Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Dataset Generalization and Explainability in AI-Based Intracranial Hemorrhage Detection from Noncontrast CT Using EfficientNet-B4

Lochan Shrestha

lochanshrestha@pahs.edu.np

Patan Academy of Health Sciences

Huma Subhani

Patan Academy of Health Sciences

Article

Keywords: intracranial hemorrhage, deep learning, EfficientNet, multilabel classification, Grad-CAM, external validation, CT imaging, domain shift, class imbalance

Posted Date: May 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9705343/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Intracranial hemorrhage (ICH) is a neurological emergency with a mortality rate exceeding 40%. Deep learning models trained on benchmark datasets rarely demonstrate robust generalizability to independent external cohorts. We developed and externally validated an open-source EfficientNet-B4 pipeline for multilabel ICH detection with integrated Grad-CAM explainability. The model was trained on 752,803 CT slices from the RSNA 2019 dataset via inverse frequency weighted loss across six outputs, with three-window RGB encoding as input. External validation was performed on the CQ500 (473 studies, New Delhi, India) without retraining. On the RSNA internal test set (112,921 slices), the model achieved a macro AUC-ROC of 0.9835 (95% CI 0.9821–0.9847), a sensitivity of 0.8192, a specificity of 0.9862, and an F1 of 0.7743. For CQ500, all sigmoid predictions fell below 0.5, yielding zero sensitivity — a primary negative finding attributable to domain shift. The threshold-independent macro AUC-ROC was 0.7276 (95% CI 0.6729–0.7815). Intraventricular hemorrhage was the most consistent cross-dataset generalization (AUC-ROC 0.8079, 95% CI 0.691–0.904). Grad-CAM heatmaps (n = 60, two independent radiologists) achieved a combined mean localization score of 3.81 ± 0.86/5 with substantial interrater agreement (κ = 0.653). Domain shift causes complete threshold failure on external data; threshold recalibration on a representative local validation set is a nonnegotiable prerequisite for deployment. The open-source pipeline is released to facilitate adaptation in resource-limited settings.

1. Introduction

Intracranial hemorrhage (ICH) is a neurological emergency associated with mortality exceeding 40% and substantial long-term disability among survivors [1,2]. Rapid and accurate diagnosis is critical: delayed identification is directly associated with worse neurological outcomes [1]. Noncontrast computed tomography (NCCT) is the primary imaging modality for the initial assessment of suspected ICH, owing to its widespread availability, speed, and high sensitivity for acute blood products. Accurate interpretation demands the recognition of five radiologically and clinically distinct subtypes (epidural, intraparenchymal, intraventricular, subarachnoid, and subdural hemorrhage), each with different etiologies, anatomical distributions, and management implications. This interpretive burden is compounded in high-volume emergency settings and in low- and middle-income countries (LMICs), where specialist neuroradiology coverage may be limited.

Deep learning has demonstrated strong diagnostic performance for ICH detection from NCCT. A 2025 systematic review and meta-analysis of 58 studies reported a pooled sensitivity of 0.92, specificity of 0.94, and AUC-ROC of 0.96 [3]. Landmark studies have achieved AUC-ROC curves as high as 0.991 [4], and reader studies have confirmed that DL assistance improves diagnostic accuracy [1]. Multiple dataset ensemble approaches have reported an AUC-ROC of 0.953 [5], which serves as the primary benchmark for the present work.

Despite this progress, five important limitations persist. First, most models are validated on a single dataset, and few open-source systems demonstrate robust generalization across heterogeneous CT

acquisitions from geographically distinct populations [6,7]. Second, epidural hemorrhage, the rarest ICH subtype (< 0.5% of slices), receives minimal attention. Third, the severe class imbalance inherent to ICH datasets is inconsistently addressed [8]. Fourth, structured radiologist scoring of model attention maps has not been reported. Fifth, open reproducible pipelines enabling replication in LMIC contexts remain scarce.

To address these gaps, we present an open EfficientNet-B4 pipeline trained on 752,803 CT slices from RSNA 2019 dataset [9] and externally validated on CQ500 [10]. Our contributions are as follows: (i) cross-dataset generalization with transparent reporting of domain shifts; (ii) Grad-CAM explainability with radiologist localization scoring; (iii) class-balanced imbalance handling with verified numerical accuracy; and (iv) full public release of code and model weights.

2. Related Work

2.1 Deep Learning Architectures for ICH Detection

The application of deep learning to ICH detection from NCCT has advanced rapidly since Chilamkurthy et al. demonstrated CNN-based detection across 313,000 head CT scans with an AUC of 0.91–0.97 on the CQ500 validation set [11]. Kuo et al. reported an AUC-ROC of 0.991 on 4,396 CT scans, demonstrating expert-level performance [4]. The RSNA 2019 Challenge [9] standardized evaluation on the largest publicly available annotated CT dataset. Burduja et al. finished in the top 2% of 1,345 teams, integrating Grad-CAM visualizations [12]. Postchallenge, D'Angelo et al. evaluated Dense-UNet on 502 NCCT scans with an accuracy of 91.24% [13]. Ensemble methods have consistently outperformed single-model approaches [14]. EfficientNet [15], designed with compound scaling, has demonstrated strong medical imaging performance [16].

2.2 Multilabel Subtype Classification and Class Imbalance

Multilabel binary cross-entropy formulations better reflect the clinical reality that a single CT study may reveal multiple concurrent hemorrhage types [14]. Class imbalance poses a compounding difficulty, particularly for epidural hemorrhage. Strategies include focal loss [8], oversampling, and weighted loss functions. Kok et al. demonstrated the effectiveness of focal loss for intraventricular hemorrhage segmentation [17]. In the present work, we adopt a class-balanced inverse frequency weighting formula and verify it numerically against logged training values.

2.3 Cross-Dataset Generalization

Generalization to external cohorts is performed in only a minority of published ICH studies [3]. Salehinejad et al. reported an AUC-ROC of 95.4% for external validation, with a drop from 98.4% internally [7]. Voter et al. identified failure modes of a validated algorithm during prospective deployment [5]. Nada et al. performed external validation on 5,600 heterogeneous CTs (AUC-ROC 0.954) [6]. The CQ500 dataset [11],

collected in New Delhi, India, has become the most widely used external benchmark for models trained on North American data.

2.4 Explainability in Brain CT AI

Grad-CAM [18] has become the most widely applied post hoc explanation method for CNN-based medical image classifiers. Lee et al. demonstrated that saliency-based explanations improved clinician trust in ICH detection [10]. Kim et al. applied Grad-CAM to all six ICH subtypes via ResNet, providing a proof-of-concept [19]. Altuve and Pérez conducted qualitative heatmap validation for ICH detection [20]. The present study extends this work by conducting formal radiologist scoring across all six output classes.

3. Materials and methods

Reporting standards. This study follows the TRIPOD-AI and CLAIM (Checklist for Artificial Intelligence in Medical Imaging) reporting guidelines for diagnostic AI studies. A supplementary checklist is available from the authors upon request.

3.1 Datasets

RSNA 2019 Intracranial Hemorrhage Detection Dataset. The primary training dataset was sourced from the RSNA 2019 Brain CT Hemorrhage Challenge [9]. After slice-level labels were extracted, the dataset comprised 752,803 CT slices across six binary labels: any_ich, epidural, intraparenchymal, intraventricular, subarachnoid, and subdural hemorrhage. The RSNA 2019 dataset provides study-level annotations; slice-level labels were generated by propagating the study-level label to all slices within a study. This introduces label noise, as hemorrhagic subtypes may be visible in only a subset of slices within a positive study (discussed in Section 5.5). The raw DICOM files were converted to JPEG format (quality factor 95) prior to training.

CQ500 Dataset. CQ500 [11] comprises 491 noncontrast head CT studies collected in New Delhi, India. Of these, 18 studies (3.7%) were excluded because of the absence of DICOM folders in the Kaggle-hosted version of the dataset; no studies were excluded on a clinical basis. The remaining 473 studies were used for external validation. ICH was present in 205 studies (43.3%) and absent in 268 (56.7%). Each study was annotated by three radiologists; the majority vote (≥ 2/3) was the ground truth. CQ500 was used exclusively for external validation.

