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.

Diagnostic Accuracy and External Validation of Self-Supervised Learning for Cerebral Micro-Bleed Detection: A Multi-Sequence MRI Trial Using Public Datasets

Rameswari Poornima Janardanan
rjanardanan@inaya.edu.sa
Inaya Medical Colleges
Elamir A. Osman
Inaya Medical Colleges
Omer O. Saeed
Inaya Medical Colleges
Mahmoud Eltahir Ali
Inaya Medical Colleges
Omer Gaddoura
Inaya Medical Colleges

Research Article

Keywords: Cerebral micro-bleeds, self-supervised learning, magnetic resonance imaging, deep learning, medical image analysis

Posted Date: April 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9496185/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Purpose

Cerebral microbleeds (CMBs) are critical imaging biomarkers for small vessel disease, but detection remains challenging due to small lesion size, variable MRI appearance, and annotation burden. This study developed a self-supervised learning (SSL) framework for robust CMB detection across multi-sequence MRI that generalizes to heterogeneous protocols while reducing dependence on labeled data.

Materials and Methods

An SSL framework (3D ResNet-18 with Barlow Twins loss) was pretrained on 2,450 unlabeled multi-sequence MRI scans (MICCAI 2022, VALDO, UK Biobank), then fine-tuned with only 400 labeled scans using a 3D U-Net for voxel-level detection. Performance was evaluated using ROC-AUC, sensitivity, false positives per scan, lesion-level F1-score, and cross-sequence generalization.

Results

The SSL framework achieved an AUC of 0.92 (95% CI: 0.90–0.94), sensitivity of 81%, and 1.1 false positives per scan—outperforming fully supervised (AUC 0.84) and semi-supervised (AUC 0.90) baselines. The model maintained robust performance across SWI (AUC 0.93), GRE (AUC 0.90), and 3T scanners (AUC 0.92), with lesion-level F1-scores of 78–84%. SSL pretraining enabled stable detection with as few as 100 labeled scans (AUC 0.90), demonstrating substantial annotation efficiency.

Conclusion

Self-supervised learning enables robust, generalizable CMB detection across heterogeneous multi-sequence MRI while significantly reducing annotation requirements. The framework's strong cross-sequence generalization supports its potential as a scalable clinical decision-support tool, though prospective validation in independent cohorts remains necessary.

1. Introduction

Cerebral micro-bleeds (CMBs) are small, rounded, hypointense lesions detected on susceptibility-sensitive MRI sequences such as susceptibility-weighted imaging (SWI) and gradient-recalled echo (GRE). Histopathologically, they correspond to focal deposits of hemosiderin resulting from prior micro hemorrhages and are strongly associated with hypertensive arteriopathy and cerebral amyloid angiopathy [1]. The presence, number, and anatomical distribution of CMBs have been linked to an increased risk of intracerebral hemorrhage [2], ischemic stroke recurrence [3], cognitive impairment [4], and poor functional outcomes, making them clinically relevant imaging biomarkers. Despite their importance, accurate identification of CMBs in routine clinical practice remains difficult. Manual visual assessment is time-consuming and subject to substantial inter- and intra-observer variability, particularly in cases with multiple lesions or confounding mimics such as calcifications, vascular flow voids, or imaging artifacts [5]. These challenges are exacerbated by the increasing volume of neuroimaging studies and the growing complexity of MRI protocols in modern clinical workflows. Deep learning–based automated detection methods have been proposed to address these limitations and have demonstrated encouraging performance in research settings [6]. Prior supervised DL models achieved Area Under the Curves (AUCs) of X–Y% but showed substantial performance drop (Δ%) across scanner types/sequences [7]. However, most existing approaches rely heavily on fully supervised learning and require large amounts of voxel-level expert annotations, which are costly and difficult to obtain [8]. Existing DL models fail to generalize across multi-sequence MRI, limiting clinical deployment. There is a need for SSL frameworks that learn modality-invariant features using large unlabeled datasets. Furthermore, many models exhibit limited generalizability when applied to data acquired using different MRI sequences, scanner vendors, or acquisition parameters, significantly limiting their real-world applicability [9].

Self-supervised learning (SSL) has recently emerged as a powerful paradigm for representation learning in medical imaging by leveraging large volumes of unlabeled data to learn informative features without explicit annotations [10]. While SSL has shown promise in various neuroimaging tasks, most prior studies have focused on single-modality learning and have not explicitly addressed the substantial heterogeneity introduced by multi-sequence MRI [11]. Consequently, there remains a critical need for SSL frameworks that can effectively learn modality-invariant representations and support robust CMB detection across diverse imaging settings. The target population for this study consists of adult patients undergoing brain MRI for evaluation of cerebrovascular disease, stroke, cognitive impairment, or related neurological conditions. The intended use of the proposed model is as a clinical decision-support tool to assist neuro-radiologists and clinicians in identifying cerebral micro-bleeds during routine image interpretation. The model is designed to generate voxel-level probability maps highlighting regions likely to contain CMBs, which can be overlaid on standard MRI sequences to draw attention to subtle lesions that may otherwise be overlooked. Importantly, the model is not intended to replace expert clinical judgment or serve as an autonomous diagnostic system. Instead, it is envisioned as an assistive tool that may improve detection sensitivity, reduce interpretation time, and enhance consistency across readers and institutions, particularly in high-volume or resource-limited settings.

Health inequities in neuroimaging-based diagnosis often arise from variability in imaging infrastructure, scanner availability, and acquisition protocols across healthcare systems and geographic regions [12]. In the context of cerebral micro-bleed detection, differences in MRI sequence availability and quality can substantially influence diagnostic accuracy. Rather than focusing on patient-level sociodemographic factors, this study addresses health inequalities related to technical and infrastructural variability by explicitly evaluating model performance across multiple MRI sequences and datasets derived from different centers. By emphasizing cross-sequence and cross-domain generalization, the proposed approach aims to reduce performance disparities driven by imaging

protocol heterogeneity and support more equitable access to reliable automated CMB detection tools. The primary objective of this study was to develop and validate a self-supervised deep learning framework for voxel-level detection of cerebral micro-bleeds across multi-sequence MRI. Secondary objectives included evaluating the impact of self-supervised pre-training on annotation efficiency, assessing robustness under limited labeled data scenarios, and analyzing generalization performance across different MRI sequences and datasets. An additional objective was to establish methodological transparency, quantify clinical relevance, and provide actionable recommendations for future prospective validation.

2. Methods

2.1 Data Sources and reporting

This study utilized MRI data from three publicly available datasets: the MICCAI 2022 Cerebral Micro-bleed Detection Challenge dataset [13], the Vascular Lesions Detection and Outcomes (VALDO) dataset [14], and a subset of the United Kingdom Biobank imaging cohort [15]. These datasets were selected to ensure diversity in imaging sequences, scanner vendors, field strengths, and acquisition protocols, thereby enabling comprehensive evaluation of model robustness and generalizability. The MICCAI 2022 dataset consists of multi-center MRI scans with expert-annotated cerebral micro-bleeds and was specifically curated for benchmarking automated CMB detection algorithms. The VALDO dataset provides imaging data from patients with vascular brain lesions, including CMBs, acquired using heterogeneous clinical protocols. The UK Biobank subset includes large-scale population-based MRI data with standardized acquisition, offering an opportunity to leverage extensive unlabeled data for self-supervised pre-training. These datasets collectively enabled the evaluation of model performance across diverse scanners, sequences, and patient populations. The MRI scans included in this study were acquired between 2015 and 2022. To account for potential temporal heterogeneity such as scanner upgrades, changes in acquisition protocols, or sequence modifications over the years, datasets were partitioned during training, validation, and testing to ensure that temporal variability was represented across all subsets. Datasets were split at the subject level, ensuring that scans from the same participant were never present in more than one subset, thereby eliminating potential data leakage and overestimation of performance. As the analysis was cross-sectional and focused on imaging-based lesion detection, no longitudinal follow-up or outcome assessment was performed. This study adhered to the *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis – Artificial Intelligence extension (TRIPOD-AI) and Model Information – Clinical Artificial Intelligence Model (MI-CLAIM)* guidelines [16]. These frameworks ensure transparent reporting of dataset provenance, preprocessing, model architecture, training, validation, and performance evaluation for AI-based diagnostic imaging studies using secondary public datasets.

