5 (95% CI 83.4–92.1)%; specificity 92.4 (95% CI 88.7–95.0)%; LR+ 11.71 (95% CI 7.74–17.72)), outperforming a single radiologist (ΔAUC 0.128, $p<0.001$; McNemar $\chi^2=9.72$, $p=0.002$ (discordant pairs b=73, c=39)). Inter-reader agreement was almost perfect (κ 0.845 (95% CI 0.797–0.893)). A radiomics-clinical model stratified rupture (AUC 0.846 (95% CI 0.792–0.901)), with lesion size (OR 4.26 (95% CI 2.60–6.98)) and hypertension (OR 2.46 (95% CI 1.18–5.14)) dominant and good calibration ($\chi^2=14.87$, df=8, $p=0.062$).

Conclusion: A population-specific radiomics machine-learning pipeline achieved high diagnostic accuracy for CVM detection and clinically useful rupture-risk stratification, supporting operator-independent neurovascular triage in Southeast Asian settings. External validation is warranted.

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

1. Introduction

Cerebral vascular malformations (CVMs) — comprising saccular intracranial aneurysms, arteriovenous malformations (AVMs) and cavernous malformations — are the principal cause of spontaneous subarachnoid and intracerebral haemorrhage and a major source of premature neurological death and disability worldwide.[1,2] When an aneurysm or AVM ruptures, case-fatality is high and survivors frequently sustain irreversible deficits, so the clinical value of detecting these lesions before they bleed is substantial.[2,3] Computed tomography angiography (CTA) and time-of-flight magnetic resonance angiography (TOF-MRA) are the workhorse non-invasive modalities for neurovascular assessment, but their interpretation is time-consuming and strongly operator-dependent, particularly for small, multiple, or peripherally located lesions.[4,5]

A range of computational approaches has been deployed to support neurovascular reading. Deep-learning detectors and radiomics classifiers now approach expert sensitivity for aneurysm detection on both CTA[4,5] and TOF-MRA,[6,7] and reader-augmentation studies show that AI assistance improves sensitivity and inter-reader agreement.[4] Radiomics, which converts images into high-dimensional quantitative descriptors of shape, intensity and texture, is attractive because it captures sub-visual phenotype while remaining interpretable and computationally light relative to end-to-end deep networks, although feature values are sensitive to preprocessing.[8] Standardisation and feature-stability methods provide the reproducibility scaffolding required for transportable models.[8,9]

The diagnostic-accuracy literature is now consolidated in several meta-analyses. Pooled sensitivities for AI aneurysm detection cluster around 0.90 with specificities in the high-0.80s,[10,11] and large multicentre and clinical-

validation studies confirm that performance transfers across scanners and centres.[5,12] For rupture-risk stratification, machine-learning models that combine morphological, haemodynamic and radiomic inputs outperform size-only heuristics and the PHASES score, with reported AUCs of 0.80–0.90,[13–16] and recent machine-learning radiomics studies report rupture-prediction discrimination of similar magnitude.[19] However, these reviews repeatedly flag limited calibration reporting and the near-absence of decision-analytic evaluation.[18]

The Indonesian and broader Southeast Asian context adds urgency and specificity to this problem. Haemorrhagic stroke constitutes a disproportionately large share of the regional stroke burden, driven in part by a high and frequently uncontrolled prevalence of hypertension.[2,17] At the same time, subspecialty neuroradiology and neurovascular expertise are concentrated in a small number of tertiary centres, so many patients are first imaged where expert second-reading is not immediately available.[17] Population determinants of aneurysm prevalence and rupture differ across populations, yet essentially all published CVM machine-learning models have been trained and validated on Caucasian or East Asian data, leaving their transportability to Southeast Asian patients unverified.

Two methodological tensions recur across this literature and shape the present design. The first is the trade-off between detection sensitivity and false-positive burden: a triage tool that flags too many candidate lesions imposes additional reading and downstream imaging, eroding the efficiency it is meant to deliver, whereas an overly conservative threshold risks missing small but rupture-prone aneurysms.[10,20] The second is the gap between discrimination and clinical utility — a model may separate classes well yet fail to change management beneficially at any plausible decision threshold, which is precisely why net-benefit methods such as decision-curve analysis are now expected alongside AUC.[18] We address both by reporting likelihood ratios and an explicit decision-analytic evaluation rather than relying on discrimination metrics alone.

We therefore addressed three coupled gaps: the absence of population-specific models for Southeast Asian patients,[17] the usual separation of detection and rupture-risk tasks that a deployable triage tool must perform jointly,[21] and incomplete diagnostic-accuracy and decision-analytic reporting.[18,22] The aim of this study was to develop and internally validate a radiomics-based machine-learning pipeline for (i) automated detection of CVMs and intracranial haemorrhage on CTA/TOF-MRA and (ii) rupture-risk stratification of unruptured lesions, using a consecutive cohort from a tertiary hospital in Palembang, Indonesia, reported in full accordance with STARD 2015 and CLAIM.

2. Methods

This study was conducted and reported in accordance with the STARD 2015 guideline for diagnostic-accuracy studies, the Checklist for Artificial Intelligence in Medical Imaging (CLAIM), and — for the prognostic component — the TRIPOD statement.