All the RSNA 2019 data were split at the study level: 70% training/15% validation/15% testing, yielding 526,962/112,920/112,921 slices, respectively. An assertion verified that no StudyInstanceUID appeared in more than one partition. StudyInstanceUID was used as the primary split key, constituting the closest available proxy for patient-level separation in this deidentified dataset.

3.3 Class distribution and imbalance handling

The RSNA training set exhibits severe class imbalance (Table 1). Binary cross-entropy loss with per-class positive weighting (BCEWithLogitsLoss, pos_weight) was applied during training. The positive class weight was computed as:

$$\mathbf{w}\, c = N/(Nc \times C)$$

where N is the total number of training slices, Nc is the number of positive slices for class c, and $C = 6$ is the number of output classes. This ensures that each class contributes equally to the aggregate loss at initialization. Weights ranged from 1.16 (any_ich) to 39.51 (epidural). The formula was confirmed against the logged training output for all six classes. Implementation: `pos_weight = total/(pos_counts * num_classes + 1e-6)`. The ε = 1e-6 term is a numerical stabilizer used to prevent division by zero for any class with zero positive samples in the training set (not applicable here but included for robustness). This approach was preferred over focal loss [8,] as it requires no additional hyperparameter tuning.

Table 1
Class distribution in the RSNA 2019 training set (n = 526,962 slices).

Hemorrhage subtype	Positive slices	Prevalence (%)	pos_weight (wc)	n_test pos.
any_ich	75,553	14.34	1.16	16,084
epidural	2,223	0.42	39.51	459
intraparenchymal	25,258	4.80	3.48	5,279
intraventricular	18,383	3.49	4.78	3,900
subarachnoid	24,952	4.74	3.52	5,431
subdural	33,108	6.29	2.65	7,043

n_test pos. : positive slices in the held-out RSNA test set. C = 6 classes.

3.4 CT preprocessing and windowing

Noncontrast CT images were preprocessed via a three-window RGB encoding strategy [10,21,22]. Three clinically standard Hounsfield unit (HU) windows were applied and stacked as separate color channels: the brain (center 40 HU, width 80 HU), subdural (center 40 HU, width 130 HU), and bone (center 400 HU, width 1800 HU) channels. For each window, raw HU pixel values were clipped to [C − W/2, C + W/2] and linearly rescaled to unsigned 8-bit integers [0,255] as pixel_out = 255 × (HU_clipped − (C − W/2))/W. The three resulting uint8 channels were concatenated to produce a 3-channel 512×512 image. RSNA DICOM files were converted to JPEG format via OpenCV (cv2.imwrite, IMWRITE_JPEG_QUALITY = 95) prior to training; this is a lossy compression step that quantifies the 8-bit windowed pixel values, and the mismatch between JPEG-compressed training data and raw DICOM external data is discussed as the primary driver of domain shift in Section 5.2. The images were downsampled to 256×256 via area interpolation (INTER_AREA) and normalized via ImageNet channel statistics (mean = [0.485, 0.456, 0.406]; std = [0.229, 0.224, 0.225]). During quality control, a lightweight Otsu-based threshold was

applied to the bone window channel to flag CT slices in which brain tissue coverage fell outside the range of 20–95%; no skull stripping was applied at the inference time.

3.5 Model Architecture

EfficientNet-B4 [15], pretrained on ImageNet [23], was loaded via the timm library [24] with the default classification head replaced by Dropout (p = 0.3) followed by Linear (1,792, 6). For training, a batch size of 128 slices per step was used. Two NVIDIA T4 GPUs were used with PyTorch DataParallel [25], with 64 slices distributed per GPU.

3.6 Training procedure

The model was trained via AdamW [26] (learning rate of 1×10^{-4}, weight decay of 1×10^{-2}) with cosine annealing scheduling (T_max = 50). Mixed-precision training [27] was enabled. Early stopping (patience = 7 epochs) was used to monitor the macro validation AUC-ROC. The best checkpoint saves the epoch index, model state, optimizer state, and validation AUC. Early termination occurred at epoch 13; the best checkpoint (epoch 6) was selected for all subsequent evaluations.

3.7 Evaluation Metrics

The per-class and macroaveraged AUC-ROC curves, sensitivity, specificity, and F1 score were computed. For the RSNA internal test set, a fixed threshold of 0.5 was applied. For CQ500, domain shift caused all sigmoid predictions to fall below 0.5 (maximum predicted probability = 0.468), yielding zero sensitivity; this is the primary negative finding. For the secondary evaluation of the CQ500, per-class Youden index thresholds were derived from the RSNA 2019 validation set and applied unchanged to the CQ500, preserving methodological independence. AUC-ROC is threshold independent and constitutes the primary cross-dataset performance measure.

The RSNA 2019 dataset is fully deidentified; StudyInstanceUIDs does not reliably map to unique patients. Study-level aggregation of RSNA test set predictions is therefore not feasible. Internal evaluation was performed at the slice level (n = 112,921), which is the native label granularity. Kang et al. 2023 benchmark [14] was reported at the study/exam level on a different dataset; direct numerical comparison is approximate and presented as contextual rather than equivalent.

3.8 Grad-CAM Explainability

Gradient-weighted class activation mapping (Grad-CAM) [18] was implemented by registering forward and backward hooks on the final convolutional block of EfficientNet-B4. For each target class, the gradient of the predicted logit with respect to the final convolutional feature map was globally average pooled to obtain per-channel importance weights. The resulting heatmap was ReLU-activated, upsampled to 256×256 via bilinear interpolation, normalized to [0,10], and overlaid as a jet colormap on the bone window channel at 50% opacity. For localization quality assessment, 10 heatmaps per class (n = 60 total) were randomly selected from true-positive test set predictions with a predicted probability ≥ 0.5, ensuring that the selection was representative rather than biased toward the highest-confidence cases.

Two board-certified radiologists independently scored all 60 heatmaps, blinded to each other's scores and used a five-point Likert scale to model architecture details: 1 = attention entirely outside the hemorrhage region; 2 = predominantly outside; 3 = partial overlap; 4 = predominantly within; and 5 = attention exclusively within the hemorrhage region. Interrater reliability was quantified via Cohen's weighted kappa (quadratic weights) and interpreted via the Landis & Koch scale [28].

3.9 External validation protocol

The CQ500 validation uses the epoch 6 checkpoint without retraining or fine-tuning. Slice-level sigmoid predictions were aggregated to the study level via mean pooling and then compared against majority-vote study-level labels. The AUC-ROC with 95% bootstrap confidence intervals (n = 1,000 study-level resamples) was the primary evaluation measure.

3.10 Statistical analysis

The 95% CI for AUC-ROC was estimated via bootstrap resampling (n = 1,000; slice level for the RSNA internal test set; study level for the CQ500). All analyses were performed in Python (scipy v1.11, NumPy v1.24).

3.11 Use of AI-assisted writing tools

During the preparation of this manuscript, the authors used Claude (claude.ai; Anthropic PBC, San Francisco, CA, USA) to assist with drafting, structural editing, and revision of the manuscript text. The AI tool was used solely to improve clarity, coherence, and language; it was not used in the design or conduct of the study, the training or evaluation of the model, the collection or analysis of data, the interpretation of results, or the clinical assessment of Grad-CAM heatmaps. All AI-assisted text was reviewed, edited, and verified by the authors, who take full responsibility for the accuracy, integrity, and scientific validity of all content in this article. Claude is not listed as an author and does not meet the ICMJE criteria for authorship.

