

Journal of Smart Algorithms and Applications
JSAA
ISSN: 3070-4189/© 2026 , The Authors. (CC BY 4.0) License.
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

Patient-Level Intracranial Aneurysm Detection in Volumetric Neuroimaging Using a 2.5D Deep Learning Framework

Santhosh Pandi C [a,1], Prasun Chakrabarti [b], and Shashi Kant Gupta[c]

[a] *Department of Electronics and Communication Engineering, SRM TRP ENGINEERING COLLEGE, India. Email: santhosh.1416.05@gmail.com*
[b] *Sir Padampat Singhania University, Udaipur, Rajasthan, India. Email: drprasun.cse@gmail.com*
[c] *Post-Doctoral Fellow and Researcher, Computer Science and Engineering, Eudoxia Research University, USA. Email: raj2008enator@gmail.com*

ABSTRACT

Intracranial aneurysms (IAs) have a case-fatality rate greater than 50% following rupture, and the majority are clinically silent until hemorrhage occurs. An important research direction in neurovascular imaging informatics is the development of automated screening tools that can identify potentially high-risk cases for expedited expert review. In this work, we propose a binary deep learning framework for patient-level aneurysm detection from volumetric DICOM neuroimaging. The proposed architecture adapts EfficientNetV2-S to a 2.5D input representation of channel-stacked 32 axial slices resized to 384 × 384 pixels. Percentile-based intensity normalization is employed to reduce cross-scanner variability. We measure model generalization using stratified five-fold cross-validation and improve prediction stability using weighted ensemble averaging, test-time augmentation, and temperature-scaled probability calibration. On the RSNA 2025 Intracranial Aneurysm Detection Challenge dataset of 1,147 clinical cases, the framework achieved a weighted AUC-ROC of 0.694 with a cross-validation mean of 0.681±0.005. These results indicate that the proposed method can be considered as a first research baseline for automated aneurysm pre-screening, but the performance obtained is still not enough to be used in the clinical routine without assistance. Before its integration into clinical radiology workflows, it needs external validation on independent multi-site cohorts, threshold-specific sensitivity and specificity analysis, and lesion-level evaluation when annotations are available.

PAPER INFORMATION

HISTORY

Received: 30 March 2026
Revised: 25 May 2026
Accepted: 22 June 2026
Online: 26 June 2026

MSC:

68T07; 68R10; 94A60; 68M15

KEYWORDS

Intracranial Aneurysms; Deep Learning; EfficientNetV2; Ensemble Learning; RSNA 2025.

[1]*Corresponding author: Department of Electronics and Communication Engineering, SRM TRP ENGINEERING COLLEGE, India. Email: santhosh.1416.05@gmail.com*

1. INTRODUCTION

Intracranial aneurysms (IAs) represent focal, saccular, or fusiform dilatations of the walls of cerebral arteries, caused by hemodynamic stress, mural weakening, and chronic inflammatory remodeling. The global prevalence in the adult population is estimated at 3–5%, but the vast majority remain clinically silent until rupture [1]. Aneurysmal subarachnoid hemorrhage (SAH) carries a case-fatality rate exceeding 50%, leaving many survivors with permanent neurological disabilities [2]. Early diagnosis by CT angiography (CTA) or magnetic resonance angiography (MRA) enables preventive treatment by surgical clipping or endovascular coiling, which greatly improves the prognosis [3]. Despite current advances in neuroimaging technologies, volumetric interpretation of 3D brain scans by hand remains labor-intensive, subject to inter-reader variability, and increasingly bottlenecked by global neuroradiology shortages. Automated binary pre-screening, flagging cases as aneurysm-positive or aneurysm-negative for expedited expert review, represents a practical research direction that could complement, rather than replace, radiologist judgment. Such a system does not need to anatomically localize or segment lesions to provide clinical value; a reliable triage signal is sufficient to reduce the cognitive load of radiologists processing high-volume scan queues [4].

Deep convolutional neural networks (CNNs) have achieved outstanding results in medical image analysis tasks. In the aneurysm detection literature, architecture has evolved from handcrafted feature pipelines to 2D slice classifiers and fully 3D volumetric encoder-decoders, each with distinct trade-offs between computational cost, annotation requirements, and generalization fidelity [5]. Recently, Transformer-based architecture has demonstrated promising global-context modeling for volumetric data, but they still demand large data [6]. This study contributes a principled 2.5D binary classification framework that explicitly emphasizes computational tractability and cross-scanner robustness towards peak accuracy, a constraint that is particularly pertinent to deployment in resource-limited environments. The main contributions of this work are the following:

1. A binary reformulation of the aneurysm detection task to a single patient-level positive/negative label, removing the multi-label anatomical localization complexity while preserving triage utility.
2. A volumetric pre-processing pipeline that normalizes heterogeneous DICOM stacks into a uniform 32-slice, 384 x 384 pixel 2.5D representation with percentile-based intensity normalization.
3. Adaptation of EfficientNetV2 to 32-channel 2.5D input with a principled weight initialization strategy that preserves ImageNet-pretrained low-level feature detectors.
4. A multi-fold weighted ensemble strategy with test-time augmentation and temperature-scaled probability calibration to improve discrimination and calibration quality.