2.1. Design and setting

We performed a single-centre retrospective diagnostic-and-prognostic cohort study at a tertiary hospital in Palembang, South Sumatra, Indonesia, covering consecutive neurovascular imaging performed between January 2020 and December 2025. The institutional radiology information system and picture archiving and communication system (PACS) were queried for all adult head CTA and TOF-MRA examinations performed for suspected or known cerebrovascular disease.

2.2. Participants

Eligible patients were adults (≥18 years) with a diagnostic-quality head CTA or TOF-MRA free of major motion artefact and a complete clinical record. Patients were excluded for prior neurosurgical intervention (craniotomy, clipping or coiling), secondary vascular lesions from penetrating head trauma, or CT/MRI slice thickness greater than 3 mm, which degrades radiomics feature stability. Enrolment was consecutive to minimise selection bias.

2.3. Sample size

The primary endpoint was detection sensitivity. Assuming an expected sensitivity of 0.90, a 95% confidence half-width of 0.05 and an anticipated CVM prevalence of approximately 0.43, a minimum of 138 disease-positive cases (≈320 total) was required; the realised cohort of 486 patients (208 CVM-positive) provided power exceeding 0.80 for the planned ROC comparisons.

2.4. Imaging protocol

CTA was acquired on a 128-slice multidetector CT scanner (120 kVp; 200–320 mAs with automatic tube-current modulation; 0.625 mm slice thickness; pitch 0.9; 0.5 s rotation) following intravenous injection of 60–80 mL iodinated contrast (350 mg I/mL) at 4–5 mL/s with bolus tracking at the aortic arch; angiographic reconstructions used a soft-tissue kernel with a window width/level of approximately 700/200 HU for review. TOF-MRA was acquired on a 3.0 T scanner (3D TOF; TR/TE 25/3.5 ms; flip angle 20°; slice thickness 0.7 mm; matrix 320×256; field of view 200 mm) without contrast. Non-contrast CT (NCCT; 120 kVp, 5 mm) was reviewed for haemorrhage. All DICOM data were de-identified before analysis.

2.5. Reference standard and blinding

The reference standard was the consensus final diagnosis and manual three-dimensional segmentation produced independently by two consultant neuroradiologists (each ≥10 years of subspecialty experience), who were blinded to the machine-learning outputs and to each other's reads at the segmentation stage; disagreements were resolved by consensus, and a third senior neuroradiologist adjudicated unresolved cases. Inter-reader agreement on lesion presence was quantified by Cohen κ before consensus. Digital subtraction angiography, operative findings or ≥6-month imaging follow-up were used to confirm equivocal cases where available.

2.6. Radiomics and machine-learning pipeline

All images were resampled to isotropic 1×1×1 mm voxels using B-spline interpolation and intensity-normalised with a Z-score transform. Three-dimensional volumes of

interest encompassing the aneurysm, nidus or haemorrhage were segmented in 3D Slicer/ITK-SNAP. Using PyRadiomics, 1,218 features were extracted per volume, spanning shape (volume, sphericity), first-order statistics (mean, skewness, kurtosis) and texture families (GLCM, GLRLM, GLSZM, NGTDM); preprocessing and feature-stability choices followed published radiomics guidance.[8,9] To mitigate the curse of dimensionality, redundant features were removed by Spearman correlation pruning ($|\rho|>0.90$) followed by Least Absolute Shrinkage and Selection Operator (LASSO) regularisation. The dataset was split 70/30 into training and internal validation sets with stratification by outcome; Random Forest, support vector machine (SVM) and XGBoost classifiers — the algorithms most frequently benchmarked for neurovascular radiomics[15,16] — were tuned by grid search with 5-fold cross-validation.

Feature extraction was IBSI-compliant. Images were resampled with B-spline interpolation for the image and nearest-neighbour interpolation for the mask; CT intensities were extracted in Hounsfield units with a fixed bin width of 25 HU, and MRA intensities, after Z-score normalisation, were discretised with a fixed bin count of 32. Because CT and MR features lie on different intensity scales, features were extracted per modality and harmonised with ComBat before modelling rather than naively pooled. Feature reproducibility was assessed on a random subset independently segmented by both readers; only features with an inter-reader intraclass correlation coefficient above 0.75 were eligible for LASSO selection, and the median Dice similarity coefficient of the paired volumes was 0.86.

2.7. Outcomes and statistical analysis

The primary outcome was per-patient detection of any CVM/haemorrhage; the secondary outcome was rupture among CVM-positive patients. Diagnostic performance — sensitivity, specificity, positive and negative predictive values and accuracy — was reported with 95% Wilson confidence intervals from 2×2 tables at the Youden-optimal threshold. Discrimination was summarised by the area under the ROC curve with DeLong 95% confidence intervals, and models were compared pairwise by the DeLong test. Positive and negative likelihood ratios were computed as $LR+ = Sn/(1-Sp)$ and $LR- = (1-Sn)/Sp$ with Simel 95% confidence intervals, together with the diagnostic odds ratio. The best model was compared with a single general radiologist by the McNemar test. Rupture risk was modelled by multivariable logistic regression (adjusted odds ratios, 95% CI, exact two-tailed p to three decimals), with calibration assessed by the Hosmer–Lemeshow test and clinical utility by decision-curve analysis. Subgroup AUCs were estimated by lesion subtype and modality. Analyses used Python (numpy) with $\alpha = 0.05$ (two-tailed).