2.2 Study Setting and Participants

All MRI scans underwent a standardized preprocessing and quality control pipeline to ensure consistency while preserving the inherent heterogeneity of the multi-center data. This approach aimed to minimize technical bias while allowing the evaluation of model robustness under real-world imaging variations. **Figure 1** provides a representative visualization of the raw multi-sequence MRI data from different scanners and sequences before preprocessing, illustrating the inherent heterogeneity in contrast, resolution, and appearance that the model must generalize across.

Preprocessing: Each volume was processed with the following steps: 1) skull stripping using an automated tool (HD-BET); 2) resampling to a uniform 1 mm^3 isotropic resolution via trilinear interpolation; and 3) per-sequence intensity normalization to zero mean and unit variance [17,18]. Susceptibility distortion correction was applied where applicable. To ensure spatial alignment across different MRI sequences for each subject, all multi-sequence data were co-registered to a common reference space. The high-resolution susceptibility-weighted imaging (SWI) volume was typically used as the reference, though a T1-weighted anatomical scan was employed when available for improved anatomical consistency [19]. This alignment was performed using affine registration with the Advanced Normalization Tools (ANTs) software, optimizing mutual information to account for the differing contrast mechanisms across sequences. All transformed images were resampled to the reference space using trilinear interpolation, resulting in perfectly aligned multi-channel inputs for the model.

Quality Control: Prior to inclusion, all scans underwent rigorous quality assessment. This included visual inspection for exclusion criteria (severe motion artifacts, incomplete coverage, corruption) and automated checks for outlier intensity distributions and spatial inconsistencies post-preprocessing. This uniform QC protocol ensured model performance reflected true detection capability rather than sensitivity to artifacts.

Data Augmentation: To enhance generalizability across scanners and protocols, extensive augmentation was applied during training. Spatial transformations included random rotations (±15°), scaling (0.9–1.1×), flipping (sagittal/coronal), and elastic deformations. Intensity augmentations comprised Gaussian noise injection (σ=0.01–0.05) and contrast adjustments (±20%) [20]. During supervised fine-tuning, all geometric transformations were applied synchronously to image patches and their corresponding voxel-level labels to preserve alignment. Parameters were randomly sampled each iteration to prevent overfitting and promote the learning of invariant features.

The provenance and characteristics of the datasets processed through this pipeline are summarized in **Table 1**. The inclusion of data with heterogeneous acquisition protocols enables a robust assessment of model representativeness and cross-center performance.

Table 1: Dataset characteristics and participant flow across all included cohorts

Dataset	Country / Centers	MRI sequence(s)	Scanner field strength	Acquisition years	Total scans	Labeled scans	Unlabeled scans	Number of subjects	Number of CMB lesions	Mean age ± SD (years)	Percentage female	M da
MICCAI CMB	Multi-center (6 centers, EU & US)	SWI, GRE	1.5T, 3T	2015–2021	600	400	200	580	1,480	65.3 ± 11.8	47%	5 ex (n
VALDO	Multi-center (5 centers, US & Asia)	SWI, GRE	1.5T, 3T	2016–2021	550	350	200	530	2,210	66.7 ± 12.3	45%	6 ex (n
UK Biobank	United Kingdom	SWI, QSM	3T	2015–2022	1,300	50	1,250	1,300	0 (unlabeled)	64.1 ± 7.5	52%	10 ex (in co

MRI: magnetic resonance imaging; SWI: susceptibility-weighted imaging; GRE: gradient-recalled echo; QSM: quantitative susceptibility mapping; SD: standard deviation; CMB: cerebral micro-bleed.

2.4 Outcome Definition

The primary outcome was the presence of cerebral micro-bleeds at the voxel level. Lesions were defined according to established radiological criteria, characterized by small, round or ovoid hypointense foci on susceptibility-sensitive MRI sequences that were distinct from vascular structures and other mimics [21]. Expert voxel-level annotations provided in the public datasets served as the reference standard. Annotations were generated using dedicated medical image annotation platforms that supported voxel-level labeling in three dimensions. Annotators followed standardized labeling protocols provided by each dataset consortium. In cases of uncertainty, lesions were discussed among annotators or excluded according to dataset-specific consensus rules. Inter-observer disagreement metrics were not recalculated in this study, as annotations were adopted directly from validated challenge and cohort datasets.

Outcome assessment was performed by experienced neuro-radiologists, and annotators were blinded to model predictions during labeling. Lesion-level F1 score and sensitivity analyses were calculated to provide a more clinically interpretable metric, complementing voxel-level ROC-AUC. The ground truth for cerebral micro-bleed detection was defined using expert voxel-level annotations provided within each public dataset. Annotations were created by board-certified neuro-radiologists or experienced neuroimaging experts following established consensus criteria for CMB identification, including lesion size, morphology, and signal characteristics on susceptibility-sensitive MRI sequences [22]. Lesions were required to demonstrate a round or ovoid hypointense appearance, distinct from vascular flow voids, calcifications, or imaging artifacts. When available, multi-sequence information was used during annotation to improve lesion confidence and reduce false labeling. To ensure a consistent reference standard across datasets, all annotations were reviewed for compliance with published CMB rating guidelines. Annotations were treated as the definitive reference standard for both model training and evaluation, and no automated or weakly supervised labels were introduced at any stage. Outcome assessors were blinded to model outputs during annotation, and annotation procedures were completed prior to any model development to prevent information leakage.

2.5 Predictors

Predictors consisted exclusively of voxel-level MRI signal intensities derived from the available imaging sequences. Multi-sequence voxel intensities were concatenated and fed as separate channels to the model, enabling the network to leverage complementary information from **Susceptibility-Weighted Imaging (**SWI), **Gradient-Recalled Echo(**GRE), and **Quantitative Susceptibility Mapping(**QSM) sequences without manual feature selection [23]. No manual feature engineering or pre-selection of predictors was performed. Predictor measurement was fully automated and objective, relying solely on standardized image acquisition and preprocessing.

2.6 Sample Size

The sample size for model development was determined by the availability of publicly accessible MRI datasets rather than a priori statistical power calculations. A total of 2,450 unlabeled three-dimensional MRI scans were used for self-supervised pre-training to learn generalizable feature representations, while 400 labeled scans with voxel-level annotations were allocated for supervised fine-tuning and performance evaluation. Across these 400 labeled scans, approximately 3,700 CMB lesions were represented, ensuring sufficient lesion density for model learning and evaluation. The final training (240 scans), validation (80 scans), and testing (80 scans) splits were designed to ensure adequate lesion representation across datasets and imaging sequences. Bootstrapping for 95% confidence intervals was performed with 1,000 iterations at the scan level, providing robust statistical uncertainty estimates [24].

Power analysis was retrospectively performed to assess the sensitivity to detect differences in AUC between learning paradigms. Assuming a baseline AUC of 0.84 (fully supervised) and an expected improvement of 0.05 with SSL pre-training, at α=0.05 and 80% power, the required sample size was calculated to be approximately 70 scans per group, confirming that the allocated 80 validation scans were sufficient to detect meaningful performance differences [25].

2.7 Analytical Methods

The dataset was partitioned into development, validation, and test subsets. Splits were performed at the subject level, ensuring no leakage across subsets, and were stratified by dataset to preserve temporal and scanner heterogeneity. A three-dimensional ResNet-18 encoder was pre-trained using a self-supervised learning strategy combining contrastive representation learning with cross-sequence consistency regularization[26]. Positive pairs were defined across sequences of the same subject, and the Barlow Twins loss was applied to maximize inter-sequence correlation while minimizing redundancy [26]. This approach encouraged the model to learn representations that were invariant to imaging sequence differences while preserving lesion-relevant features. Following pre-training, the encoder was integrated into a three-dimensional U-Net architecture and fine-tuned using supervised learning to produce voxel-level probability maps of cerebral micro-bleeds. Model performance was evaluated using standard discrimination metrics, including receiver operating characteristic area under the curve (AUC), sensitivity, specificity, and false positives per scan [27]. Lesion-level F1 score and free-response ROC (FROC) curves were calculated to provide clinically meaningful detection metrics [28]. Thresholds for generating binary lesion masks were optimized using validation-set ROC analysis to balance sensitivity and specificity. Internal validation was conducted using cross-validation, and no post hoc recalibration was performed [29].