4. Results

4.1 Model training

Training converged rapidly, with the validation macro AUC-ROC reaching 0.9837 at epoch 6 (training loss = 0.0553). Training loss decreased monotonically across logged epochs from 0.2002 (epoch 1) to 0.0154 (epoch 13). The validation AUC-ROC rose steeply from 0.9726 (epoch 1) to 0.9837 (epoch 6) and then plateaued while training loss continued to decline, which was consistent with the onset of overfitting and the rationale for early stopping. Early stops at epoch 13 after seven consecutive epochs without improvement (Fig. 2, Table 2).

Table 2
Validation performance across training epochs. Metrics from training logs. Early stopping is best marked with ★.

Epoch	Train loss	Val macro AUC	Sensitivity	Specificity
1	0.2002	0.9726	0.7606	0.9798
2	0.1420	0.9787	0.7970	0.9806
3	0.1149	0.9813	0.7970	0.9838
4	0.0910	0.9822	0.7940	0.9858
5	0.0692	0.9831	0.7912	0.9879
6 ★	**0.0553**	**0.9837**	**0.8207**	**0.9864**
7	0.0418	0.9827	0.7784	0.9904
8	0.0333	0.9831	0.8085	0.9877
9	0.0267	0.9821	0.7649	0.9917
10	0.0238	0.9825	0.7756	0.9917
11	0.0196	0.9835	0.7825	0.9909
12	0.0172	0.9834	0.7948	0.9906
13	0.0154	0.9822	0.7612	0.9930

★ *Best checkpoint selected for all subsequent evaluations.*

4.2 Internal Validation — RSNA Test Set

On the RSNA held-out test set (112,921 slices, threshold = 0.5), the model achieved a macro AUC-ROC of 0.9835 (95% CI 0.9821–0.9847), a sensitivity of 0.8192, a specificity of 0.9862, and an F1 of 0.7743 (Table 3). The per-class AUC-ROC ranged from 0.9758 (subarachnoid) to 0.9921 (intraventricular). Epidural hemorrhage, the most severely imbalanced class (w = 39.51), had an AUC-ROC of 0.9847, with a specificity of 0.9979. The macro AUC-ROC of 0.9835 contextually exceeds the Kang et al. 2023 benchmark of 0.953 [14], although this comparison is approximate given different evaluation units.

Table 3
Per-class performance on the RSNA 2019 internal test set (n = 112,921 slices, threshold = 0.50).

Class	AUC-ROC	95% CI	Sens.	Spec.	F1	n_pos
any_ich	0.9820	0.9811–0.9829	0.8434	0.9806	0.8605	16,084
epidural	0.9847	0.9777–0.9908	0.7821	0.9979	0.6845	459
intraparenchymal	0.9868	0.9855–0.9882	0.8511	0.9870	0.8043	5,279
intraventricular	0.9921	0.9910–0.9931	0.8769	0.9908	0.8221	3,900
subarachnoid	0.9758	0.9740–0.9775	0.7535	0.9808	0.7060	5,431
subdural	0.9793	0.9778–0.9807	0.8082	0.9804	0.7685	7,043
MACRO	**0.9835**	**0.9821–0.9847**	**0.8192**	**0.9862**	**0.7743**	—

95% CI: bootstrap (n = 1,000, slice level). Sens. = Sensitivity; Spec. = specificity; n_pos = positive slices in the test set.

4.3 External Validation — CQ500

Primary evaluation (threshold = 0.5): The maximum predicted sigmoid probability across all CQ500 slices was 0.468. All outputs fell below the fixed 0.5 threshold, yielding zero sensitivity across all classes (specificity = 1.0). This constitutes a primary negative finding: the model produces no positive predictions on the CQ500 data without recalibration.

Discriminative performance (AUC-ROC)

The macro AUC-ROC for the CQ500 was 0.7276 (95% CI 0.6729–0.7815; Table 4). Intraventricular hemorrhage demonstrated the highest degree of cross-dataset generalizability (AUC-ROC 0.8079, 95% CI 0.691–0.904). The subdural (0.6762) and epidural (0.6772, 95% CI 0.496–0.858) methods showed the poorest cross-dataset performance; the wide epidural CI reflects only 12 positive studies.

Secondary evaluation (RSNA-derived Youden thresholds)

At the RSNA validation set, the Youden index threshold (range 0.018–0.070), the macro sensitivity was 0.347, and the macro specificity was 0.937, reflecting systematic conservative predictions under domain shift. Intraventricular hemorrhage was the most clinically useful balance (sensitivity 0.731, specificity 0.808).

Table 4
Per-class performance on the CQ500 external validation dataset (n = 473 studies). Primary: AUC—ROC. Second, sensitivity, specificity, and F1 at the RSNA validation set Youden index thresholds were assessed, and no changes were detected for the CQ500.

Class	AUC	95% CI	Sens.	Spec.	F1	Thresh.	n_pos
any_ich	0.7298	0.6775–0.7802	0.4721	0.9348	0.6039	0.070	197
epidural	0.6772	0.4956–0.8582	0.0833	0.9610	0.0645	0.000	12
intraparenchymal	0.7098	0.6508–0.7646	0.1550	0.9884	0.2614	0.068	129
intraventricular	0.8079	0.6905–0.9044	0.7308	0.8076	0.2901	0.018	26
subarachnoid	0.7646	0.6852–0.8327	0.3333	0.9808	0.4524	0.043	57
subdural	0.6762	0.5703–0.7681	0.3061	0.9505	0.3529	0.053	49
MACRO	**0.7276**	**0.6729–0.7815**	**0.3468**	**0.9372**	**0.3375**	—	**205**

AUC-ROC: threshold-independent primary measure; 95% CI by study-level bootstrap (n = 1,000). Sens., Spec., F1: At the RSNA validation set, the Youden thresholds applied unchanged to the CQ500. Epidural threshold 0.000: Under domain shift, no RSNA validation slice prediction exceeded the Youden operating point for this class, indicating that the model's epidural output distribution shifted below the RSNA-derived threshold; de facto, no CQ500 slice was classified as epidural using this threshold. n_pos: positive studies in CQ500. The prevalence of the CQ500 class was as follows: any_ich, 43.3%; intraparenchymal, 28.3%; subarachnoid, 12.7%; subdural, 11.2%; intraventricular, 5.9%; and epidural, 2.7%.

4.4 Grad-CAM Explainability

Two board-certified radiologists independently scored Grad-CAM heatmaps for 60 randomly selected true-positive test set predictions (10 per class, predicted probability ≥ 0.5) via a five-point localization scale (Fig. 4). The combined mean localization score was 3.81 ± 0.86/5 (Table 5). Interrater reliability was κ = 0.653, indicating substantial agreement (Landis & Koch scale) [29]. Exact agreement was achieved on 47 of 60 images (78.3%), and all 60 images (100%) had within-one-point agreement, confirming scale stability. Intraparenchymal hemorrhage achieved the highest localization scores (mean of 4.40/5 for both raters), which is consistent with its large, visually distinct parenchymal signal. Subdural hemorrhage showed perfect interrater agreement (κ = 1.000), which was consistent with the anatomically unambiguous crescent-shaped convexity pattern. Any_ich showed the lowest mean score (3.30/5) and lowest per-class kappa (0.467), which was expected given its heterogeneous composite nature spanning all five subtypes. Epidural hemorrhage showed near-perfect interrater agreement (κ = 0.808), which was consistent with the anatomically constrained peripheral location.

Table 5
Radiologist Grad-CAM localization scores by two independent board-certified radiologists (n = 60 heatmaps; 10 per class; randomly selected true-positive predictions; threshold ≥ 0.50).