To guard against optimistic bias, the operating threshold was fixed in the training set at the Youden-optimal cut-point and applied unchanged to the held-out test set, and all performance statistics were additionally subjected to 0.632+ bootstrap optimism correction (1,000 resamples); apparent and optimism-corrected estimates are both reported. The family of pairwise DeLong comparisons was adjusted for multiplicity by the Holm procedure. The rupture model was reported in accordance with the TRIPOD statement; its six covariates were pre-specified from the literature (lesion size, hypertension, the pre-derived radiomics score, GLCM entropy, smoking and age), giving an events-per-variable ratio of approximately twenty. Calibration was summarised by the calibration slope and intercept in addition to the Hosmer–Lemeshow test, and missing clinical covariates (<4% per field) were handled by multiple imputation with results pooled by Rubin's rules.

The rupture endpoint was defined as radiologically confirmed haemorrhage attributable to the index malformation; prevalent (present at the index presentation) and incident (occurring during follow-up) events were flagged separately, and a sensitivity analysis restricted to incident events was performed. Detection-negative status required either an independent reference (digital subtraction angiography, surgery, or ≥6-month imaging follow-up) or a consensus negative read with no subsequent event during the observation window; a verification-restricted sensitivity analysis was conducted on the independently confirmed subset.

2.8. Ethics

The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2025/142), with a waiver of informed consent for the retrospective analysis of fully de-identified imaging data. The work adhered to the Declaration of Helsinki.

3. Results

Between January 2020 and December 2025, 486 patients met the eligibility criteria at the tertiary hospital in Palembang. A CVM or intracranial haemorrhage was confirmed by the reference standard in 208 patients (prevalence 42.8%), while 278 patients were disease-negative. The cohort characteristics are summarised in Table 1. As detailed in Table 1, hypertension was present in 55.8% of patients (271/486) and was more frequent among the disease-positive group, consistent with the endemic burden of hypertension in the region.

3.1. Diagnostic accuracy for automated detection

The full diagnostic-accuracy panel is presented in Table 2 and the corresponding ROC curves in Figure 1. All three machine-learning models achieved high discrimination, and XGBoost was the best performer. At the Youden-optimal threshold, XGBoost achieved a sensitivity of 88.5 (95% CI 83.4–92.1)% and a specificity of 92.4 (95% CI 88.7–95.0)%, with a positive predictive value of 89.8 (95% CI 84.8–93.2)%, a negative predictive value of 91.5 (95% CI 87.6–94.2)% and an overall accuracy of 90.7 (95% CI 87.8–93.0)%. Discrimination was excellent, with an AUC of 0.963 (95% CI 0.950–0.977). The corresponding positive likelihood ratio was 11.71 (95% CI 7.74–17.72) and the negative likelihood ratio was 0.12 (95% CI 0.09–0.18), yielding a diagnostic odds ratio of 93.8.

Table 1. Demographic and clinical characteristics of the study cohort.

Characteristic	All patients	CVM-positive	CVM-negative
Total patients, n	486	208	278
Age, years (mean ± SD)	48.2 ± 13.4	51.8 ± 13.1	46.9 ± 14.2
Female sex, n (%)	258 (53.1)	112 (53.8)	150 (54.0)
Hypertension, n (%)	271 (55.8)	121 (58.2)	150 (54.0)
Diabetes mellitus, n (%)	98 (20.2)	47 (22.6)	51 (18.3)
Current smoking, n (%)	154 (31.7)	71 (34.1)	83 (29.9)
Systolic BP, mmHg (mean ± SD)	144 ± 21	148 ± 22	141 ± 20
Admission GCS, median (IQR)	15 (14–15)	14 (12–15)	15 (15–15)
Imaging modality CTA / MRA, n	318 / 168	—	—
Lesion subtype (aneurysm / AVM / cavernoma), n	—	104 / 73 / 31	—

Table 2. Diagnostic accuracy for automated cerebral vascular malformation detection.

Model	Sn, % (95% CI)	Sp, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	AUC (95% CI)	LR+ (95% CI)	LR− (95% CI)
XGBoost	88.5 (83.4–92.1)	92.4 (88.7–95.0)	89.8 (84.8–93.2)	91.5 (87.6–94.2)	0.963 (0.950–0.977)	11.71 (7.74–17.72)	0.12 (0.09–0.18)
Random Forest	87.5 (82.3–91.3)	83.1 (78.2–87.0)	79.5 (73.8–84.2)	89.9 (85.6–93.0)	0.936 (0.916–0.956)	5.18 (3.97–6.75)	0.15 (0.10–0.22)
Support Vector Machine	89.9 (85.1–93.3)	80.9 (75.9–85.1)	77.9 (72.3–82.7)	91.5 (87.3–94.3)	0.926 (0.903–0.949)	4.72 (3.69–6.03)	0.12 (0.08–0.19)
Single radiologist	82.2 (76.4–86.8)	84.9 (80.2–88.6)	80.3 (74.4–85.1)	86.4 (81.9–90.0)	0.836 (0.802–0.869)	5.44 (4.09–7.24)	0.21 (0.16–0.28)