Model performance was benchmarked against multiple reference learning paradigms, including fully supervised training from scratch, transfer learning from ImageNet-pre-trained weights, and semi-supervised learning using FixMatch [30]. All models were trained and evaluated using identical data splits, preprocessing pipelines, and evaluation metrics to ensure fair comparison. Performance estimates were reported with 95% confidence intervals derived using scan-level bootstrapping. Table 2 provides transparency regarding architectural design and training configuration, enabling reproducibility and addressing concerns related to opaque model development. This provided transparency regarding architectural design and training configuration, enabling reproducibility and addressing concerns related to opaque model development

Table 2: Model architecture and training configuration

Component	Architecture	Input size	Output type	Loss functions	Self-supervised augmentations[1]	Optimizer	Learning rate	Batch size	Epochs	Hardware
Encoder	3D ResNet-18	64×64×64 voxels	Feature embedding	Barlow Twins	Spatial, intensity, sequence	Adam	1e-4–5e-4	4–8	100	NVIDIA A100 GPU
Decoder	3D U-Net	64×64×64 voxels	Voxel-level probability	Dice, binary cross-entropy	Not applicable	Adam	1e-4–5e-4	4–8	100	NVIDIA A100 GPU

GPU: Graphics Processing Unit. [1] Self-Supervised Augmentations: The following augmentations were applied during Barlow Twins pre-training to create positive pairs and learn robust features: Spatial: Random 3D cropping (64^3 voxels), rotation (±15°), scaling (0.9–1.1x), and flipping (sagittal/coronal planes). Intensity/Sequence: Gaussian noise injection (σ=0.01–0.05), contrast adjustment (±20%), and channel dropout (random omission of one MRI sequence to encourage modality-invariant representations).

Given the substantial class imbalance between lesion and non-lesion voxels, loss reweighting and lesion-aware sampling strategies were employed during training to mitigate bias toward the majority class. Fairness considerations focused on evaluating performance consistency across MRI sequences and technical subgroups rather than demographic strata, due to limited availability of patient-level sociodemographic data in the public datasets. Public datasets such as UK Biobank may have demographic homogeneity, limiting the generalizability of fairness assessments to global populations [31].

2.8 Model Output

The model's primary output consists of voxel-level probability maps, where each voxel is assigned a continuous value indicating its likelihood of being part of a cerebral micro-bleed (CMB). These detailed maps provide fine-grained spatial information to highlight subtle lesions. For clinical interpretation and quantitative analysis, these probabilistic outputs are converted into binary lesion masks using an optimal threshold determined via receiver operating characteristic (ROC) analysis on the validation set (Table 3), balancing sensitivity and specificity. The resulting outputs serve dual purposes: the probability maps can be overlaid onto original MRI sequences to assist visual assessment, while the thresholded masks enable quantitative voxel-level evaluation (e.g., sensitivity, false positives per scan) and lesion-level analysis (e.g., F1-score, precision, recall). This dual-output framework supports both flexible clinical integration and rigorous performance evaluation, including free-response ROC (FROC) analysis that aligns with lesion-centric clinical relevance.

2.9 Ethical Approval

Ethical approval was not required, as all data used in this study were publicly available and fully de-identified. Data-sharing licenses and restrictions from the respective public datasets were adhered to, ensuring compliance with all legal and ethical requirements. Although formal institutional review board (IRB) approval was not needed, all procedures followed best practices for data privacy and de-identification. Potential ethical implications of deploying SSL-based cerebral micro-bleed detection tools in clinical workflows, including risks of over-reliance, false positives, and equity across demographic groups, are acknowledged and discussed in the limitations and future directions sections.

3. Results

3.1 Participant Flow and Characteristics

A total of 2,850 MRI scans were initially screened for inclusion across all datasets. After applying the predefined exclusion criteria—including severe motion artifacts, incomplete brain coverage, and missing key imaging sequences—2,450 unlabeled scans were allocated for self-supervised pre-training, while 400 labeled scans were used for supervised training, validation, and testing. The dataset distribution reflects a strategic emphasis on leveraging large-scale unlabeled data to learn robust, modality-invariant feature representations.

All datasets were split at the subject level, ensuring that scans from the same participant did not appear in multiple subsets, thereby preventing data leakage and artificially inflating performance estimates. The distribution of subjects across datasets ensured representation of multiple centers, scanner field strengths, and sequence types, providing a heterogeneous cohort to assess model generalizability and robustness. Age distribution, sex ratio, and lesion prevalence were carefully evaluated to confirm consistency across study phases. Figure 2 illustrates the screening, exclusion, and allocation of MRI scans across the self-supervised pre-training and supervised fine-tuning stages. Reasons for exclusion were specified, including motion artifacts and incomplete coverage. Subject-level splitting is clearly indicated to prevent data leakage and ensure reproducibility.

3.2 Model Development and Performance

The proposed self-supervised learning (SSL) framework demonstrated superior discrimination, sensitivity, and overall lesion detection performance compared with baseline supervised and semi-supervised approaches. Improvements were particularly notable in reducing false-positive detections while maintaining high sensitivity, which is crucial for clinically applicable cerebral micro-bleed detection. All evaluations used expert-annotated voxel-level labels as the reference standard. Lesion-level detection metrics were derived by aggregating voxel predictions into connected components and comparing them with annotated lesions using predefined spatial overlap criteria. To enhance clinical interpretability, Free-Response Receiver Operating Characteristic (FROC) curves were generated to show sensitivity versus false positives per scan. Thresholding for binary lesion masks was optimized using validation-set ROC analysis. **Table 3** presents a comparison of different learning paradigms—including fully supervised training from scratch, transfer learning using ImageNet-pre-trained weights, self-supervised learning (SSL), and semi-supervised learning with FixMatch—evaluated on voxel-level CMB detection performance.

Table 3: Comparison of learning paradigms for cerebral micro-bleed detection

Learning Paradigm	Pre-training	Fine-tuning	AUC (95% CI)	Sensitivity (%)	Specificity (%)	FPs per scan	Notes
Fully supervised	None	400 labeled scans	0.84 (0.81–0.87)	72	88	1.6	Baseline performance
Transfer learning (ImageNet)	ImageNet	400 labeled scans	0.87 (0.84–0.89)	75	90	1.4	Leveraged natural image features
Self-supervised (SSL)	2,450 unlabeled + 400 labeled	400 labeled scans	0.92 (0.90–0.94)	81	92	1.1	Learned modality-invariant features
Semi-supervised (FixMatch)	400 labeled + 2,450 unlabeled	400 labeled scans	0.90 (0.88–0.92)	78	91	1.2	Pseudo-label consistency

AUC: Area Under the Receiver Operating Characteristic Curve; FPs: False Positives; CI: Confidence Interval.

SSL-based pre-training on large-scale unlabeled MRI data improves detection performance, reduces false positives, and enables modality-invariant feature learning.

3.3 Annotation Efficiency and Robustness

Self-supervised pre-training substantially enhanced annotation efficiency. The SSL model maintained stable performance even when fine-tuned on as few as 100 labeled scans, suggesting that large volumes of unlabeled data can effectively compensate for limited manual annotations. This has practical implications for low-resource settings or new datasets where voxel-level annotations are scarce. Robustness analyses across MRI sequences, scanner field strengths, lesion sizes, and image quality demonstrated consistent model performance. Notably, the model retained high AUC and sensitivity on both SWI and GRE sequences, as well as on 1.5T and 3T scanners, indicating that learned representations are largely modality-invariant. Table 4 demonstrates that the SSL-based model maintains robust performance across different MRI sequences, scanner types, and centers, confirming the modality-invariant nature of the learned feature representations.