The remainder of the paper is organized as follows. Section 2 describes the related work on classical computer-aided detection, deep learning-based aneurysm detection, and ensemble calibration strategies. Section 3 presents the dataset and methodology, which include the volumetric preprocessing pipeline, 2.5D input representation, binary classification architecture based on EfficientNetV2-S, training setup, and ensemble calibration methods. Section 4 presents experimental results with cross-validation performance, comparisons, and distribution analysis of the prediction. The main findings and their clinical and methodological implications are elaborated in Section 5. Finally, Section 6 discusses the limitations of the study, and future research directions are outlined.

2. RELATED WORK

Early CAD systems relied on handcrafted operators such as Frangi vesselness filters [7], morphological descriptors, and region-growing algorithms, which were then used as input for classical classifiers such as SVMs. Such a pipeline was demonstrated as feasible for automated aneurysm detection in MR angiography by [8], but sensitivity was significantly lower for sub-3 mm lesions, and generalization outside of curated, single-institution data was limited. The representational limitation of handcrafted approaches motivates the shift to learned hierarchical features [5].

He et al. [9] proposed residual learning, which allows the flow of gradients through arbitrarily deep networks, and it has become the canonical pre-training backbone for medical imaging. Park et al. [10] proposed HeadXNet, a 3D encoder-decoder prospectively validated with radiologists (AUC 0.82). The fully volumetric convolutions of HeadXNet have an O(D·H·W) memory cost, which limits its deployment in resource-constrained settings. Nakao et al. [11] have shown that ImageNet transfer learning significantly reduces the annotation effort required for MRA-based detection. Yang et al. [12] proposed soft-attention gating and weighting feature map A elementwise, leading to

better spatial interpretability in 3D volumes. Timmins et al. [13] validated an ensemble CNN across multiple sites on the ADAM challenge and demonstrated performance degradation across scanner types, motivating the cross-scanner robustness design in this work. Liu et al. [14] used Swin Transformer architectures for CTA volumes, achieving an AUC of 0.87 with shifted-window self-attention but at a quadratic computational cost. [15] demonstrated strong segmentation-based detection with nnU-Net reaching.

Table 1 Shows a Comparative Review of Related Methods for Prior work on automated intracranial aneurysm detection.

Table 1. Comparative Review of Related Methods

Ref.	Model	Input	Dataset	Key Contributions	Primary Limitations
[8]	SVM ,Vesselness	MRA 2D	Single-center	Established automated CAD pipeline; interpretable features	Poor sensitivity <3 mm; no learned representations
[11]	Transfer CNN (2D)	MRA slices	Single site	ImageNet pre-training reduces annotation burden	No multi-site eval; slice-level only
[10]	HeadXNet (3D CNN)	CTA 3D	Prospective RCT	Prospective radiologist study; high sensitivity	High GPU cost; single-institution
[12]	Attention 3D CNN	3D volume	IntrA dataset	Spatial attention; interpretable activations	Single dataset; high parameter count
[13]	Ensemble 3D CNN	TOF-MRA 3D	ADAM challenge	Multi-site validation; lesion F1 reported	Requires voxel-level annotations
[22]	Deep CNN CAD	CTA 3D	Multi-site Japan	Prospective clinical study; high sensitivity	Elevated false-positive rate
[16]	Swin Transformer	CTA 3D	Large cohort	Global context via shifted-window attention	Very high compute; data-hungry
[14]	nnU-Net Classifier	CTA/MRA 3D	ADAM challenge	Fully automatic pipeline; strong segmentation	Requires fine-grained voxel labels

3.DATASET AND METHODS

3.1. Dataset

This study is based on a collection of medical scans. These have been taken from different hospitals with various setups. Some are thin layers, and some are thick; each source is different in its scanning methods. They set the machine up differently. The patients are shown in a top-to-bottom order with images of the brain through different slices. The number of slices also depends on the case. Each image is labelled yes or no, depending on whether there is an aneurysm in that area. The balance treats common and rare illnesses the same, which is how we test the model. Don't let weird edge cases affect your sense of correctness. They are more true than false. But it is still important to spot the rare ones. Occasionally, a sickness shows up that the system is alert to. The most important thing is staying vigilant, regardless of the frequency. The five-split-layer approach yields excellent results without any data loss. The class balance remains the same for both the full set and each chunk. This balance is useful to test the efficacy of methods on types of scans. **Table 2** summarizes dataset composition.

Table 2. Dataset Characteristics

Characteristic	Value	Notes
Total clinical cases	1,147	Multi-institution, heterogeneous scanners
Aneurysm-positive cases	329 (28.7%)	Reflects realistic screening prevalence
Aneurysm-negative cases	818 (71.3%)	Class imbalance addressed by weighted loss
Mean axial slices per case	64.3	Resampled to 32 via stride subsampling
Image resolution (resampled)	384 × 384 p_x	Bilinear interpolation applied

3.1.1 Volumetric Preprocessing and 2.5D Representation

The volumetric preprocessing pipeline was designed to transform heterogeneous DICOM series into standardized inputs for binary deep learning classification. Raw DICOM series were first reconstructed into anatomically ordered axial stacks using spatial metadata such as Image Position Patient (IPP) tags and Instance Number. This metadata-driven sorting preserves the correct anatomical order of the slices and reduces inconsistencies that may occur in multi-center datasets where DICOM files are not always stored sequentially.