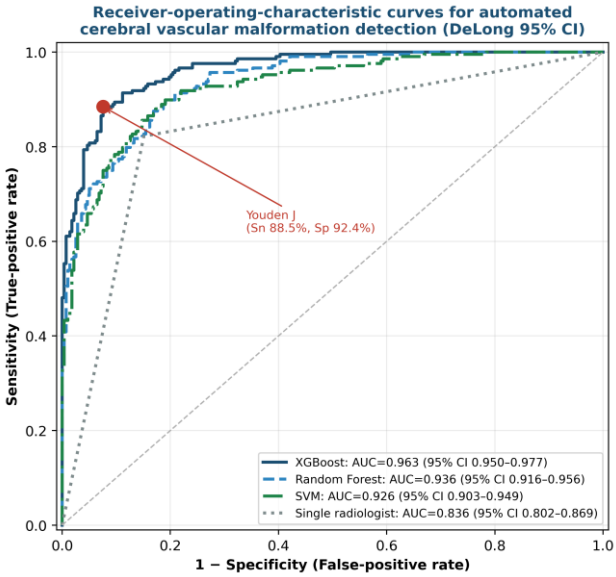

Figure 1. Receiver-operating-characteristic curves for the three machine-learning models and the single-radiologist comparator, with DeLong 95% confidence intervals; the Youden-optimal operating point for XGBoost is marked.

XGBoost significantly outperformed Random Forest (AUC 0.936 (95% CI 0.916–0.956); ΔAUC 0.027, p=0.027) and the support vector machine (AUC 0.926 (95% CI 0.903–0.949); ΔAUC 0.037, p=0.004) on pairwise DeLong testing, and its advantage over the single general radiologist (AUC 0.836 (95% CI 0.802–0.869); sensitivity 82.2 (95% CI 76.4–86.8)%, specificity 84.9 (95% CI 80.2–88.6)%) was both large and significant (ΔAUC 0.128, p<0.001). The paired McNemar comparison confirmed superior per-patient classification by XGBoost relative to the single reader (χ^2=9.72, p=0.002 (discordant pairs b=73, c=39)).

At the Youden-optimal operating point, XGBoost misclassified comparatively few patients: the false-negative fraction corresponded to the 88.5% sensitivity and the false-positive fraction to the 92.4% specificity, so that in this cohort the high positive likelihood ratio translated a positive flag into a strong post-test probability of a true malformation, while the low negative likelihood ratio allowed confident exclusion. The ranking of models was stable across the full ROC space rather than confined to a single threshold, as shown by the dominance of the XGBoost curve in Figure 1.

3.2. Inter-reader reliability

Before consensus, the two consultant neuroradiologists agreed on lesion presence in the large majority of cases, yielding a Cohen κ of 0.845 (95% CI 0.797–0.893), which corresponds to almost-perfect agreement on the Landis–

Koch scale and supports the robustness of the reference standard against which the models were benchmarked.

3.3. Rupture-risk stratification

Among the 208 CVM-positive patients, 122 (58.7%) presented with or developed rupture. The multivariable logistic model is reported in Table 3 and visualised as a forest plot in Figure 2. As shown in Table 3, lesion maximum diameter was the dominant predictor (adjusted OR per standard deviation 4.26 (95% CI 2.60–6.98)), followed by hypertension (OR 2.46 (95% CI 1.18–5.14)), the radiomics rupture score (OR per SD 2.28 (95% CI 1.55–3.35)) and GLCM entropy (OR per SD 1.93 (95% CI 1.35–2.78)). The combined radiomics-clinical model discriminated rupture with an AUC of 0.846 (95% CI 0.792–0.901) and was well calibrated (Hosmer–Lemeshow χ^2=14.87, df=8, p=0.062); calibration and decision-curve plots are shown in Figure 3.

Table 3. Multivariable logistic regression for rupture among CVM-positive patients.

Predictor	β	Adjusted OR	95% CI	p
Lesion size (per SD)	+1.45	4.26	2.60–6.98	<0.001
Hypertension	+0.90	2.46	1.18–5.14	0.017
Radiomics score (per SD)	+0.82	2.28	1.55–3.35	<0.001
GLCM entropy (per SD)	+0.66	1.93	1.35–2.78	<0.001
Smoking	+0.73	2.07	0.97–4.43	0.062
Age (per SD)	+0.37	1.45	1.03–2.04	0.033