Class	Rater 1 (mean ± SD)	Rater 2 (mean ± SD)	Combined mean	Cohen's κ	Interpretation
any_ich	3.20 ± 1.08	3.40 ± 1.28	3.30	0.467	Moderate
epidural	3.60 ± 0.49	3.50 ± 0.67	3.55	0.808	Almost perfect
intraparenchymal	4.40 ± 0.49	4.40 ± 0.49	4.40	0.583	Moderate
intraventricular	3.90 ± 0.70	3.60 ± 0.80	3.75	0.531	Moderate
subarachnoid	4.00 ± 1.00	4.10 ± 0.94	4.05	0.565	Moderate
subdural	3.80 ± 0.40	3.80 ± 0.40	3.80	1.000	Perfect
Overall	**3.82 ± 0.83**	**3.80 ± 0.89**	**3.81 ± 0.86**	**0.653**	**Substantial**

Scores: 1 = attention entirely outside the hemorrhage; 5 = attention exclusively within the hemorrhage. Interrater reliability: Cohen's weighted kappa (quadratic weights); Landis & Koch (1977) [28] scale. Exact agreement: 47/60 (78.3%); within-one-point agreement: 60/60 (100%). κ interpretation: <0.20 slight; 0.21–0.40 fair; 0.41–0.60 moderate; 0.61–0.80 substantial; 0.81–1.00 almost perfect.

5. Discussion

5.1 Internal performance in the context of the literature

The model achieved a macro AUC-ROC of 0.9835 (95% CI 0.9821–0.9847) on the RSNA internal test set. This value is numerically greater than the Kang et al. 2023 benchmark of 0.953 [14] and falls within the pooled meta-analysis estimate of 0.96 [3]. As discussed in Section 4.2, this comparison is not equivalent (slice-level versus exam-level evaluation) and should be interpreted as contextual rather than confirmatory. The high macro specificity of 0.9862 indicates a low false-positive rate, which is important for clinical screening workflows [1,23]. Epidural hemorrhage, the most severely imbalanced class (0.42% prevalence, w = 39.51), achieved an AUC-ROC of 0.9847, demonstrating that the class-balanced weighting formula effectively prevents minority class collapse. The sensitivity of 0.7821 for the epidural method reflects the intrinsic difficulty of detecting a class represented by only 459 positive test slices, which is consistent with the literature [3,5].

5.2 Domain Shift and Cross-Dataset Generalization

The most notable finding was complete threshold failure on the CQ500: all sigmoid outputs fell below 0.5, rendering the model clinically innovative without recalibration. The discriminative AUC-ROC of 0.7276 confirms the retained separation power. Domain shift is attributable to multiple concurrent

factors: (i) image format differences—JPEG-compressed RSNA data versus raw DICOM CQ500 acquisitions, producing systematically lower sigmoid outputs; (ii) scanner heterogeneity—CQ500 acquired on Siemens and Philips scanners across varied slice thicknesses (1.25–5 mm) and reconstruction kernels versus the curated RSNA multi-institutional collection; (iii) differences in slice thickness altering partial volume effects, particularly for thin extra-axial collections; (iv) population-level differences in skull morphology, brain atrophy, and hemorrhage etiology; and (v) structural label noise differences—RSNA labels propagated from the study level versus CQ500's majority-vote study-level labels, potentially contributing to conservative sigmoid outputs. These findings reinforce the conclusions of Voter et al. [5] and Salehinejad et al. [7] that benchmark performance does not predict deployment performance and highlight threshold recalibration as a mandatory predeployment step. Intraventricular hemorrhage showed the most consistent cross-dataset performance (AUC-ROC 0.8079, sensitivity 0.731 at the RSNA-derived threshold) and represented the most clinically deployable output class under domain shift. The anatomically distinctive periventricular distribution of intraventricular blood produces a high-contrast signal that is relatively robust to scanner and acquisition variation, a finding with practical implications for triage prioritization in LMIC deployment. The epidural AUC-ROC for the CQ500 (0.6772, 95% CI 0.496–0.858) is statistically unreliable given that only 12 positive studies exist and should not be used for subtype-specific comparisons.

5.3 Epidural Hemorrhage: Rare Subtype Performance

Epidural hemorrhage presents a dual challenge: extreme class imbalance (0.42%, w = 39.51) and a high dependence on anatomically precise skull–brain interface signals. The class-balanced weighting formula successfully suppressed false positives (specificity of 0.9979 internally), but the sensitivity (0.7821) reflects the difficulty of correctly classifying a small number of positive slices. A dataset with substantially greater epidural incidence is needed for robust cross-domain assessment.

5.4 Grad-CAM and clinical interpretability

The two-rater Grad-CAM evaluation (n = 60, κ = 0.653, substantial agreement) provides credible evidence that the model's attention is broadly consistent with the expected anatomical distribution of each ICH subtype. The combined mean score of 3.81 ± 0.86 / 5 (ranging from 3.30–4.40 across subtypes) indicates that both radiologists consistently judged model attention predominantly within the hemorrhage region. These findings are consistent with those of Lee et al. [10] and Kim et al. [19], who demonstrated anatomically grounded Grad-CAM attention in CNN-based ICH classifiers. Notably, the intraparenchymal subtype had the highest mean (4.40 / 5), which was consistent with its large, visually distinctive parenchymal signal; the subdural subtype achieved perfect interrater agreement (κ = 1.000), which was consistent with the unambiguous crescent-shaped convexity pattern; and any_ich scored lowest (3.30 / 5, κ = 0.467), which was expected given its heterogeneous composite nature. The evaluation is limited to predictions with probabilities ≥ 0.5; attention near the decision boundary was not assessed and remains a direction for future work.

5.5 Limitations

Label noise (slice-level label propagation). The RSNA 2019 dataset provides only study-level labels. Slice-level training labels are generated by propagating the study-level label to every slice in a study, meaning that every slice in a positive study is labeled positive regardless of whether hemorrhage is visible on that slice, and every slice in a negative study is labeled negative. This introduces bidirectional label noise. Given the focal and spatially limited nature of most intracranial hemorrhages relative to standard CT slice thickness (typically 5 mm), it is clinically reasonable to expect that only a minority of slices within a positive study contain visible hemorrhage, implying that the majority of 'positive' training slices may not exhibit visible pathology. This noise almost certainly inflates slice-level specificity and sensitivity estimates and likely inflates the AUC-ROC relative to what a model trained on slice-confirmed labels would achieve. It may also encourage the model to learn study-level features (e.g., patient-specific anatomy, scanner artifacts) rather than slice-level pathology. The internal AUC-ROC of 0.9835 should therefore be interpreted as an upper bound estimate. A fully slice-annotated dataset, such as a DICOM-native dataset with per-slice radiologist annotation, is needed to obtain unbiased slice-level performance estimates. The model operates at the slice level without interslice volumetric context, limiting detection of small or thin hemorrhages more apparent across consecutive slices. A lightweight Otsu-based skull mask was used for quality control rather than a validated deep learning skull stripping tool (e.g., HD-BET), which may affect parenchymal signal isolation. The study-level split is a proxy for patient-level separation in this deidentified dataset. JPEG compression at a quality factor of 95 was applied to training data only; the mismatch with raw DICOM external data was a primary contributor to domain shift and threshold failure. The Grad-CAM radiologist assessment (n = 60, two raters) is limited to predictions with probabilities ≥ 0.5; model attention near the decision boundary was not assessed. The external validation was limited to a single external cohort (CQ500).

5.6 Future directions

Domain adaptation techniques should be evaluated to mitigate the JPEG-to-DICOM intensity distribution mismatch. Candidate approaches include histogram equalization (matching CQ500 intensity histograms to the RSNA training distribution before inference), standard unit normalization (z score normalization applied to HU values prior to windowing), and test-time adaptation methods that adjust the model's feature statistics to the target domain. Threshold calibration on a small institutional validation set should be formalized as a mandatory predeployment step. Extension to three-dimensional volumetric models would provide the interslice context currently unavailable. Prospective clinical studies at PAHS and other Nepali emergency departments would directly assess the pipeline's utility in LMIC settings. The preliminary Grad-CAM radiologist assessment should be extended to a larger multirater study with formal interrater reliability measurements across the full range of predicted probabilities. Validation on additional external cohorts with independently acquired DICOM data would further characterize cross-domain generalization.