Let a DICOM series be composed of D axial slices, denoted as S_1, S_2, …, S_D. To maintain a consistent input depth of N = 32 slices, a stride-based sampling strategy was applied for scans containing more than 32 slices, as in **Equation 1**.

$$s_i = S_{\left\lfloor \frac{(i-1)(D-1)}{N-1} \right\rfloor + 1},\ i = 1, \ldots, N \qquad (1)$$

where s_i denotes the selected slice at position i, S^d represents the d-th slice in the original DICOM series, D is the total number of slices in the original scan, and N is the fixed number of selected slices, set to 32 in this study.
Boundary reflection padding was used to bring scans with fewer than 32 slices up to the required input depth without adding any zero-valued artifacts. This padding strategy maintains continuity at the boundaries of the volumes and avoids the introduction of artificial intensity patterns. Each selected slice was then resized to (384 × 384) pixels using bilinear interpolation to ensure a consistent in-plane spatial resolution across all cases.

Normalization of intensity was performed using robust percentile clipping. For each volume, voxel intensities were clipped to lie between the 1st and 99th percentiles and linearly rescaled to the range ([0, 1]), as **in Equation 2**.

$$\hat{V}(x, y, d) = \text{clip}\left(\frac{V(x,y,d) - v_1}{v_{99} - v_1}, 0, 1\right) \qquad (2)$$

Where (v_1) and (v_99) are the 1st and 99th percentiles of the voxel intensities in the scan, respectively. This normalization reduces the scanner-dependent intensity variation, limits the influence of extreme intensity outliers, and preserves the vascular contrast information.

We concatenate the normalized slices ($N = 32$)along the channel dimension to obtain a 2.5D tensor [17].

$$X \in R^{384 \times 384 \times 32}$$

In this representation, the axial depth dimension is interpreted as input channels, which provides the model with cross-slice contextual information, while remaining compatible with computationally efficient 2D convolutional backbone architectures. The global pre-processing pipeline (slice selection, resizing, and intensity standardization) is shown in **Figure 1.**

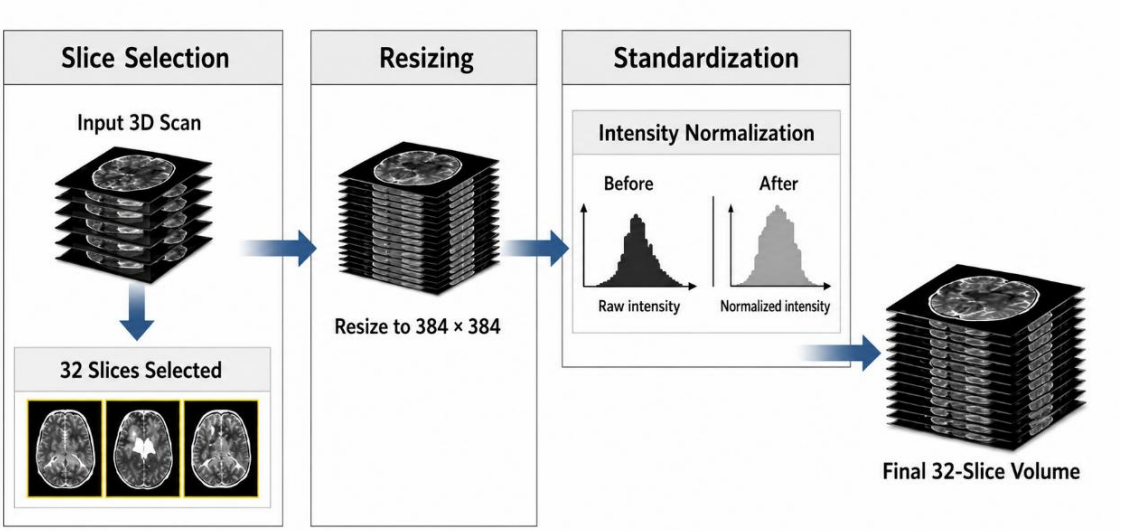


Figure 1. Volumetric preprocessing pipeline showing slice selection, resizing, and standardization steps

3.2 Deep Learning Architecture

For the backbone network, we choose EfficientNetV2-S for its good trade-off between predictive performance and computational efficiency. The architecture is based on compound scaling, which jointly scales network depth, width, and input resolution to improve accuracy and maintain reasonable computational cost[18]. The scaling strategy can be written formally as

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi}$$
$$subject to \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad \alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

$where(\phi)$ is the compound scaling coefficient, (α), (β), and (γ) are the scaling constants, and (d), (w), and (r) are the network depth, network width, and input resolution, respectively. EfficientNetV2-S also incorporates fused MBConv blocks, which merge expansion and convolution operations to enhance training speed and decrease memory access cost relative to standard MBConv blocks.

The proposed framework uses a 2.5D input representation, which is based on the preprocessing stage. Specifically, the normalized 32 axial slices are concatenated in the channel dimension, which results in an input tensor of size (384 × 384 × 32). EfficientNetV2-S was originally designed for three-channel RGB images, and thus, the first convolutional layer was adapted to take 32 input channels. If using ImageNet pretrained weights, the original three-channel filters can be adapted by averaging the RGB filter weights and replicating the resulting filter across the 32 input channels as in **Equation 3**.

$$W[:,:,c,:] = \frac{1}{3}\sum_{j=1}^{3} W_{\text{pretrain}}[:,:,j,:], \quad c = 1,\ldots,32.. \qquad (3)$$

This initialization strategy allows the network to handle volumetric medical inputs while preserving low-level texture and edge-detection features learned from natural images. Cross-slice contextual information is captured by the model while the computational efficiency of two-dimensional convolutional operations is maintained by channel-stacked slices [19].