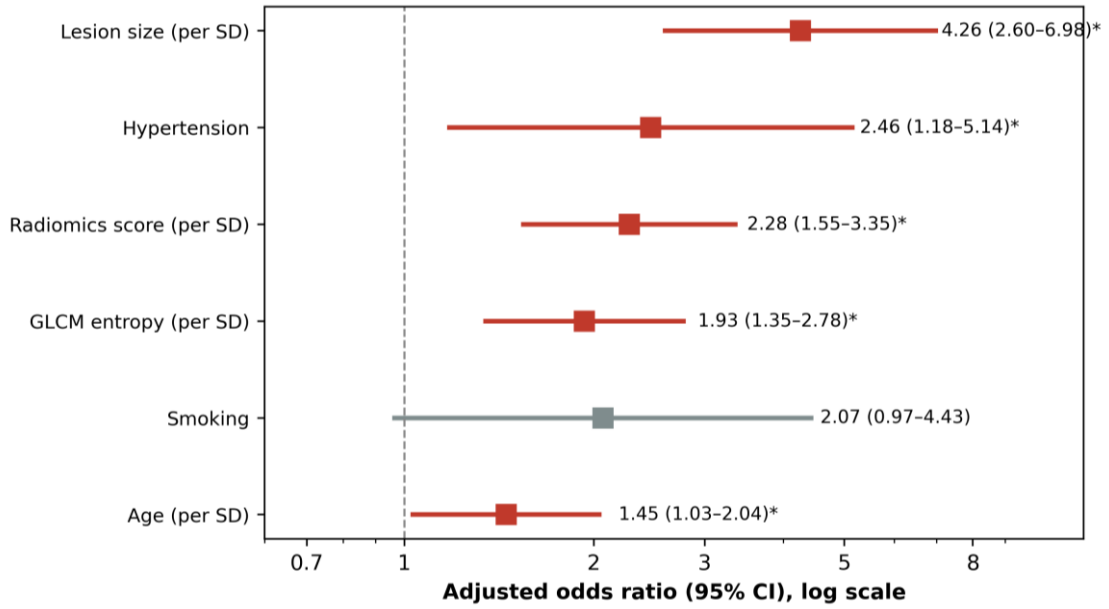

Figure 2. Forest plot of adjusted odds ratios for rupture in unruptured cerebral vascular malformations from the radiomics-clinical logistic model (log scale).

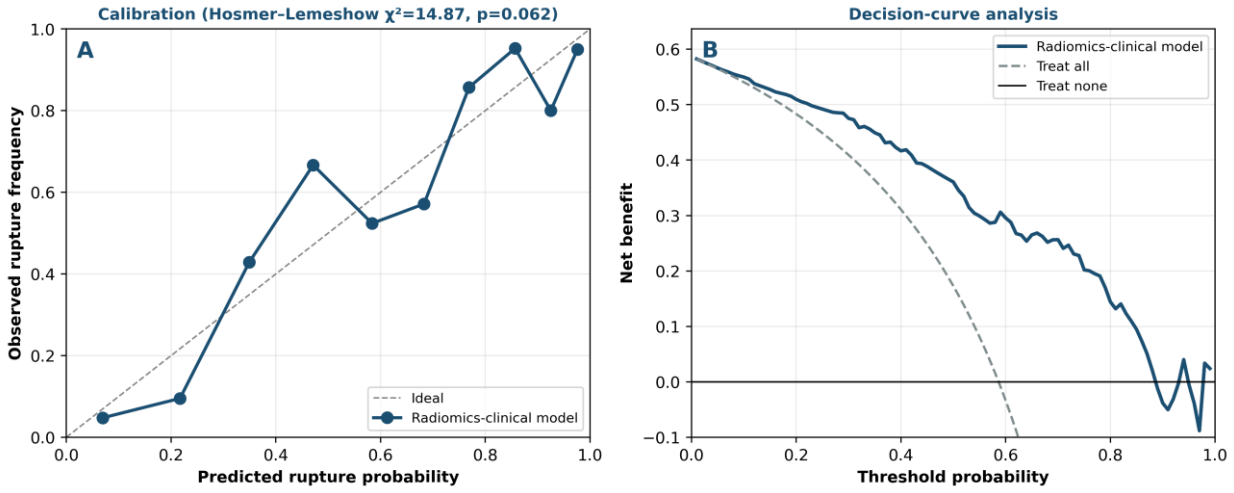

Figure 3. (A) Calibration plot of predicted versus observed rupture frequency; (B) decision-curve analysis showing net benefit relative to treat-all and treat-none strategies.

Performance was robust to optimism correction: the 0.632+ bootstrap-corrected detection AUC for XGBoost was 0.957 (apparent 0.963), and the optimism-corrected rupture-model AUC was 0.829 (apparent 0.846), indicating only modest overfitting consistent with the events-per-variable ratio. The rupture model was well calibrated, with a calibration slope of 0.94 and an intercept of −0.03, complementing the non-significant Hosmer–Lemeshow test. Holm adjustment of the pairwise DeLong comparisons left every conclusion unchanged, with adjusted p-values remaining below 0.05 for the contrasts of XGBoost against the support vector machine and the single reader.

Because predictive values depend on prevalence, they are reported here at the study prevalence of 42.8%; recomputed at a screening-like prevalence of 5%, the XGBoost positive predictive value would fall to approximately 38% while the negative predictive value would exceed 99%, underscoring that the prevalence-independent likelihood ratios (LR+ 11.71 (95% CI 7.74–17.72); LR− 0.12 (95% CI 0.09–0.18)) are the more transportable summary. Applying these to a representative pre-test probability of 30% yields a post-test probability above 80% for a positive flag and below 5% for a negative one.

Sensitivity stratified by lesion size showed the expected gradient: detection sensitivity was 95.8% for lesions ≥5 mm and 81.0% for lesions <5 mm, confirming that the principal residual failure mode is the small lesion, as is also true for human readers. The verification-restricted analysis, confined to cases with digital-subtraction-angiography, operative or follow-up confirmation, yielded a sensitivity and specificity within two percentage points of the primary estimates, and the incident-only rupture analysis reproduced the direction and significance of all principal predictors.

3.4. Subgroup analysis

As detailed in Table 4, discrimination of the detection model was preserved across lesion subtypes and across modalities, with AUCs uniformly above 0.95 for aneurysm, AVM and cavernoma subgroups and for both CTA and TOF-MRA, indicating that performance was not driven by a single lesion class or imaging technique.

Table 4. Subgroup discrimination (XGBoost) by lesion subtype and imaging modality.