6. Conclusion

EfficientNet-B4, trained on 752,803 RSNA 2019 CT slices with class-balanced weighted loss, achieves a macro AUC-ROC of 0.9835 (95% CI 0.9821–0.9847) on the internal test set. External validation of the CQ500 reveals complete threshold failure under domain shift—a critical negative finding that must be foregrounded in any deployment context. The threshold-independent AUC-ROC of 0.7276 demonstrated retained discriminative capacity, with intraventricular hemorrhage showing the most consistent cross-dataset generalization. The Grad-CAM attention maps received a radiologist localization score of 3.81/5 across all six ICH subtypes, confirming anatomically plausible model attention. These findings establish threshold recalibration on a representative local validation set as a nonnegotiable prerequisite for deployment, and the open-source pipeline is released to facilitate adaptation in resource-limited healthcare settings.

Declarations

Ethics statement: This study used only publicly available, fully deidentified datasets (RSNA 2019 and CQ500). No patient contact or new data collection was performed. Ethical approval was not needed.

Data availability: RSNA 2019: https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection. CQ500: http://headctstudy.qure.ai. Code and model weights: https://github.com/lochanshrestha-dev/ich-detection-efficientnet-

Author contributions (CRediT): Lochan Shrestha: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Huma Subhani:** Conceptualization, Resources, Writing – Review & Editing, Supervision, Project administration.

Generative AI and AI-assisted technologies in the writing process During the preparation of this work, the authors used Claude (Anthropic) to assist with manuscript drafting, structural editing, and revision. The authors reviewed, edited, and critically evaluated all AI-assisted content and take full responsibility for the integrity, accuracy, and scientific validity of the published article. No AI tool was used in the design, conduct, or analysis of the study, in the interpretation of results, or in the clinical assessment of Grad-CAM heatmaps.

Competing interests: The authors declare that they have no competing interests.

Funding: No external funding was received for this study.

Acknowledgments: The authors thank the RSNA and Qure.ai for making their datasets publicly available and Kaggle for providing computational resources.

References

1. Caceres, J. A. & Goldstein, J. N. Intracranial hemorrhage. *Emerg. Med. Clin. North. Am.* **30**, 771–794 (2012).
2. van Asch, C. J. et al. Incidence, case fatality, and functional outcome of intracerebral hemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. *Lancet Neurol.* **9**, 167–176 (2010).
3. Karamian, A. & Seifi, A. Diagnostic Accuracy of Deep Learning for Intracranial Hemorrhage Detection in Non-Contrast Brain CT Scans: A Systematic Review and Meta-Analysis. *J. Clin. Med.* **14**, 2377 (2025).
4. Kuo, W., Häne, C., Mukherjee, P., Malik, J. & Yuh, E. L. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc. Natl. Acad. Sci.* 116, 22737–22745 (2019).
5. Voter, A. F., Meram, E., Garrett, J. W. & Yu, J. P. J. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Intracranial Hemorrhage. *J. Am. Coll. Radiol. JACR.* **18**, 1143–1152 (2021).
6. Nada, A. et al. External validation and performance analysis of a deep learning-based model for the detection of intracranial hemorrhage. *Neuroradiol. J.* **38**, 312–321 (2025).
7. Salehinejad, H. et al. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. *Sci. Rep.* **11**, 17051 (2021).
8. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. in 2980–2988 (2017).
9. Flanders, A. E. et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiol. Artif. Intell.* **2**, e190211 (2020).
10. Lee, H. et al. An explainable deep-learning algorithm for the detection of acute intracranial hemorrhage from small datasets. *Nat. Biomed. Eng.* **3**, 173–182 (2019).
11. Chilamkurthy, S. et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* **392**, 2388–2396 (2018).
12. Burduja, M., Ionescu, R. T. & Verga, N. Accurate and Efficient Intracranial Hemorrhage Detection and Subtype Classification in 3D CT Scans with Convolutional and Long Short-Term Memory Neural Networks. *Sensors* **20**, 5611 (2020).
13. D'Angelo, T. et al. Accuracy and time efficiency of a novel deep learning algorithm for Intracranial Hemorrhage detection in CT Scans. *Radiol. Med. (Torino)*. **129**, 1499–1506 (2024).
14. Kang, D. W. et al. Strengthening deep-learning models for intracranial hemorrhage detection: strongly annotated computed tomography images and model ensembles. *Front. Neurol.* **14**, 1321964 (2023).
15. Tan, M., Le, Q. & EfficientNet Rethinking Model Scaling for Convolutional Neural Networks. in *Proceedings of the 36th International Conference on Machine Learning* 6105–6114PMLR, (2019).

16. Park, S. H. et al. Comparison between single and serial computed tomography images in classification of acute appendicitis, acute right-sided diverticulitis, and normal appendix using EfficientNet. *PLOS ONE*. **18**, e0281498 (2023).
17. Kok, Y. E. et al. Semantic Segmentation of Spontaneous Intracerebral Hemorrhage, Intraventricular Hemorrhage, and Associated Edema on CT Images Using Deep Learning. *Radiol. Artif. Intell.* **4**, e220096 (2022).
18. Selvaraju, R. R. et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. in 618–626 (2017).
19. Kim, K. H., Koo, H. W., Lee, B. J., Yoon, S. W. & Sohn, M. J. Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning. *J. Korean Phys. Soc.* **79**, 321–327 (2021).
20. Altuve, M. & Pérez, A. Intracerebral hemorrhage detection on computed tomography images using a residual neural network. *Phys. Med. PM Int. J. Devoted Appl. Phys. Med. Biol. Off J. Ital. Assoc. Biomed. Phys.* **99**, 113–119 (2022).
21. Inkeaw, P. et al. Automatic hemorrhage segmentation on head CT scan for traumatic brain injury using 3D deep learning model. *Comput. Biol. Med.* **146**, 105530 (2022).
22. Vidhya, V. et al. YOLOv5s-CAM: A Deep Learning Model for Automated Detection and Classification for Types of Intracranial Hematoma in CT Images. *IEEE Access.* **11**, 141309–141328 (2023).
23. Deng, J. et al. ImageNet: A Large-Scale Hierarchical Image Database.
24. Wightman, R. & PyTorch *Image Models* https://doi.org/10.5281/zenodo.4414861 (2026).
25. Paszke, A. et al. Curran Associates, Inc.,. PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* vol. 32 (2019).
26. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. in (2018).
27. Micikevicius, P. et al. *Mixed Precision Train.* in (2018).
28. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33**, 159–174 (1977).
29. Cohen, J. A. Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).