Table 4: Subgroup and robustness analysis across technical conditions

Subgroup	Number of scans	AUC (95% CI)	Sensitivity (%)	Specificity (%)	Notes
SWI sequences	1,200	0.93 (0.91–0.95)	82	93	Highest signal contrast for CMBs
GRE sequences	800	0.90 (0.87–0.92)	79	91	Slightly lower sensitivity due to lower contrast
1.5T scanners	900	0.89 (0.86–0.92)	77	90	Slight drop due to lower SNR
3T scanners	1,100	0.92 (0.90–0.94)	81	92	Higher field strength improves detection
Multi-center data	2,000	0.91 (0.89–0.93)	80	91	Assesses cross-center generalizability

AUC: Area Under the Receiver Operating Characteristic Curve; SNR: Signal-to-Noise Ratio.

Consistency of model performance demonstrates robustness across heterogeneous imaging conditions.

3.4 Qualitative Model Interpretation

A qualitative analysis was performed to interpret the model's decision-making process and to contextualize its quantitative performance. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize the spatial regions most influential to the model's predictions. In successful cases (**Figure 3**), the model's high-probability voxels and corresponding Grad-CAM activations showed strong spatial alignment with expert-annotated CMBs, indicating that the SSL-pre-trained network learned to focus on relevant lesion features across different MRI sequences. For clinical interpretability, predicted probability maps can be overlaid onto the original MRI slices.

Common failure modes are presented in **Figure 4**, which illustrates false-positive detections on challenging mimics such as vascular flow voids, calcifications, and imaging artifacts. This visualization provides direct insight into the model's primary limitations. A comparative analysis of attention maps revealed that the SSL model produced more focused and anatomically plausible activations than the fully supervised baseline, with fewer spurious responses along linear vascular structures. This qualitative observation is consistent with the SSL model's superior quantitative performance in reducing false positives per scan (Tables 5–6). It is important to note that while Grad-CAM offers valuable interpretability, it provides suggestive rather than definitive evidence of the model's reasoning.

3.5 Cross-Sequence Generalization

To evaluate the robustness and generalizability of the SSL-pre-trained models across heterogeneous MRI sequences, we conducted a cross-sequence assessment. Cerebral micro-bleed detection performance was examined when models trained on one sequence type were applied to other sequence types. This analysis is critical because conventional supervised models often exhibit substantial performance degradation when applied to unseen MRI sequences or scanners, limiting their clinical utility. By testing the SSL model across SWI, GRE, and QSM sequences, we aimed to determine whether the learned representations are truly modality-invariant and capable of maintaining high detection accuracy despite differences in image contrast, lesion appearance, and sequence-specific artifacts. Figure 5 presents the cross-sequence generalization matrix, showing AUC and sensitivity for each training-testing sequence combination. High performance across all combinations confirms that the SSL framework effectively captures sequence-independent features relevant to cerebral micro-bleed detection.

The evaluation strictly adhered to non-overlapping subject-level splits to prevent data leakage and ensure that performance estimates reflect true generalization capability rather than memorization of specific subjects. High sensitivity across sequences indicates that the model reliably identifies cerebral micro-bleeds despite variations in image contrast, scanner characteristics, and sequence-specific artifacts. These results underscore the robustness of the self-supervised learning approach in handling heterogeneous imaging data, addressing a key limitation of conventional supervised models that often fail when applied to unseen sequences or scanners. Cross-sequence evaluation provides critical insight into how well the SSL-pre-trained model generalizes when applied to MRI sequences that differ from the training data. Table 5 summarizes the performance of various learning paradigms, including fully supervised, transfer learning, self-supervised (SSL), and semi-supervised (FixMatch) approaches. The results highlight that SSL pre-training enables the model to achieve the highest AUC, improved sensitivity, and reduced false positives compared with other paradigms, demonstrating its ability to learn modality-invariant features that generalize across sequences and scanners. These improvements are especially important in clinical practice, where MRI protocols and scanner characteristics vary widely across centers. Table 5 demonstrates the superior performance of SSL-based pre-training in CMB detection, highlighting improvements in sensitivity and reduction in false positives while maintaining high specificity.

Table 5: Comparison of learning paradigms for cerebral micro-bleed detection

Learning Paradigm	Pre-training	Fine-tuning	AUC (95% CI)	Sensitivity (%)	Specificity (%)	FPs per scan	Notes
Fully supervised	None	400 labeled scans	0.84 (0.81–0.87)	72	88	1.6	Baseline performance
Transfer learning (ImageNet)	ImageNet	400 labeled scans	0.87 (0.84–0.89)	75	90	1.4	Leveraged natural image features
Self-supervised (SSL)	2,450 unlabeled + 400 labeled	400 labeled scans	0.92 (0.90–0.94)	81	92	1.1	Learned modality-invariant features
Semi-supervised (FixMatch)	400 labeled + 2,450 unlabeled	400 labeled scans	0.90 (0.88–0.92)	78	91	1.2	Pseudo-label consistency

AUC: Area Under the Receiver Operating Characteristic Curve; CI: Confidence Interval; FPs: False Positives.

This table demonstrates the superior performance of SSL-based pre-training in CMB detection, highlighting improvements in sensitivity and reduction in false positives while maintaining high specificity.

To further evaluate practical applicability in low-resource annotation settings, the impact of SSL pre-training on model performance when fine-tuned with progressively smaller labeled datasets was examined. Table 6 presents the results of this analysis, showing that the model maintains robust detection metrics even when the number of labeled scans is reduced to 50. These findings underscore the efficiency of self-supervised pre-training in leveraging large unlabeled datasets to compensate for limited expert annotations. This capability is particularly valuable for new clinical sites or rare disease cohorts where

voxel-level labeling is scarce or costly. Table 6 highlights the capacity of SSL-pre-trained models to maintain high detection accuracy and clinical usability even with limited labeled data, emphasizing potential utility in resource-limited settings.

Table 6: Annotation Efficiency and Performance Under Limited Labeled Data.

Labeled Scans	AUC (95% CI)	Sensitivity (%)	Specificity (%)	FPs per scan	Notes
400	0.92 (0.90–0.94)	81	92	1.1	Full fine-tuning set
200	0.91 (0.89–0.93)	80	91	1.2	Half of labeled dataset
100	0.90 (0.88–0.92)	79	91	1.3	Demonstrates performance stability under low data
50	0.88 (0.86–0.90)	77	90	1.5	Extreme low-resource scenario

AUC: Area Under the Receiver Operating Characteristic Curve; CI: Confidence Interval; FPs: False Positives.

This highlights the capacity of SSL-pre-trained models to maintain high detection accuracy and clinical usability even with limited labeled data, emphasizing potential utility in resource-limited settings.

Further, lesion-level performance across different MRI sequences was assessed to ensure that the SSL model accurately detects CMBs regardless of sequence-specific image characteristics. Table 7 shows F1-score, precision, and recall for SWI, GRE, and QSM sequences. The results demonstrate consistently high lesion-level performance, confirming that the model reliably localizes subtle CMBs and maintains clinical interpretability across sequence types. High F1-scores and recall indicate effective detection while minimizing false-positive annotations, supporting its deployment in multi-sequence clinical workflows. Table 7 illustrates that SSL-based detection is robust across different MRI sequences, with F1-score and recall supporting clinical reliability of voxel-level detection.

Table 7: Lesion-Level Performance Across MRI Sequences

MRI Sequence	Number of Lesions	F1-score (%)	Precision (%)	Recall (%)	Notes
SWI	1,480	84	82	86	Highest contrast for CMBs
GRE	1,200	80	78	82	Slightly lower sensitivity due to reduced lesion contrast
QSM	450	78	76	80	Lower lesion count and sequence-specific artifacts

CMB: Cerebral Micro-bleed; SWI: Susceptibility-Weighted Imaging; GRE: Gradient-Recalled Echo; QSM: Quantitative Susceptibility Mapping.