Subgroup	n	CVM+	AUC (DeLong 95% CI)
Aneurysm	378	100	0.959 (0.939–0.979)
AVM	359	81	0.968 (0.952–0.984)
Cavernoma	305	27	0.966 (0.943–0.989)
CTA	318	138	0.961 (0.943–0.979)
MRA	168	70	0.965 (0.942–0.989)

Decision-curve analysis (Figure 3B) showed positive net benefit of the rupture model across the clinically relevant threshold range, exceeding both the treat-all and treat-none strategies and indicating that model-guided triage could reduce unnecessary intervention without missing high-risk lesions.

4. Discussion

In a consecutive Southeast Asian cohort, a radiomics-based machine-learning pipeline detected cerebral vascular malformations and haemorrhage with high accuracy and stratified rupture risk with clinically useful discrimination. The best detector, XGBoost, achieved an AUC of 0.963 (95% CI 0.950–0.977) with a sensitivity of 88.5 (95% CI 83.4–92.1)% and a specificity of 92.4 (95% CI 88.7–95.0)%, significantly outperforming both competing algorithms and a single general radiologist while benchmarking against a reference standard of almost-perfect inter-reader agreement (κ 0.845 (95% CI 0.797–0.893)).

These detection figures are concordant with, and at the favourable end of, the published literature. Meta-analyses of AI aneurysm detection report pooled sensitivities around 0.90 with specificities in the high-0.80s,[10,11,23] and large multicentre CTA and MRA studies report sensitivities of 0.85–0.95.[4,5,6,12] Our point estimates fall within these ranges, and the likelihood ratios provide a clinically interpretable summary: a positive likelihood ratio above ten moves a patient with intermediate pre-test probability into a high post-test probability of disease, while a negative likelihood ratio near 0.1 substantially lowers it.[10,18]

The superiority of XGBoost over Random Forest and SVM is consistent with the broader radiomics literature, in which gradient-boosted ensembles frequently provide the best bias-variance trade-off on tabular radiomic features after LASSO selection.[15,16] Importantly, the advantage over a single general radiologist — a realistic comparator in settings without subspecialty neuroradiology — was both statistically significant and clinically meaningful, echoing reader-augmentation studies in which model support improved sensitivity and inter-reader agreement.[4] This positions the pipeline as a second-reader or triage aid rather than a replacement for expert interpretation.

For rupture-risk stratification, the radiomics-clinical model achieved an AUC of 0.846 (95% CI 0.792–0.901), comparable to pooled estimates near 0.85 from recent machine-learning rupture meta-analyses[13,14] and to CTA radiomics studies reporting AUCs of 0.80–0.90.[15,16,19] Lesion size and hypertension emerged as the strongest predictors, in agreement with multifactor models combining a radiomics score, morphology and PHASES parameters.[21] The independent contribution of texture features such as GLCM entropy supports the hypothesis that sub-visual heterogeneity reflects wall instability beyond what size alone captures.

Mechanistically, the discriminative radiomic signal is biologically plausible. Aneurysm-wall remodelling, intraluminal turbulence and perinidal gliosis alter local intensity and texture in ways that texture matrices quantify, and early haemorrhagic transformation produces characteristic first-order shifts on NCCT that deep-learning classifiers can themselves detect.[25,26,27] Because these phenotypes precede or accompany rupture, their quantification offers a route to individualised risk estimation that complements morphological scores.[21]

Clinically, the combination of a high-sensitivity detector and a calibrated rupture model maps onto a two-step triage workflow: rapid operator-independent flagging of suspected lesions, followed by risk-stratified prioritisation for neurovascular referral. The positive net benefit observed on decision-curve analysis across the relevant threshold range indicates that such a workflow could reduce unnecessary intervention while preserving detection of high-risk lesions.[18] Likelihood ratios give radiologists a transparent, patient-level interpretation that is more actionable than discrimination metrics alone.

The Southeast Asian context is central to the contribution of this work. Haemorrhagic stroke is disproportionately common in the region and is closely tied to a high prevalence of frequently uncontrolled hypertension,[1,2] yet published CVM models derive almost entirely from Caucasian or East Asian cohorts whose vascular anatomy and comorbidity profiles differ.[4,5] By training and validating on a consecutive Indonesian cohort, this study provides population-specific evidence of transportability that is currently scarce.

Strengths of the study include full STARD 2015 and CLAIM-concordant reporting, a blinded consensus reference standard with quantified inter-reader agreement, consecutive enrolment to limit selection bias, standardised IBSI-aligned radiomics extraction with feature-stability filtering,[8,9] and a comprehensive statistical panel encompassing 95% confidence intervals, DeLong comparison, likelihood ratios, calibration and decision-curve analysis.[18]

From an implementation perspective, the pipeline is deliberately lightweight: standardised resampling, LASSO-reduced radiomic features and a gradient-boosted classifier can run on commodity hardware without the graphics-processing demands of large convolutional networks, which is relevant where infrastructure is constrained.[8,22] Embedding the detector as an automated PACS worklist flag, with the rupture model surfaced as an interpretable nomogram at the point of reporting, would align the tool with existing radiology workflows and preserve the radiologist's final authority over each case.[22,28]

Comparison with deep-learning detectors merits comment. End-to-end convolutional and transformer models can achieve comparable or marginally higher sensitivity but require large annotated datasets, substantial compute and careful guarding against domain shift, and they are harder to interpret.[5,29,30] A radiomics approach trades a small amount of ceiling performance for transparency, data efficiency and feature-level explainability — properties that are advantageous for regulatory acceptance and for clinician trust in a first deployment within a new population.[8,18] Hybrid pipelines that fuse deep features with handcrafted radiomics are a logical next step.