Figures

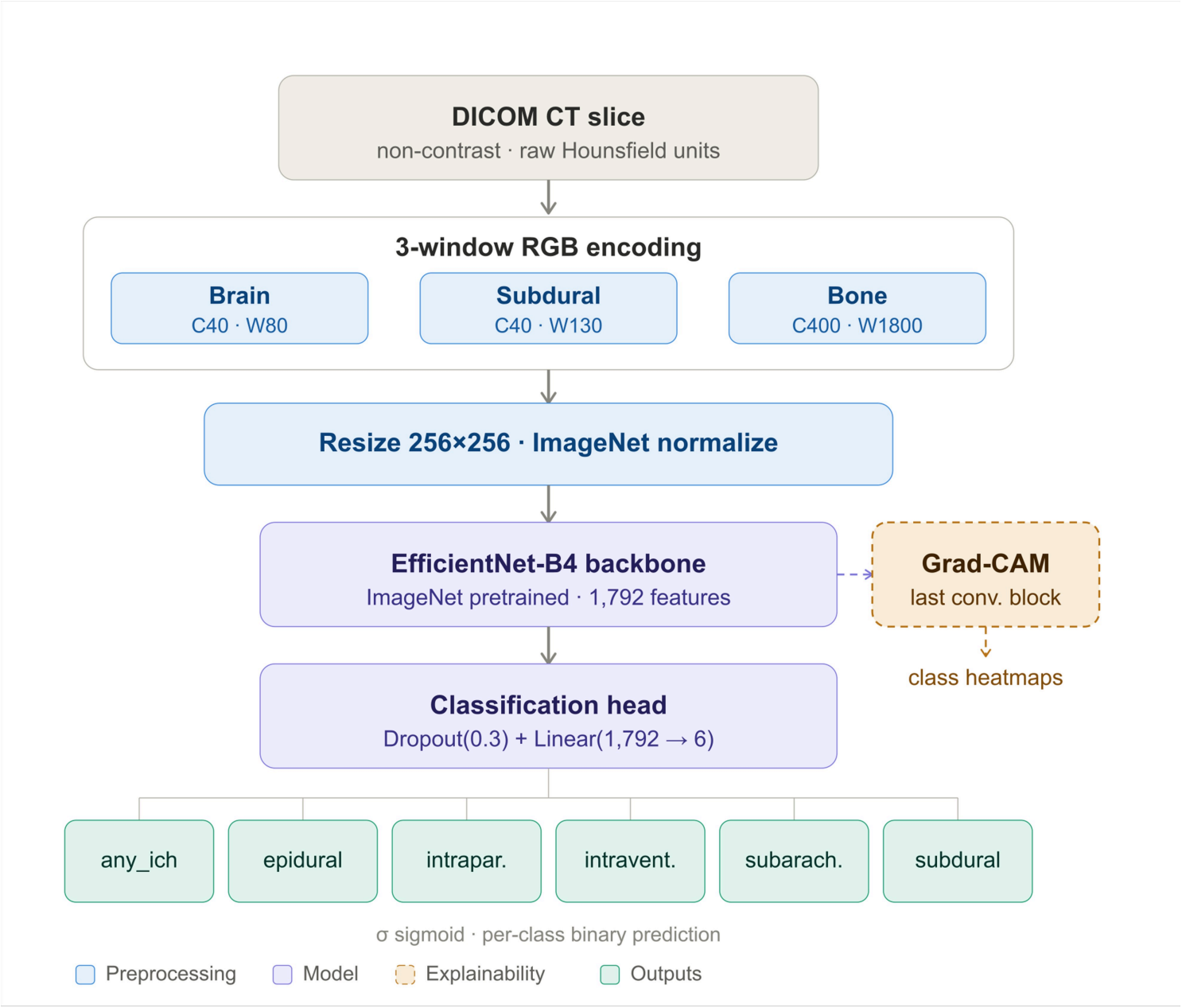


Figure 1

Intracranial hemorrhage (ICH) classification pipeline. he system processes raw noncontrast DICOM CT slices through a 3-window RGB encoding strategy (brain, subdural, and bone windows) and standardizes them via 256 × 256 resizing and ImageNet normalization. These processed images are analyzed by an EfficientNet-B4 backbone to extract 1,792 features, which are then passed through a classification head—comprising a 0.3 dropout layer and a linear transformation—to predict six distinct hemorrhage categories via sigmoid activation. Model transparency is provided by Grad-CAM class heatmaps generated from the final convolutional block.

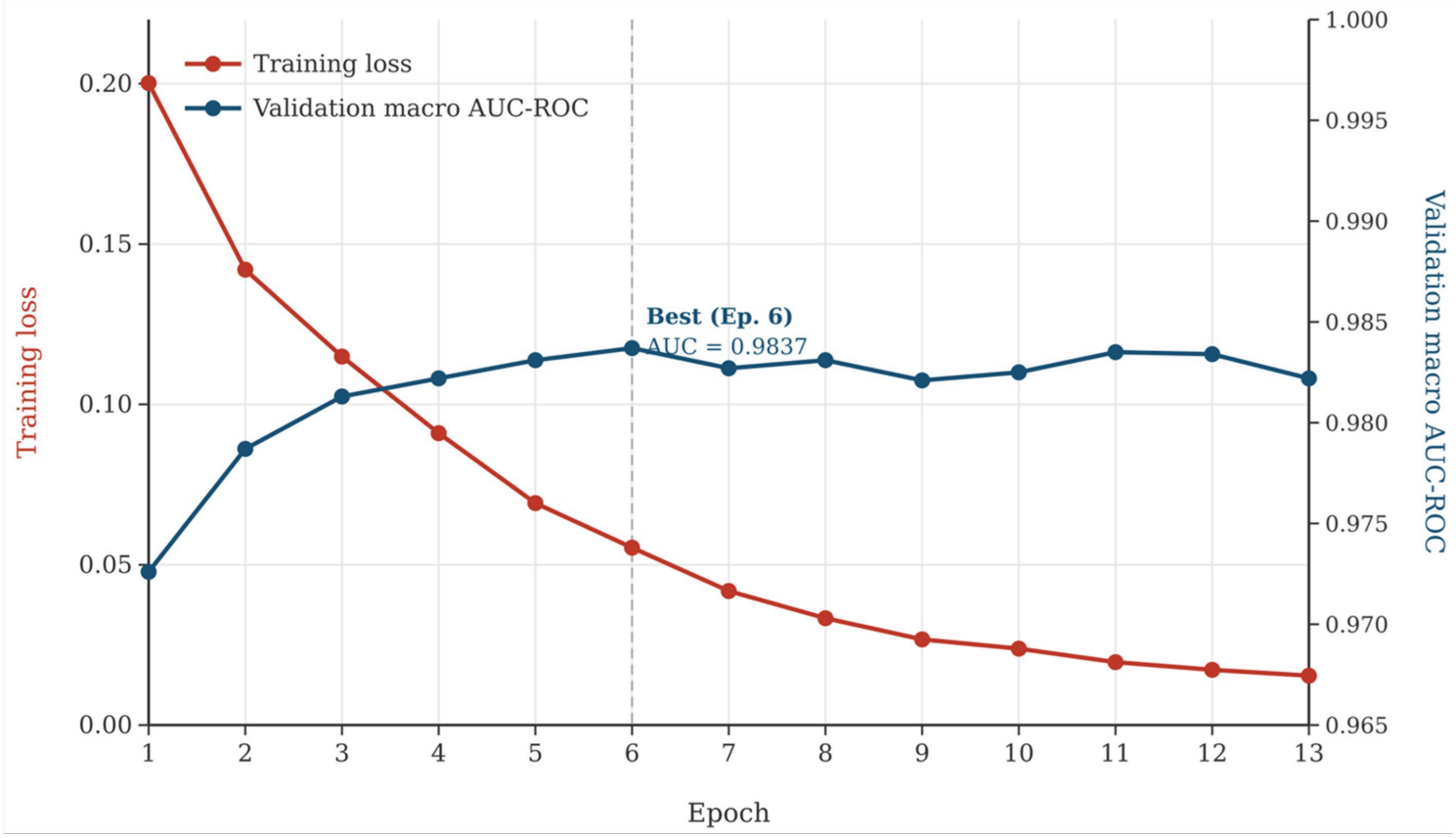


Figure 2

Training dynamics: EfficientNet-B4 on RSNA 2019. Left y-axis (red): training loss. Right y-axis (blue): validation macro AUC-ROC curve. Dashed vertical line: best checkpoint (epoch 6, AUC = 0.9837). Early stopping at epoch 13.