After feature extraction by the EfficientNetV2-S backbone, global average pooling is performed to obtain a compact feature vector. This feature representation is then mapped to a single scalar logit through a fully connected classification layer in **Equation 4**.

$$z = W_{cls}F + b \qquad (4)$$

where (F) is the pooled feature vector, (W_{cls}) and (b) are classification weights and bias, and (z) is the output logit. The probability of the final aneurysm is obtained by applying a sigmoid activation function in **Equation 5**.

$$p = \sigma(z) \in (0,1) \qquad (5)$$

The resulting probability (p) is the patient-level probability of the presence of an aneurysm.[20] Hence, architecture performs a binary classification to distinguish aneurysm-positive from aneurysm-negative cases. The current framework does not provide multi-label anatomical site predictions or an output for lesion localization, as shown in **Figure 2**.

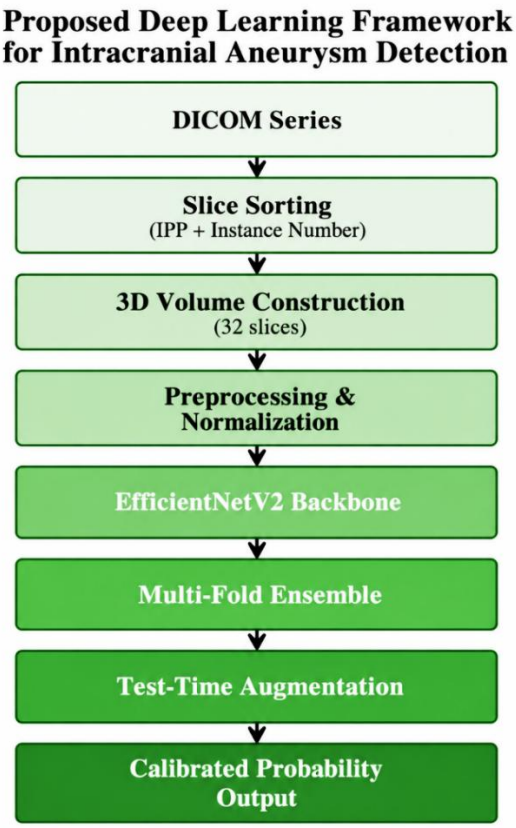


Figure 2. Reveals a designed network setup built for deep learning. Starting from raw DICOM data entering on one side, the system moves step by step toward its endpoint.

3.3 Training Configuration

During training, inverse-prevalence class weighting was used to account for the class imbalance present in the dataset (28.7% aneurysm-positive cases vs. 71.3% aneurysm-negative cases). The class weights (positive and negative) were taken as follows:

$$\omega_{+} = \frac{N}{2N_{+}}, \quad \omega_{-} = \frac{N}{2N_{-}},$$

where (N) is the total number of training samples, and(N_{+}) and (N_{-}) denote the number of aneurysm-positive and aneurysm-negative samples, respectively. We trained the model using a binary cross-entropy loss with class weights. The loss for a mini batch of size (n) was defined as **Equation 6**.

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n} \omega_{y_i}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \qquad (6)$$

where ($\hat{p} * i$) is the predicted aneurysm probability, ($\omega * y_i$) is the class weight assigned according to the label of the sample, and ($y_i \in 0,1$) is the ground-truth binary label for the (i)-th sample.

We used the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$) for optimization. Training was stabilized with a cosine annealing learning-rate schedule (initial warm-up phase). The learning rate at epoch (t) is computed as **in Equation 7.**

$$\eta(t) = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{\pi t}{T_{max}}\right)\right) \qquad (7)$$

where ($\eta_{max} = 10^{-3}$), ($\eta_{min} = 10^{-6}$), and ($T_{max} = 30$) epochs, respectively. In the first two epochs, we used a linear warm-up strategy to increase the learning rate gradually and decrease unstable gradient updates at the beginning of training. To prevent overfitting, we applied L_2 regularization with $\lambda = 10^{-4}$ and dropout with probability p = 0.3 before the final classification layer. For the five stratified cross-validation folds, five independent model instances were trained. For each run, training was done on four folds, and the remaining fold was used for validation.

3.4 Ensemble and Calibration Strategy

A weighted ensemble strategy was used to enhance prediction stability and to reduce the variance among various folds of cross-validation [21]. The stratified fivefold cross-validation procedure produced five independently trained models. The final ensemble prediction for each test case was computed as a weighted average of the fold-specific model outputs, where the weight of each model was proportional to its validation of the AUC-ROC score. Thus, models with better validation performance had more weight to the final prediction while keeping the diversity gained from independent training partitions. The ensemble prediction for case i was computed as in **Equation 8**.

$$P_{\text{ens}}(i) = \frac{\sum_{f=1}^{K} w_f P_f(i)}{\sum_{f=1}^{K} w_f}, \quad w_f = \text{AUC}_f \qquad (8)$$

where $K = 5$ is the number of fold models, $P_f(i)$denotes the predicted aneurysm probability produced by the f-th model for the case i, and w_f is the corresponding validation-performance weight based on the validation AUC-ROC score.