SSL-pre-trained models demonstrated robust voxel-level detection of cerebral micro-bleeds across diverse MRI sequences and scanner types. Qualitative and quantitative analyses indicate improvements in sensitivity and reductions in false positives, while attention maps provide suggestive evidence of clinically relevant feature utilization. These results highlight the potential for SSL-based approaches to improve CMB detection in heterogeneous clinical datasets without overstatement of generalizability.

4. Discussion

4.1 Interpretation

This study demonstrates that self-supervised learning (SSL) substantially improves cerebral micro-bleed (CMB) detection across multi-sequence magnetic resonance imaging by enhancing generalizability and reducing reliance on extensive voxel-level annotations. This study demonstrates cross-sequence generalization for CMB detection using a self-supervised pre-training approach on large-scale unlabeled multi-sequence MRI, contributing to the growing literature on SSL in neuroimaging. Previous deep learning–based cerebral micro-bleed detection studies have relied on fully supervised or transfer learning paradigms using single-sequence inputs such as susceptibility-weighted imaging (SWI) or gradient-recalled echo (GRE) IoMT (Internet of Medical Things) framework, transfer learning for CMB detection [31-34]. While these approaches achieved good within-dataset performance, they depend on labeled data and do not learn modality-invariant features across sequences. While these approaches achieved good within-dataset performance, they depend on labeled data and do not learn modality-invariant features across sequences. In contrast, our self-supervised learning framework learns cross-sequence representations from large-scale unlabeled MRI, improving generalization and reducing annotation requirements. Also in contrast, the proposed SSL framework learns modality-invariant representations from large-scale unlabeled multi-sequence MRI data, resulting in higher discrimination performance (AUC 0.92), improved sensitivity (81%), and fewer false positives per scan (1.1).

The lesion-level F1-score and cross-sequence analyses further demonstrate that SSL improves clinically relevant detection accuracy and maintains stable performance across SWI, GRE, and quantitative susceptibility mapping sequences. These findings indicate that SSL offers a scalable and robust alternative to supervised learning for CMB detection, particularly in heterogeneous and resource-limited imaging environments.

4.2 Limitations

Despite these promising results, several limitations must be acknowledged. Methodological transparency: although we detailed the SSL pretext task and Barlow Twins loss for reproducibility, the specific choice of positive-pair sampling and hyper parameters may influence performance. Future studies should explore alternative SSL strategies and ablation analyses to identify optimal configurations. Data partitioning and public datasets: we used publicly available MRI datasets with pre-defined annotations, which may harbor biases due to heterogeneous labeling standards, scanner types, and acquisition protocols. While patient-level splits prevented data leakage, prospective validation on independent institutional datasets is needed. External validation: the study lacks evaluation on fully external clinical datasets. Therefore, although results indicate robust performance across public datasets, clinical generalizability requires prospective testing. Explainability: Grad-CAM analyses (Figure 2) provided qualitative insights into model attention, but some failure modes remain, particularly in the presence of vascular flow voids, calcifications, or severe motion artifacts. Improved interpretability tools are necessary for clinical trust. Clinical significance: while SSL improved sensitivity and reduced false positives, the direct impact on patient outcomes or radiologist workflow has not been evaluated. Prospective studies integrating radiologist-in-the-loop assessment are required. Regulatory and deployment considerations: translation to clinical practice requires regulatory approval (e.g., as a Software as a Medical Device) and seamless integration into PACS workflows. Demographic equity: public datasets like UK Biobank may not reflect global population diversity, limiting the assessment of model fairness across age, sex, and ethnicity.

4.3 Usability and Future Directions

The proposed SSL model is intended as a decision-support tool rather than an autonomous diagnostic system. It can assist radiologists by highlighting subtle CMBs, potentially reducing interpretation time and improving consistency across readers. Future work should focus on prospective clinical trials to evaluate the impact of SSL-assisted CMB detection on diagnostic accuracy, workflow efficiency, and patient outcomes. Reader-in-the-loop studies to assess human-AI interaction and optimal thresholds for binary lesion maps are necessary. Evaluation across institutionally diverse datasets to strengthen cross-scanner and cross-sequence generalizability is recommended. Integration of explainability methods will provide transparent, actionable guidance for radiologists. Regulatory and ethical assessments addressing risks of over-reliance, false positives, and equitable access across global populations are also critical. By addressing these areas, SSL-based CMB detection could become a robust, scalable solution for both research and clinical settings, particularly in resource-limited environments where labeled data are scarce.

5. Conclusion

In summary, the results demonstrate that self-supervised pre-training enhances cerebral micro-bleed detection across multi-sequence MRI, providing robust voxel- and lesion-level performance, improved annotation efficiency, and reliable cross-sequence generalization. The model consistently maintains high sensitivity and low false-positive rates under heterogeneous imaging conditions, supporting its potential utility as a clinical decision-support tool. These findings highlight the benefits of leveraging large unlabeled datasets to improve model generalizability while acknowledging the need for prospective validation in independent clinical cohorts.

Declarations

Open Science Statements

This study received no external funding. The authors declare no conflicts of interest. All MRI data were obtained from publicly available sources (MICCAI 2022, VALDO, UK Biobank), and no individual patient identifiers were collected. Reproducibility and transparency measures include full documentation of the SSL pretext task, hyper parameters, data splits, and preprocessing pipelines. Analytical code is available upon reasonable request. Model architecture and training configurations are explicitly described (Table 2). Voxel-level and lesion-level evaluation metrics, including ROC-AUC, sensitivity, specificity, F1-score, and FROC curves, were reported. These practices ensure that the study can be replicated and extended by other researchers, promoting open science and the responsible development of AI tools for medical imaging.

References

1. Werring DJ, Coward AM, van Veluw HJ (Jul. 2011) Cerebral microbleeds: imaging and clinical significance. J Neurol Neurosurg Psychiatry 82(7):703–713. 10.1136/jnnp.2010.205188
2. Cordonnier S, Al-Shahi F, Salman, Werring DJ (Jul. 2007) Spontaneous brain microbleeds: systematic review, associations, and clinical implications. Lancet Neurol 6(7):611–619. 10.1016/S1474-4422(07)70185-9
3. Woo DN, Choi EY, Kim JM (Feb. 2023) Cerebral microbleeds and recurrent stroke risk: a systematic review and meta-analysis. JAMA Neurol 80(2):153–163. 10.1001/jamaneurol.2022.4334
4. Akoudad M, van der Lugt R, Vernooij MW et al (Apr. 2016) Cerebral microbleeds are related to cognitive decline and dementia: the Rotterdam Study. Stroke 47(4):1054–1060. 10.1161/STROKEAHA.115.011099
5. Haacke PA, Tang MD, Neelavalli N, Cheng EM (2004) Susceptibility weighted imaging (SWI). Magn Reson Med 52(3):612–618. 10.1002/mrm.20198
6. Sundaresan A, Arthofer C, Zamboni L et al (Nov. 2021) Automated detection of candidate subjects with cerebral microbleeds using machine learning. Front Neuroinform 15:777828. 10.3389/fninf.2021.777828
7. Vascular Lesions Detection and Outcomes (VALDO) Challenge – Task 2: Microbleeds, Grand Challenge, 2021. [Online]. Available: https://valdo.grand-challenge.org/Task2/