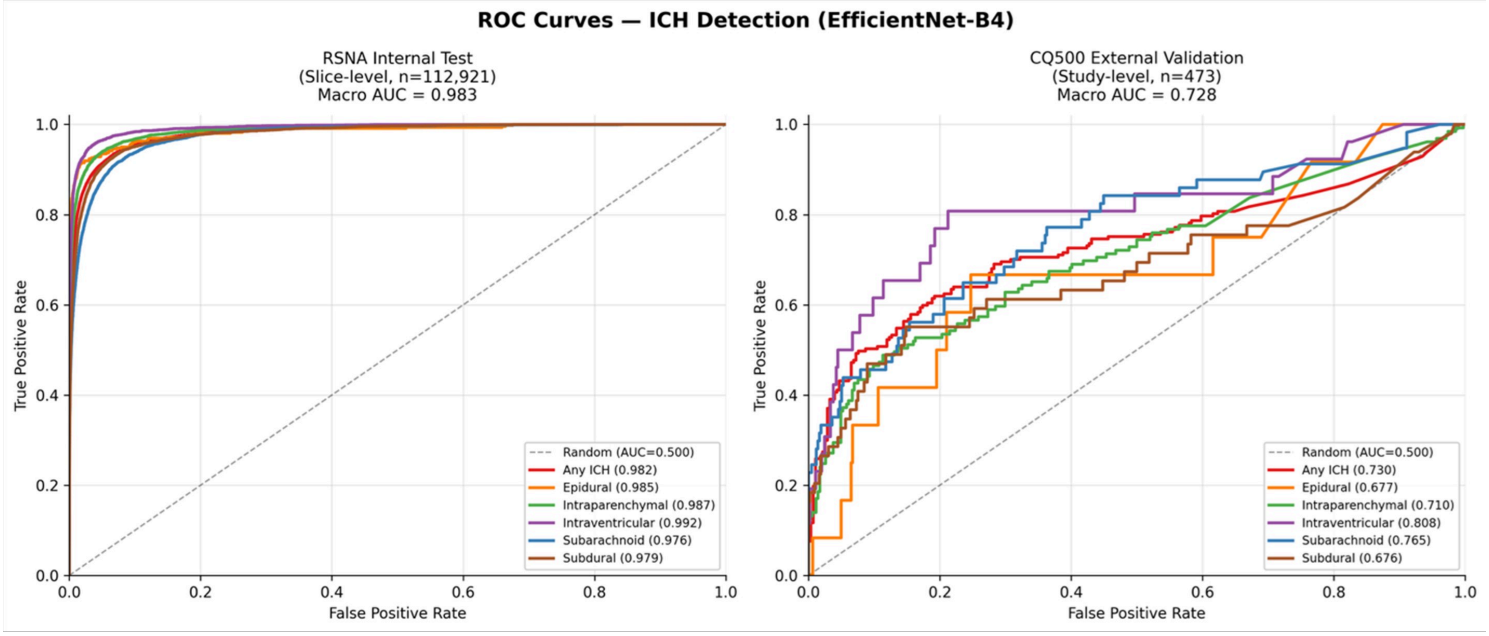


Figure 3

Empirical ROC curves. Left: RSNA 2019 internal test set (slice level, n = 112,921), macro AUC = 0.983. Right: CQ500 external validation (study-level, n = 473), macro AUC = 0.728. The stepped appearance of the CQ500 curves reflects small positive study counts per class. Diagonal dashed line: reference classifier (AUC = 0.50).

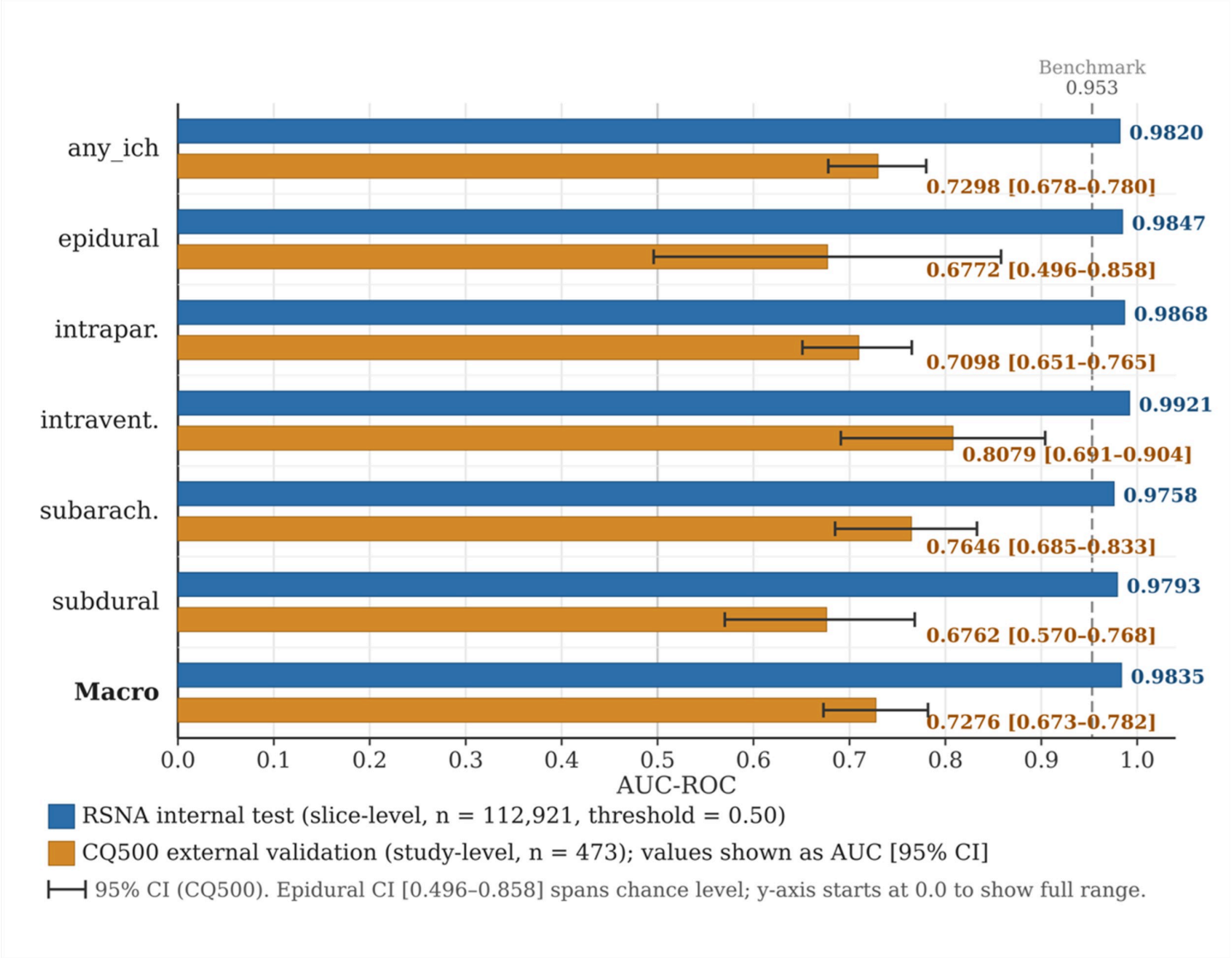


Figure 4

Figure 5. *AUC-ROC comparison: RSNA internal test set (blue bars, slice level, threshold = 0.50) vs CQ500 external validation (amber bars, study level, AUC-ROC only). Error bars: 95% bootstrap CI (CQ500). Horizontal dashed line: Kang* et al. *2023 benchmark (AUC = 0.953).*