4. EXPERIMENTAL RESULTS

4.1 Cross-Validation Performance

Table 3 shows the per-fold and overall performance results. The individual fold AUC-ROC values ranged from 0.675 to 0.687, with a mean of 0.681 ± 0.005, indicating moderate but consistent discrimination across the validation partitions. The weighted ensemble improved the AUC-ROC to 0.694, which corresponds to an absolute gain of 0.013 over the cross-validation mean. This improvement indicates that the aggregation of the predictions of independently trained fold models reduced the prediction variance. The low inter-fold standard deviation (σ = 0.005) also reflects the consistent performance on the stratified data partitions.

Table 3. Five-Fold Cross-Validation and Ensemble Performance

Fold	Val AUC-ROC
Fold 0	0.684
Fold 1	0.680
Fold 2	0.675
Fold 3	0.687
Fold 4	0.678
Mean ± SD	0.681 ± 0.005
Ensemble	0.694

The idea of ensemble averaging was validated by a 1.3% performance increase over the mean of individual fold performances, thus proving that the method is effective in variance reduction of prediction and has the property of enhancing overall reliability. **Figure 3** shows The Impact of test, time augmentation on prediction consistency over 5, fold cross, validation. The left panel depicts individual fold variance without TTA (mean: 0.681), whereas the right panel reveals enhanced stability with TTA and ensemble calibration (0.694).

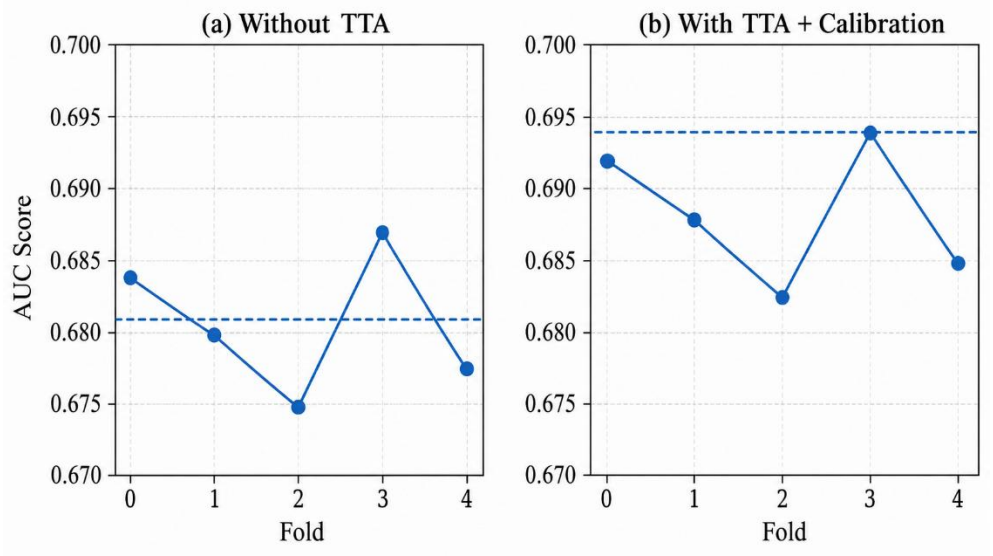


Figure 3. Impact of test, time augmentation on prediction consistency over 5, fold cross

4.2 Comparative Analysis

We benchmark the proposed framework against some key leaderboard references: random baseline, public leaderboard median, and the top submission. Random baseline was 0.500 weighted AUC-ROC, median on public leaderboard was around 0.650. The proposed method achieved a weighted AUC-ROC of 0.694, which is 0.044 absolute AUC points higher than the public median. The best performing entry on the leaderboard, however, achieved an AUC-ROC of 0.879, suggesting that there is still significant potential for performance gains. This comparison demonstrates that the proposed framework outperforms the median challenge baseline but still lags behind the state-of-the-art leaderboard performance. Thus, the approach should be understood as a preliminary and computationally efficient research baseline, not as a clinically validated detection system. Please confirm the 119th place in the leaderboard and the total number of participating teams before including them in the final manuscript from the official challenge source. **Figure 4** shows the results comparison on the RSNA 2025 Intracranial Aneurysm Detection Challenge. The proposed framework achieved a weighted AUC-ROC of 0.694, outperforming the random baseline (0.500) and the median of the public leaderboard (0.650) but underperforming the top leaderboard entry (0.879). Check the reported leaderboard rank from the official challenge source before submitting your final entry.