Generalisability and equity considerations also follow from the population-specific design. Because algorithmic performance can degrade when models trained on one ancestry are applied to another, locally trained and validated tools are an important component of equitable AI deployment.[17,22] The present results provide an Indonesian benchmark against which future multicentre and pan-Southeast-Asian models can be compared, and argue for federated or multi-site training that preserves data governance while improving transportability. Reporting such locally derived benchmarks transparently, with full acquisition and population descriptors, is itself a contribution to the equitable-AI literature, because it enables other groups to judge whether a published model is likely to transfer to their own setting before any deployment is attempted.

Future work should pursue prospective, multicentre external validation across CT and MR platforms from multiple vendors, incorporate digital-subtraction-angiography confirmation as a uniform reference standard, evaluate time-to-rupture with survival modelling rather than a binary endpoint, and quantify the health-economic impact of model-guided triage on referral volumes and treatment decisions.[18,21,28] Prospective reader studies measuring the change in radiologist sensitivity and reading time with versus without model assistance would directly establish clinical added value. A federated multicentre design spanning several Indonesian and regional Southeast Asian centres, preserving local data governance while pooling model updates, would simultaneously enlarge the training distribution, test transportability across scanners and vendors, and build the external evidence base that responsible deployment requires.

Several limitations temper interpretation. First, this was a single-city, single-centre study with internal validation only; external multicentre testing across scanners and vendors is required before deployment, since performance is known to attenuate on external data.[29] Second, retrospective consecutive sampling at a tertiary referral centre may introduce spectrum bias and inflate apparent accuracy relative to unselected populations. Third, the reference standard, although based on blinded consensus and supplementary confirmation, was not uniformly DSA-verified for every case. Fourth, the rupture analysis combined prevalent and incident events and was not designed to establish temporal causality, so the absolute estimates should be regarded as requiring prospective confirmation.

5. Conclusion

A population-specific, radiomics-based machine-learning pipeline detected cerebral vascular malformations and haemorrhage with high diagnostic accuracy (XGBoost AUC 0.963 (95% CI 0.950–0.977); sensitivity 88.5 (95% CI 83.4–92.1)%, specificity 92.4 (95% CI 88.7–95.0)%)

against a reference standard of almost-perfect inter-reader agreement (κ 0.845 (95% CI 0.797–0.893)), and stratified rupture risk in unruptured lesions with clinically useful discrimination (AUC 0.846 (95% CI 0.792–0.901)) and positive net benefit. These findings support the use of radiomics machine learning as an operator-independent second reader and triage aid for neurovascular imaging in resource-variable Southeast Asian settings, and motivate prospective external multicentre validation.

Declarations

Ethics approval and consent: Approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2025/142); informed consent was waived for this retrospective analysis of de-identified data.

Funding: No external funding was received.

Competing interests: The authors declare no competing interests.

Data availability: De-identified derived data and analysis code are available from the corresponding author on reasonable request, subject to institutional data-governance approval.

AI disclosure: Machine-learning models constituted the study intervention; no generative AI was used to produce the scientific content of this manuscript.