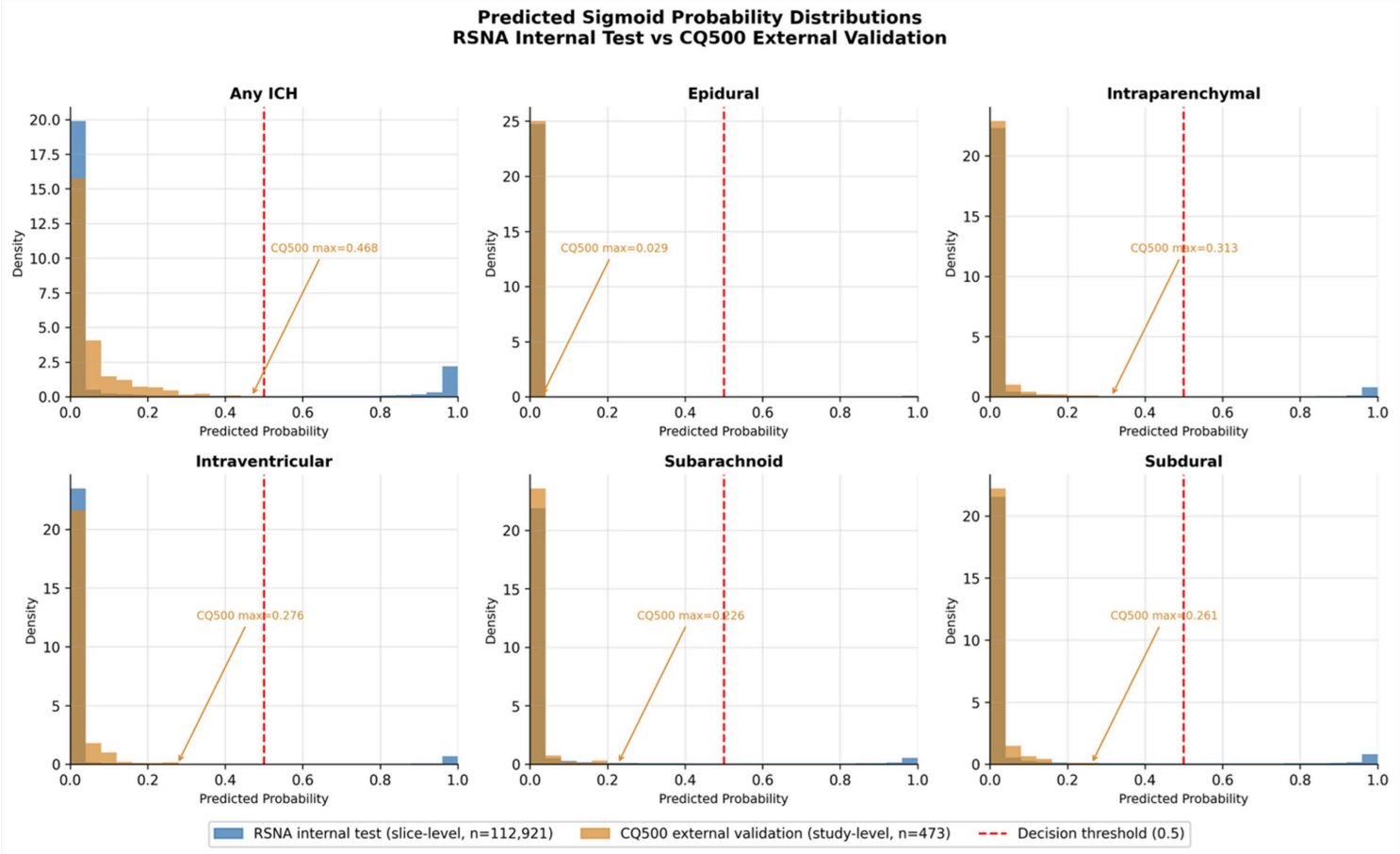


Figure 5

Figure 6. *Predicted sigmoid probability distributions for the RSNA 2019 internal test set (blue, slice-level, n = 112,921) and the CQ500 external validation set (amber, study-level, n = 473) per ICH subtype. The RSNA distribution spans the full [0, 1] range, indicating a well-calibrated internal classifier. The CQ500 predictions are entirely concentrated near zero, with no class exceeding the decision threshold of 0.5 (red dashed line). The per-class maximum CQ500 values were as follows: any_ich 0.468, epidural 0.029, intraparenchymal 0.313, intraventricular 0.276, subarachnoid 0.226, and subdural 0.261. The epidural maximum of 0.029 directly explains the degenerate Youden threshold of 0.000 in Table 4. This systematic compression of CQ500 predictions is consistent with the JPEG-to-DICOM intensity distribution mismatch identified as the primary driver of domain shift (Section 5.2).*

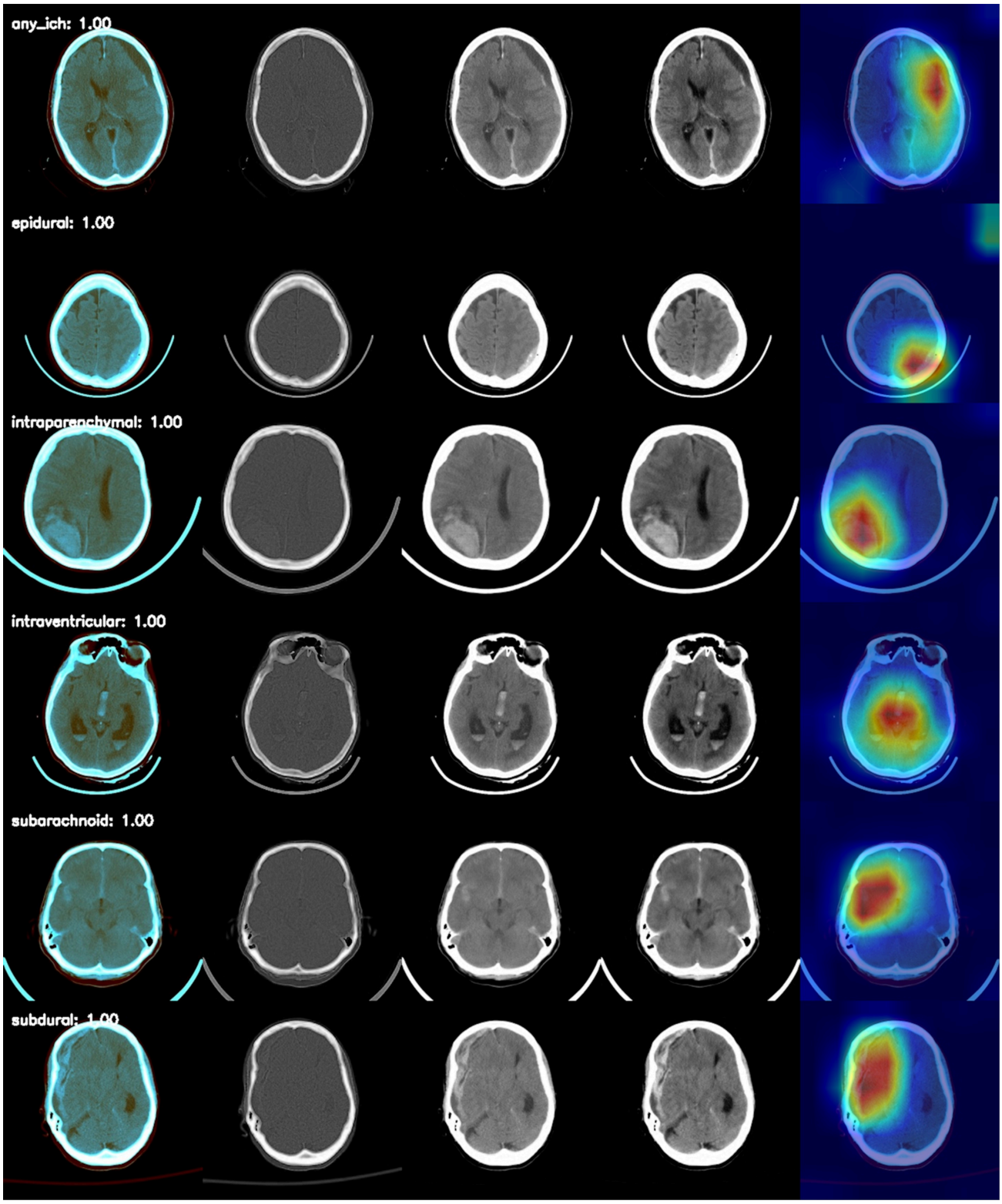


Figure 6

Figure 4. *Grad-CAM class-specific attention heatmaps (highest-confidence prediction per ICH subtype, predicted probability ≈ 1.00). Jet color map (blue = low, red = high activation) overlaid on the bone window channel at 50% opacity. Expected anatomical localization confirmed for all subtypes. Full evaluation: Sixty randomly selected true-positive predictions (10 per class) were scored by two independent radiologists; the combined mean score was 3.81 ± 0.86 / 5 (κ = 0.653). See Table 5 for per-class breakdown.*