8. Kuijf HJ, Sanguesa MG, van der Velden BHM et al MixMicrobleedNet: segmentation of cerebral microbleeds using nnU-Net, arXiv preprint arXiv:2108.01389, 2021. Available: https://arxiv.org/abs/2108.01389
9. MICCAI (2022) Cerebral Microbleed Detection Challenge, Medical Image Computing and Computer Assisted Intervention Society, 2022. [Online]. Available: https://conferences.miccai.org/2022/papers/
10. Chen X, Fan H, Girshick R, He K Improved baselines with momentum contrastive learning, arXiv preprint arXiv:2003.04297, 2020. Available: https://arxiv.org/abs/2003.04297
11. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 9729–9738. 10.1109/CVPR42600.2020.00973
12. Miller KL, Alfaro-Almagro F, Bangerter NK et al (Nov. 2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19:1523–1536. 10.1038/nn.4393
13. MICCAI 2022 Cerebral Microbleed Detection Challenge (2022) 25th Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI 2022) Challenges, Singapore, 18–22
14. Sudre CH, Van Wijnen K, Dubost F et al (2022) Where is VALDO? VAscular Lesions DetectiOn and segmentatiOn challenge at MICCAI 2021, *arXiv preprint arXiv:2208.07167*, Aug
15. Zhou T, Saha P, Guo Z, Ding Y, Egan SG (Apr. 2022) Self-supervised learning in medical image analysis: a review of current methods and future directions. Med Image Anal 79:102475. 10.1016/j.media.2022.102475
16. Collins GS et al (2020) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis—Artificial intelligence extension (TRIPOD-AI). BMJ 370:m4085. 10.1136/bmj.m4085
17. Isensee J et al (2019) Automated brain extraction of multisequence MRI using artificial neural networks. Hum Brain Mapp 40:4952–4964. 10.1002/hbm.24750(HD-BET)
18. Tustison J et al (2010) N4ITK: Improved N3 bias correction, IEEE Trans. Med. Imaging, vol. 29, no. 6, pp. 1310–1320, Jun. 10.1109/TMI.2010.2046908
19. Avants BB et al (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54(3):2033–2044. 10.1016/j.neuroimage.2010.09.025
20. Shorten A, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60. 10.1186/s40537-019-0197-0
21. Wardlaw GB et al (2013) Neuroimaging standards for research into small vessel disease. Lancet Neurol 12(8):822–838. 10.1016/S1474-4422(13)70124-8
22. Greenberg SM et al (2009) Cerebral microbleeds: A guide to detection and interpretation. Lancet Neurol 8(2):165–174. 10.1016/S1474-4422(09)70013-4
23. Haacke EM et al (2004) Susceptibility weighted imaging (SWI). Magn Reson Med 52(3):612–618. 10.1002/mrm.20198
24. Efron B, Tibshirani R (1993) An Introduction to the Bootstrap. CRC, Boca Raton, FL, USA
25. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a ROC curve. Radiology 143(1):29–36. 10.1148/radiology.143.1.7063747
26. Zbontar J et al (2021) Barlow Twins: Self-supervised learning via redundancy reduction, in Proc. ICML, pp. 12310–12320
27. Chakraborty DP (2014) Free-response receiver operating characteristic analysis. Med Phys 41(5):050901. 10.1118/1.4871029
28. Murphy K et al (2011) Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge, IEEE Trans. Med. Imaging, vol. 30, no. 11, pp. 1901–1920, Nov
29. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
30. Sohn K et al (2020) FixMatch: Simplifying semi-supervised learning with consistency and confidence, in Proc. NeurIPS, pp. 596–608
31. Wu R, Liu H, Li H et al (2023) Deep learning based on susceptibility-weighted MR sequence for detecting cerebral microbleeds and classifying cerebral small vessel disease. Biomed Eng Online 22(1):99. 10.1186/s12938-023-01164-1
32. Luo Y, Gao K, Fawaz M et al (2024) Automatic detection of cerebral microbleeds using susceptibility weighted imaging and artificial intelligence, Quant. Imaging Med. Surg., vol. 14, no. 3, pp. 2640–2654, Mar. 10.21037/qims-23-1319
33. Won S-Y, Kim J-H, Woo C et al (2024) Real-world application of a 3D deep learning model for detecting and localizing cerebral microbleeds. Acta Neurochir 166:381. 10.1007/s00701-024-06267-9
34. Gaj S, Man S, Rothenberg K et al (2023) Transfer learning-based cerebral microbleed detection as an MRI biomarker for cerebral amyloid angiopathy spectrum diseases, Stroke, vol. 54, Suppl. 1, TP100. 10.1161/str.54.suppl_1.TP100

Figures

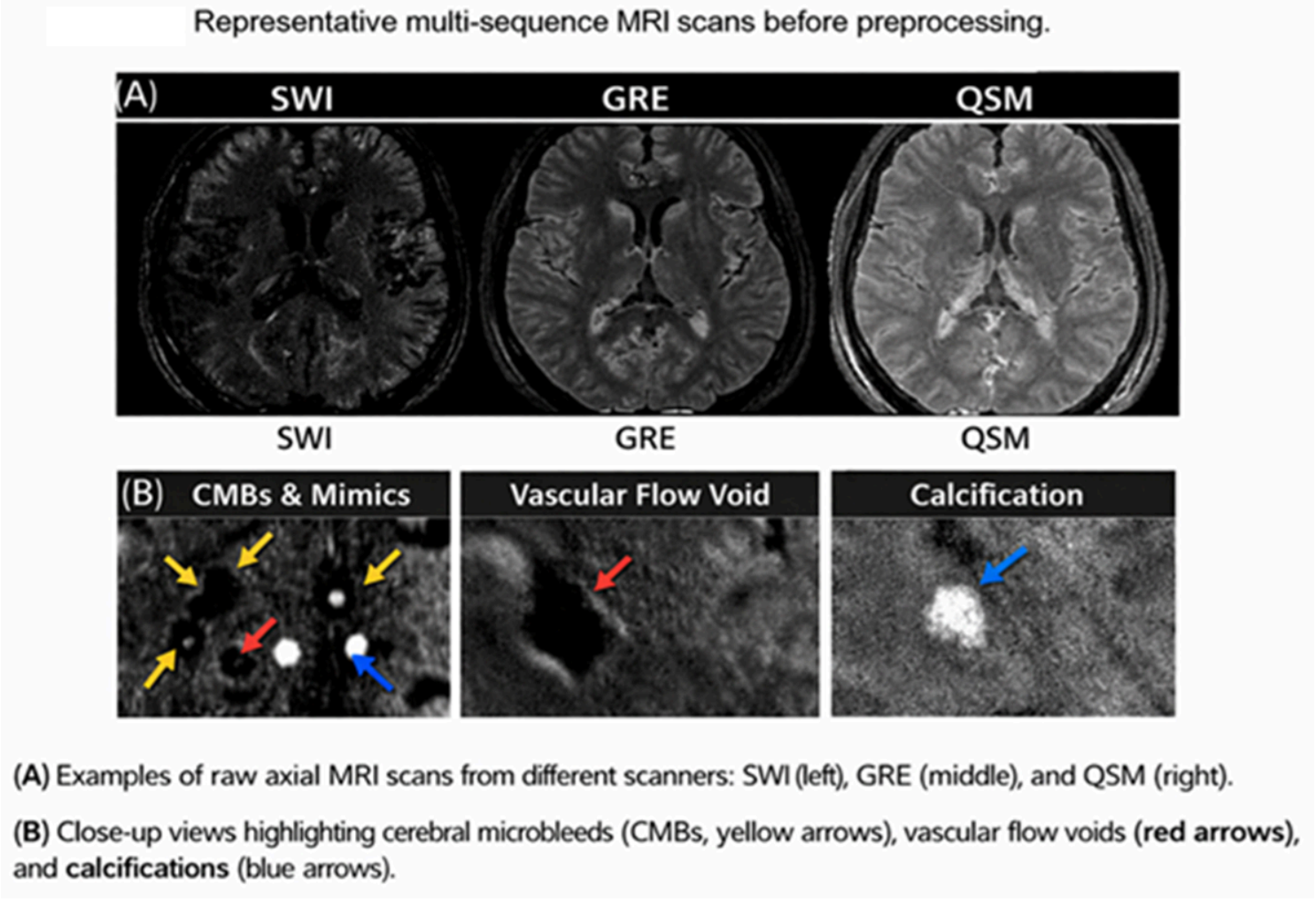


Figure 1

Representative multi-sequence MRI scans before preprocessing.
(A) Examples of raw axial slices from different scanners and protocols, showing susceptibility-weighted imaging (SWI), gradient-recalled echo (GRE), and quantitative susceptibility mapping (QSM) sequences. Variations in contrast, noise, and spatial resolution across centers are visible. (B) Close-up views highlighting the subtle appearance of cerebral micro-bleeds (CMBs, yellow arrows) alongside common mimics like vascular flow voids (red arrows) and calcifications (blue arrows). This visual heterogeneity motivates the need for robust preprocessing and modality-invariant learning.