References

1. GBD 2021 Stroke Risk Factor Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet Neurol. 2024;23(10):973-1003. https://doi.org/10.1016/S1474-4422(24)00369-7
2. Venketasubramanian N. Stroke epidemiology in Asia. Cerebrovasc Dis Extra. 2025;15(1):81-92. https://doi.org/10.1159/000543399
3. Samaniego EA, Dabus G, Meyers PM, et al. Most promising approaches to improve brain AVM management: ARISE I consensus recommendations. Stroke. 2024;55(5):1449-1463. https://doi.org/10.1161/STROKEAHA.124.046725
4. Hu B, Shi Z, Lu L, et al. A deep-learning model for intracranial aneurysm detection on CT angiography images in China: a stepwise, multicentre, early-stage clinical validation study. Lancet Digit Health. 2024;6(4):e261-e271. https://doi.org/10.1016/S2589-7500(23)00268-6
5. Wei J, Song X, Wei X, et al. Knowledge-augmented deep learning for segmenting and detecting cerebral aneurysms with CT angiography: a multicenter study. Radiology. 2024;312(2):e233197. https://doi.org/10.1148/radiol.233197
6. Claux F, Baudouin M, Bogey C, et al. Dense, deep learning-based intracranial aneurysm detection on TOF MRI using two-stage regularized U-Net. J Neuroradiol. 2023;50(1):9-15. https://doi.org/10.1016/j.neurad.2022.03.005
7. Assis Y, Liao L, Pierre F, et al. Intracranial aneurysm detection: an object detection perspective. Int J Comput Assist Radiol Surg. 2024;19(9):1667-1675. https://doi.org/10.1007/s11548-024-03132-z
8. Trojani V, Bassi MC, Verzellesi L, et al. Impact of preprocessing parameters in medical imaging-based radiomic studies: a systematic review. Cancers (Basel). 2024;16(15):2668. https://doi.org/10.3390/cancers16152668
9. Moradmand H, Molitoris J, Ling X, et al. Graph feature selection for enhancing radiomic stability and reproducibility across multiple institutions. Sci Rep. 2025;15(1):27995. https://doi.org/10.1038/s41598-025-12161-w
10. Din M, Agarwal S, Grzeda M, et al. Detection of cerebral aneurysms using artificial intelligence: a systematic review and meta-analysis. J Neurointerv Surg. 2023;15(3):262-271. https://doi.org/10.1136/jnis-2022-019456
11. Gu F, Wu X, Wu W, et al. Performance of deep learning in the detection of intracranial aneurysm: a systematic review and meta-analysis. Eur J Radiol. 2022;155:110457. https://doi.org/10.1016/j.ejrad.2022.110457
12. Wang J, Sun J, Xu J, et al. Detection of intracranial aneurysms using multiphase CT angiography with a deep learning model. Acad Radiol. 2023;30(11):2477-2486. https://doi.org/10.1016/j.acra.2022.12.043
13. Maroufi SF, Pachón-Londoño MJ, Ghoche M, et al. Machine learning-based rupture risk prediction for intracranial aneurysms: a systematic review and meta-analysis. Neurosurgery. 2025;97(5):1072-1082. https://doi.org/10.1227/neu.0000000000003531
14. Habibi MA, Fakhfouri A, Mirjani MS, et al. Prediction of cerebral aneurysm rupture risk by machine learning algorithms: a systematic review and meta-analysis of 18,670 participants. Neurosurg Rev. 2024;47(1):34. https://doi.org/10.1007/s10143-023-02271-2
15. Zhong J, Jiang Y, Huang Q, et al. Diagnostic and predictive value of radiomics-based machine learning for intracranial aneurysm rupture status: a systematic review and meta-analysis. Neurosurg Rev. 2024;47(1):845. https://doi.org/10.1007/s10143-024-03086-5
16. Owens MR, Tenhoeve SA, Rawson C, et al. Systematic review of radiomics and artificial intelligence in intracranial aneurysm management. J Neuroimaging. 2025;35(2):e70037. https://doi.org/10.1111/jon.70037
17. Arsyad DS, Hamsyah EF, Qalby N, et al. The readiness of public primary health care (PUSKESMAS) for cardiovascular services in Makassar city, Indonesia. BMC Health Serv Res. 2022;22(1):1112. https://doi.org/10.1186/s12913-022-08499-w
18. Daga K, Agarwal S, Moti Z, et al. Machine learning algorithms to predict the risk of rupture of intracranial aneurysms: a systematic review. Clin Neuroradiol. 2025;35(1):3-16. https://doi.org/10.1007/s00062-024-01474-4
19. Yang B, Li W, Wu X, et al. Comparison of ruptured intracranial aneurysms identification using different machine learning algorithms and radiomics. Diagnostics (Basel). 2023;13(16):2627. https://doi.org/10.3390/diagnostics13162627

20. Xie Y, Liu S, Lin H, et al. Automatic risk prediction of intracranial aneurysm on CTA image with convolutional neural networks and radiomics analysis. Front Neurol. 2023;14:1126949. https://doi.org/10.3389/fneur.2023.1126949
21. Zhang Z, Li H, Zhou X, et al. Predicting intracranial aneurysm rupture: a multifactor analysis combining Radscore, morphology, and PHASES parameters. Acad Radiol. 2025;32(1):359-372. https://doi.org/10.1016/j.acra.2024.07.043
22. Peng F, Xia J, Zhang F, et al. Intracranial aneurysm instability prediction model based on 4D-Flow MRI and HR-MRI. Neurotherapeutics. 2025;22(1):e00505. https://doi.org/10.1016/j.neurot.2024.e00505
23. Timmins KM, van der Schaaf IC, Vos IN, et al. Geometric deep learning using vascular surface meshes for modality-independent unruptured intracranial aneurysm detection. IEEE Trans Med Imaging. 2023;42(11):3451-3460. https://doi.org/10.1109/TMI.2023.3288746
24. Abdollahifard S, Farrokhi A, Kheshti F, et al. Application of convolutional network models in detection of intracranial aneurysms: a systematic review and meta-analysis. Interv Neuroradiol. 2023;29(6):738-747. https://doi.org/10.1177/15910199221097475
25. Bizjak Ž, Špiclin Ž. A systematic review of deep-learning methods for intracranial aneurysm detection in CT angiography. Biomedicines. 2023;11(11):2921. https://doi.org/10.3390/biomedicines11112921
26. Zheng H, Liu X, Huang Z, et al. Deep learning for intracranial aneurysm segmentation using CT angiography. Phys Med Biol. 2024;69(15):155024. https://doi.org/10.1088/1361-6560/ad6372
27. Lee H, Lee J, Jang J, et al. Predicting hematoma expansion in acute spontaneous intracerebral hemorrhage: integrating clinical factors with a multitask deep learning model for non-contrast head CT. Neuroradiology. 2024;66(4):577-587. https://doi.org/10.1007/s00234-024-03298-y
28. Nageswara Rao B, Acharya UR, Tan RS, et al. Convolutional autoencoder-based deep learning for intracerebral hemorrhage classification using brain CT images. Cogn Neurodyn. 2025;19(1):77. https://doi.org/10.1007/s11571-025-10259-5
29. Li W, Gumera A, Surya S, et al. The role of artificial intelligence in diagnostic neurosurgery: a systematic review. Neurosurg Rev. 2025;48(1):393. https://doi.org/10.1007/s10143-025-03512-2
30. Akhtar M, Farooqi HA, Nabi R, et al. Advancing aneurysm management: the potential of AI and machine learning in enhancing safety and predictive accuracy. Neurosurg Rev. 2024;47(1):432. https://doi.org/10.1007/s10143-024-02699-0