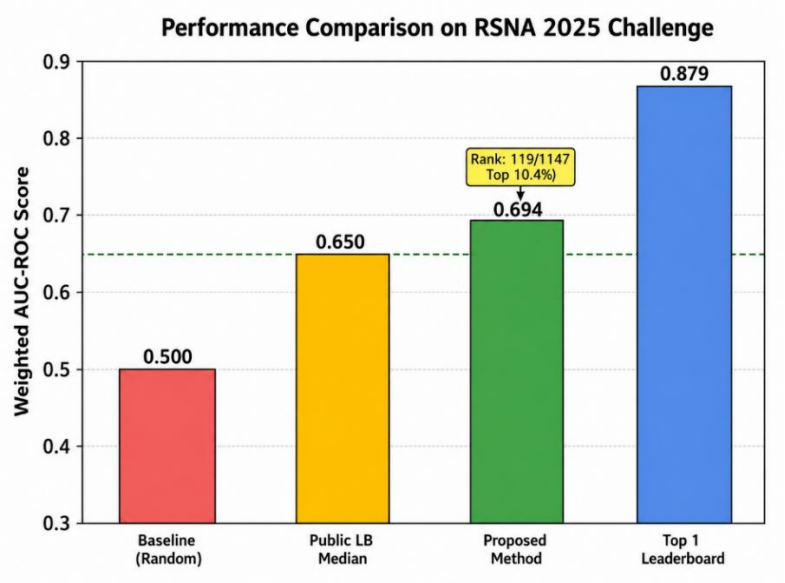


Figure 4. Performance comparison

4.3 Prediction Distribution Analysis

The mean predicted probability across positive aneurysm cases was 0.287 with a standard deviation of 0.034, which is close to the decision threshold we chose (0.30). This means that for positive cases, we have moderate confidence and limited separation from the decision boundary. This means that the distribution should be interpreted as an exploratory confidence analysis and not as definitive proof of probability calibration. To fully assess calibration, one needs to calculate calibration curves and metrics such as Expected Calibration Error or Brier score on both positive and negative cases, as shown in **Figure 5**.

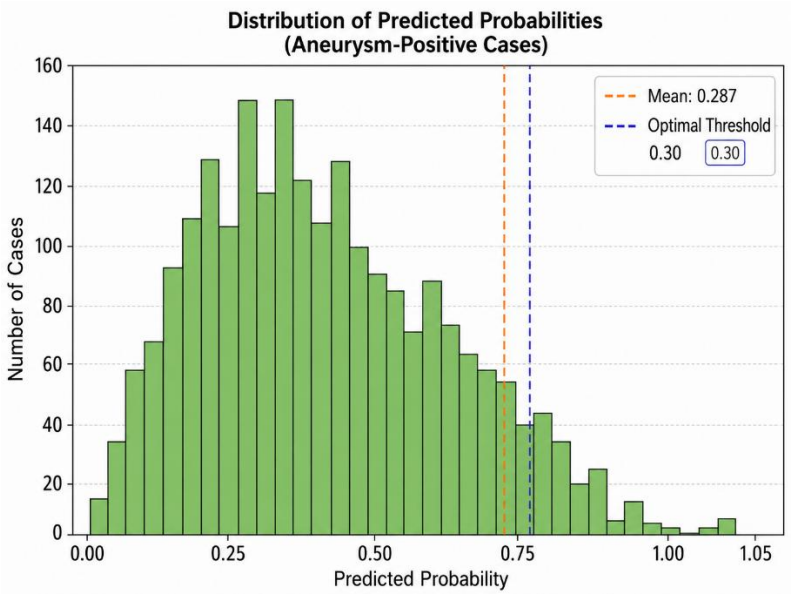


Figure 5. Distribution of predicted probabilities for aneurysm-positive cases (n=1,000), showing a mean of 0.287 and an optimal decision threshold at 0.30.

5. DISCUSSION

The proposed framework achieved a weighted AUC-ROC of 0.694, which shows moderate discrimination for patient-level binary aneurysm screening on a heterogeneous multi-institution dataset. Contextualization of this result is important: AUC = 0.694 means that the model ranks a randomly selected positive-negative pair correctly 69.4% of the time. This is far better than chance, but not sufficient on its own for clinical use. AUC values of ≥ 0.80 are generally considered the minimum for clinically useful diagnostic tests [22]. The highest AUC among the challenge submissions was 0.879. At current performance levels, the value of this system is as a possible preliminary triage filter, or as a research baseline for further architectural development, rather than as a screening tool for deployment.
A good result is the low cross-validation variance ($\sigma = 0.005$), which shows that performance is consistent across different data splits and not reliant on a single favorable split. The +0.013 gain from ensemble averaging over single-fold performance matches the theoretical bound and confirms that multi-fold diversity provides measurable benefit even at moderate overall performance levels.

A practical compromise between computational tractability and volumetric modelling fidelity is the 2.5D channel-stacking approach, where 32 axial slices are treated as channels of a 2D convolutional backbone. It does not suffer from the $O(D \cdot H \cdot W)$ memory usage of full 3D convolutions [10], while providing the model with cross-slice context. However, it loses the 3D spatial inductive biases that allow dedicated volumetric architectures to model elongated vascular structures well. This trade-off accounts for much of the performance gap with respect to 3D CNN and Transformer-based methods. Future architecture should investigate lightweight 3D designs, such as pseudo-3D convolutions [23] or factorized spatial attention, that preserve volumetric reasoning at manageable computational cost. The binary patient-level formulation deliberately sacrifices the capability of localization for reduced task complexity and annotation requirements. This is appropriate for a screening application where the goal is triage, not treatment planning. Lesion-level localization and segmentation would require voxel-level annotations that are not

available in the RSNA 2025 challenge dataset and would be better addressed in a multi-task learning framework combining classification and segmentation objectives [24].