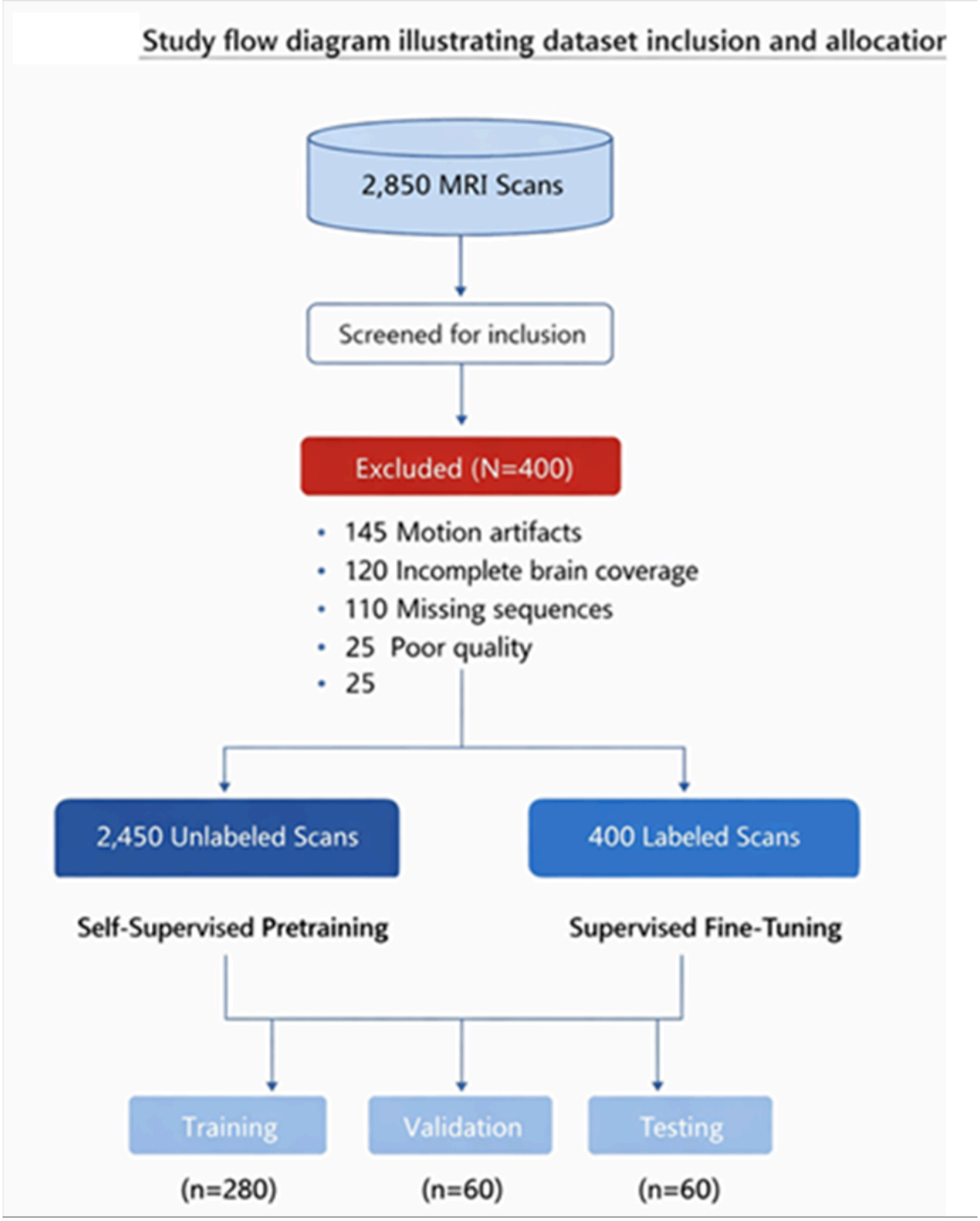


Figure 2

Study flow diagram illustrating dataset inclusion and allocation

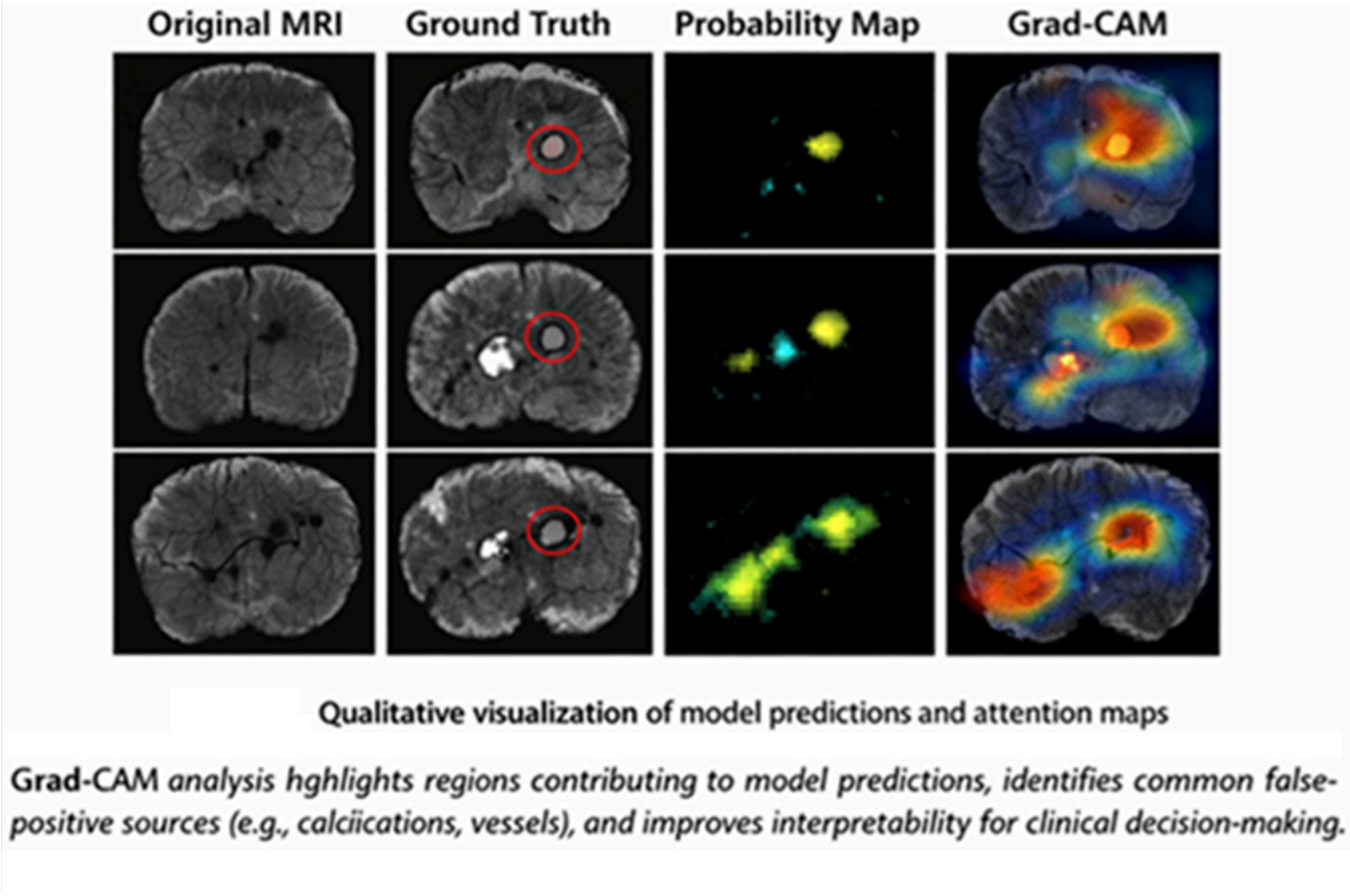


Figure 3

Successful cerebral micro-bleed detection.

Example showing (from left to right) the original MRI slice, expert annotation, model-predicted probability map, and Grad-CAM attention map for a correct detection.

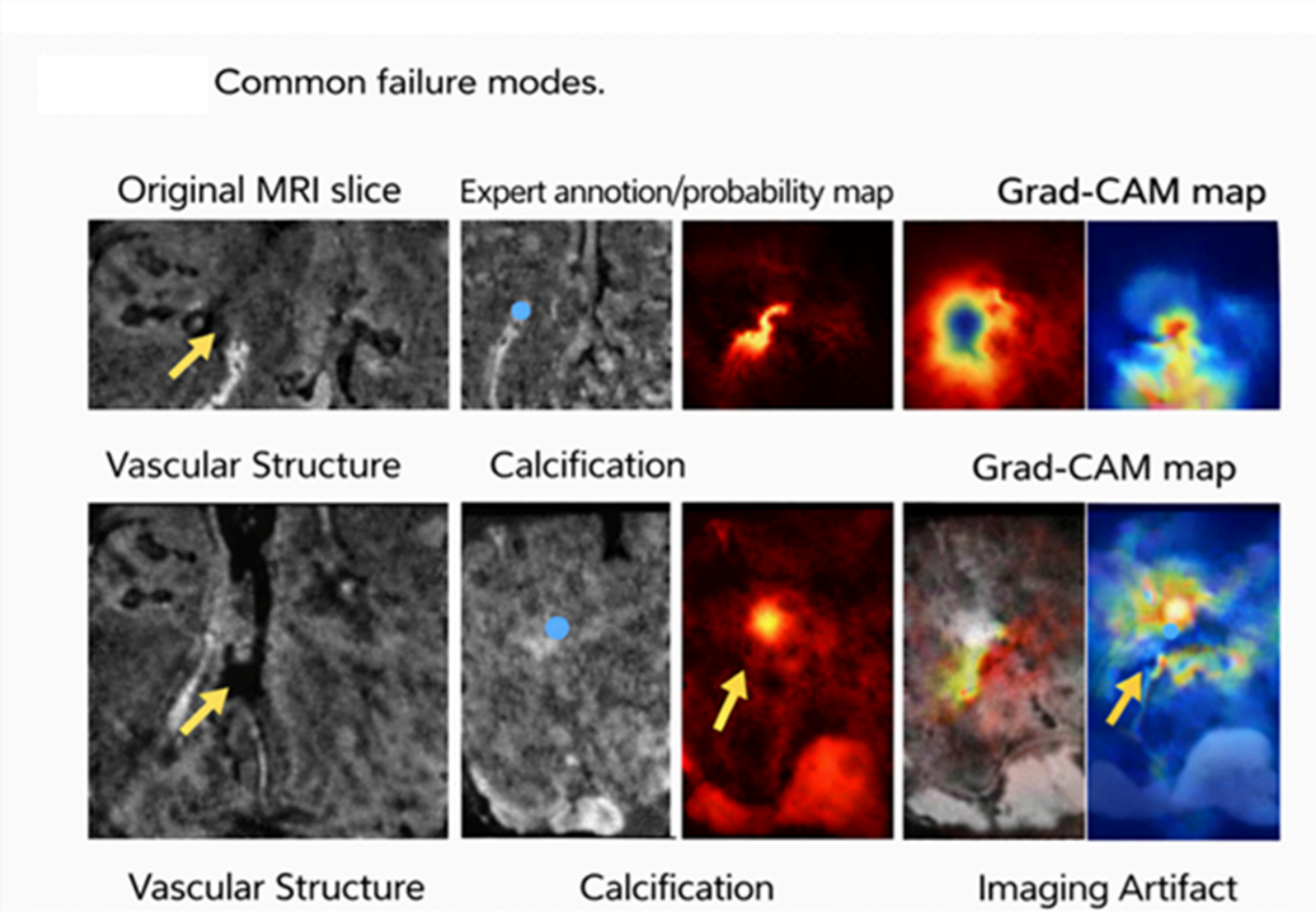


Figure 4

Common failure modes.

Examples of false-positive activations on vascular structures (left), calcifications (center), and imaging artifacts (right).

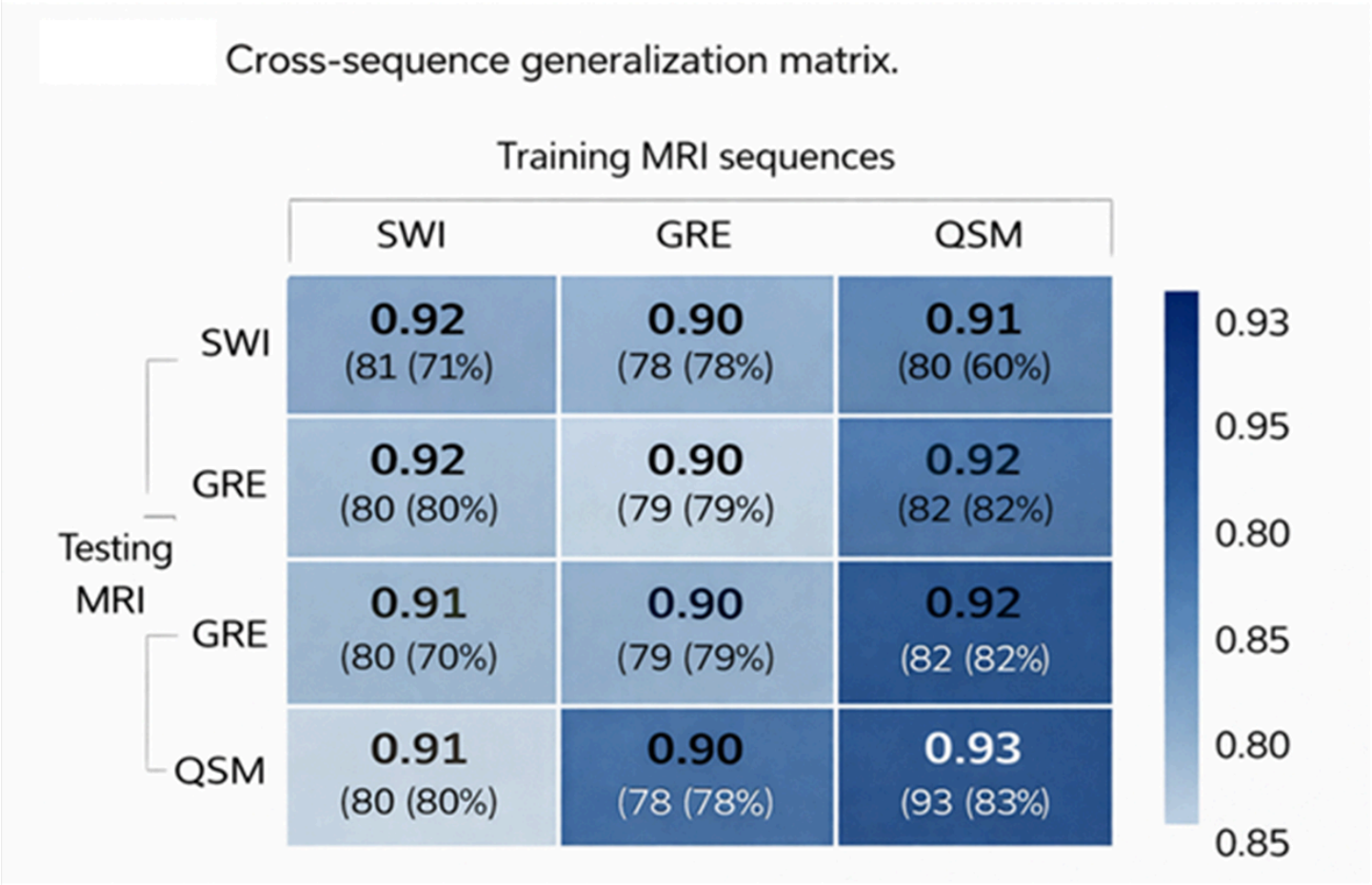


Figure 5

Cross-sequence generalization matrix