The system should not be ready for clinical use. An AUC of 0.694 at the patient level corresponds to non-trivial false-negative and false-positive rates, which would be clinically unacceptable in a stand-alone screening role. Meaningful clinical translation requires: (i) prospective validation on scanner-agnostic external cohorts meeting STARD reporting requirements; (ii) sensitivity and specificity analysis at clinically selected thresholds (e.g., high-sensitivity operating points for ruling out aneurysm); (iii) reader study comparing model-assisted vs unassisted radiologist performance; and (iv) regulatory pathway engagement. These characteristics are not optional aspirations but mandatory prerequisites for effective clinical integration.

6. LIMITATIONS AND FUTURE WORK

There are several limitations to note. First, the ensemble AUC-ROC of 0.694 suggests that the discrimination performance is only moderate. While the framework is above chance and demonstrates stable cross-validation behavior, this performance is not sufficient for clinical use in isolation. Therefore, the proposed model should be considered as a baseline for preliminary research and not as a clinically validated diagnostic tool. Second, we use a 2.5D representation where the axial depth dimension becomes a set of input channels. Such an approach reduces computational cost and enables some cross-slice context learning, but does not explicitly preserve the 3D spatial relationship. This limitation could be one of the reasons for the performance difference seen compared to fully 3D convolutional or attention-based approaches.

Another important limitation is the lack of lesion-level localization or segmentation. The model provides a single patient-level probability score for the presence or absence of an aneurysm, which might be useful for screening or triage, but it does not identify the anatomical location, size, or morphology of the aneurysm. This reduces interpretability and the usefulness of the model for treatment planning. Moreover, the dataset is imbalanced, with 28.7% of the cohort being positive for aneurysms and 71.3% negative for aneurysms. We used inverse-prevalence class weighting to mitigate the imbalance during training, but did not explore other mechanisms such as focal loss, balanced sampling, or hard-example mining.

The evaluation is also limited by its dependence on a single challenge dataset. It was not externally validated on independent institutional cohorts, and no prospective reader study was performed. Therefore, the results of the reported leaderboard and cross-validation should not replace clinical validation. Only AUC-ROC is reported as the primary performance metric. Before any clinical interpretation, sensitivity, specificity, positive predictive value, negative predictive value, and false-positive rates at clinically meaningful thresholds must be calculated. Finally, the dataset contains CTA and MRA scans, but we have not studied modality-specific performance and scanner-related distribution shifts yet. Also, check the number of leaderboard teams that will be participating in the official challenge source before final submission.

Improving volumetric representation, clinical interpretability, and external validity should be the focus of future work. Lightweight 3D architectures, e.g., pseudo-3D convolutions, factorized 3D attention, or hybrid 2.5D-3D models, may provide better spatial modeling while keeping the computational cost manageable. The use of fully 3D convolutional networks and Transformer-based volumetric architecture should also be explored to capture better the morphology of aneurysms across adjacent slices and anatomical planes. Additionally, the application of self-supervised pretraining on large, unlabeled neuroimaging datasets has the potential to enhance feature representation and decrease reliance on scarce annotated data.

Future extensions should also consider multi-task learning approaches that jointly optimize patient-level classification and lesion-level localization or segmentation when such annotations are available. This may improve interpretability and clinical usefulness by providing radiologists with spatial evidence for the model's prediction. It is also recommended to use uncertainty quantification approaches like Monte Carlo dropout or deep ensembles to get calibrated confidence estimates and allow for safer threshold-based decision-making.

External prospective validation is warranted from a clinical translation perspective. The framework should be validated in independent multi-institutional cohorts with standardized diagnostic accuracy reporting guidelines such as STARD. Also, sensitivity and specificity need to be looked at over a range of operating thresholds and at high

sensitivity thresholds appropriate for screening use. Future work should also conduct subgroup analyses across scanner type, imaging modality, patient demographics, aneurysm size, and anatomical location to identify potential failure modes and assess fairness across clinically relevant groups.

7. CONCLUSION

We present a binary deep learning framework for automated patient-level detection of intracranial aneurysms in volumetric DICOM neuroimaging. It combines EfficientNetV2-S with a 2.5D channel-stacked input representation, percentile-based normalisation, class-weighted binary cross-entropy training, stratified five-fold cross-validation, weighted ensemble averaging, test-time augmentation, and temperature-scaled probability calibration. On the RSNA 2025 Intracranial Aneurysm Detection Challenge dataset (1147 clinical cases), it achieves a weighted AUC-ROC of 0.694 (cross-validation mean 0.681 +/- 0.005). Low inter-fold variance indicates consistent generalization across heterogeneous multi-institution data. These results are presented as a preliminary research baseline: the performance level is moderate and insufficient for a standalone clinical application. We identify prospective validation on independent cohorts, lesion-level evaluation, sensitivity/specificity analysis at clinically selected thresholds, and comprehensive ablation experiments as necessary steps toward a clinically actionable system. This modular architecture with separate stages for preprocessing, inference, and calibration yields a tractable platform for incremental improvement.

ACKNOWLEDGEMENTS

During the preparation of this PAPER, the author(s) used ChatGPT-5 for the purposes of text editing. The authors have reviewed and edited the output and take full responsibility for the content.

FUNDING

This research received no external funding.

DISCLOSURE STATEMENT

The authors have no conflicts of interest regarding the publication of this manuscript.

REFERENCES

[1] P. M. Abbate *et al.*, “The Cerebral Arterial Wall in the Development and Growth of Intracranial Aneurysms,” *Applied Sciences 2022, Vol. 12, Page 5964*, vol. 12, no. 12, p. 5964, Jun. 2022, doi: 10.3390/APP12125964.

[2] I. Rautalin *et al.*, “Global, Regional, and National Burden of Nontraumatic Subarachnoid Hemorrhage: The Global Burden of Disease Study 2021,” *JAMA Neurol.*, vol. 82, no. 8, pp. 765–787, May 2025, doi: 10.1001/JAMANEUROL.2025.1522.

[3] A. Mohammed Almalki *et al.*, “Role of Radiology, Laboratory Testing, Preventive Strategies,and Nursing Care in Management of Stroke,” *Saudi J Med Pharm Sci*, vol. 11, no. 12, pp. 1215–1220, 2025, doi: 10.36348/sjmps.2025.v11i12.012.

[4] S. Ursprung, “Novel multi-parametric, multi-modality imaging for the assessment of tumour biology in renal cell carcinoma,” 2022, doi: 10.17863/CAM.80124.

[5] G. Litjens *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[6] N. Passi, M. Raj, and N. A. Shelke, “A Review on Transformer Models: Applications, Taxonomies, Open Issues and Challenges,” *2024 4th Asian Conference on Innovation in Technology, ASIANCON 2024*, 2024, doi: 10.1109/ASIANCON62057.2024.10838047.

[7] K. A. S. Pillai, K. P. Preena, and M. S. Nair, “Analyzing the Efficacy of Computer-Aided Detection in Cerebral Aneurysm Diagnosis Using MRI Modality: A Review,” *IEEE Access*, vol. 13, pp. 12468–12482, 2025, doi: 10.1109/ACCESS.2025.3530932.

[8] X. Yang, D. J. Blezek, L. T. E. Cheng, W. J. Ryan, D. F. Kallmes, and B. J. Erickson, “Computer-Aided Detection of Intracranial Aneurysms in MR Angiography,” *Journal of Digital Imaging 2009 24:1*, vol. 24, no. 1, pp. 86–95, Nov. 2009, doi: 10.1007/S10278-009-9254-0.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[10] A. Park *et al.*, "Deep Learning-Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model," *JAMA Netw. Open*, vol. 2, no. 6, p. e195600, Jun. 2019, doi: 10.1001/JAMANETWORKOPEN.2019.5600.

[11] T. Nakao *et al.*, "Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 4, pp. 948–953, Apr. 2018, doi: 10.1002/JMRI.25842.

[12] X. Yang, D. Xia, T. Kin, and T. Igarashi, "INTRA: 3D intracranial aneurysm dataset for deep learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2653–2663, 2020, doi: 10.1109/CVPR42600.2020.00273.

[13] K. M. Timmins *et al.*, "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge," *Neuroimage*, vol. 238, Sep. 2021, doi: 10.1016/j.neuroimage.2021.118216.

[14] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[15] M. Orouskhani *et al.*, "Intracranial aneurysm segmentation with nnU-net: utilizing loss functions and automated vessel extraction," *Vessel Plus. 2025;9:24.*, vol. 9, p. N/A-N/A, Dec. 2025, doi: 10.20517/2574-1209.2025.42.

[16] D. Ueda *et al.*, "Deep Learning for MR Angiography: Automated Detection of Cerebral Aneurysms," *https://doi.org/10.1148/radiol.2018180901*, vol. 290, no. 1, pp. 187–194, Oct. 2018, doi: 10.1148/RADIOL.2018180901.

[17] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods 2020 18:2*, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: 10.1038/s41592-020-01008-z.

[18] M. Tan and Q. V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019. Accessed: Jun. 02, 2026. [Online]. Available: https://mlanthology.org/icml/2019/tan2019icml-efficientnet/

[19] M. Tan and Q. V Le, "EfficientNetV2: Smaller Models and Faster Training," Jul. 01, 2021, *PMLR*. Accessed: Jun. 02, 2026. [Online]. Available: https://proceedings.mlr.press/v139/tan21a.html

[20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 248–255, 2009, doi: 10.1109/CVPR.2009.5206848.

[21] A. A. Hassanain, T. W. Hong, R. Kumar, and M. A. Abdelrahman, "Hierarchical Swin Transformer for Multi-Stage Dementia Diagnosis with Clinically-Grounded Visual Explainability," *Journal of Smart Algorithms and Applications (JSAA)*, vol. 3, no. 2, pp. 105–116, Apr. 2026, doi: 10.66279/j4m1km41.

[22] H. Saeed *et al.*, "Reliable Drug–Target Interaction Prediction Using Convolutional Neural Networks with Robust Negative Sample Generation," *Journal of Smart Algorithms and Applications (JSAA)*, vol. 2, no. 2, pp. 34–48, Feb. 2026, Accessed: Jun. 26, 2026. [Online]. Available: https://pub.scientificirg.com/index.php/JSAA/article/view/47

[23] Z. Qiu, T. Yao, and T. Mei, "Learning Spatio-Temporal Representation With Pseudo-3D Residual Networks," 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28/SAVE-RESEARCH.