A Winning Strategy for the RSNA 2025 Intracranial Aneurysm Detection Challenge: An Integrated, State-of-the-Art Approach

Strategic Foundations: Learning from the Giants of Medical Imaging Competitions

The pursuit of a top-tier solution in a highly competitive arena such as the RSNA 2025 Intracranial Aneurysm (IA) Detection Challenge necessitates a strategy built upon the foundational principles of past successes, while simultaneously innovating beyond established paradigms. An analysis of analogous, large-scale medical imaging competitions reveals an evolutionary trajectory of techniques, from which a robust and forward-looking strategy can be distilled. The goal is not to merely replicate historical methods but to understand the underlying causal factors for their success and adapt them to the unique challenges of aneurysm detection.

The RSNA 2019 Intracranial Hemorrhage Challenge: A Starting Point, Not a Destination

The 2019 RSNA Intracranial Hemorrhage (ICH) Detection Challenge serves as a valuable, albeit dated, reference point.[1] The competition's objective was to detect and classify acute intracranial hemorrhage and its subtypes from head CT scans. Submissions were evaluated using a weighted multi-label logarithmic loss, where the presence of

any hemorrhage was weighted more heavily than specific subtypes, a design choice that emphasized sensitivity for the primary finding.[1]

A review of the top-performing solutions from 2019 reveals a consistent architectural pattern: a "2.5D" approach that combined 2D Convolutional Neural Networks (CNNs) with sequential models.[2] In this paradigm, each axial slice of a 3D CT volume was processed independently by a 2D CNN to extract intra-slice spatial features. The resulting sequence of feature vectors, representing the entire scan, was then fed into a recurrent neural network (RNN), typically a Bidirectional Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM) network, to model the inter-slice context.[2] This method was a clever and effective workaround for the technical limitations of the time. Full 3D CNNs were computationally prohibitive for many teams, and standard 2D CNNs were incapable of capturing the crucial through-plane information necessary for accurate diagnosis. The 2.5D approach offered a pragmatic compromise. Furthermore, ensembling was a ubiquitous strategy for performance enhancement, with leading teams combining the outputs of anywhere from seven to thirty-one individual models to achieve their final scores.[2]

However, the technological landscape of deep learning has undergone a significant transformation since 2019, primarily due to the ascendancy of Transformer-based architectures. The 2.5D CNN+RNN pipeline, while innovative for its time, is now conceptually and computationally superseded. This older approach was fundamentally a solution to a problem that Transformers solve natively: modeling long-range dependencies. The self-attention mechanism at the core of the Transformer architecture is explicitly designed to capture relationships between all elements in a sequence, regardless of their distance.[3] In the context of a 3D medical scan, this means a 3D Transformer can directly model the relationships between voxels across the entire volume, effectively learning the inter-slice context that the RNN was tasked with approximating.

Recent research has explicitly demonstrated the superiority of Transformer-based models for the very same ICH detection task, achieving better performance with a fraction of the parameters and computational cost of the 2019 winning solutions.[2] This progress signals a clear directive: a winning strategy for 2025 must not be anchored to the methods of 2019. It must be fundamentally 3D and must leverage architectures capable of natively and efficiently processing volumetric data. The 2.5D approach, born of necessity, is no longer necessary.

The LUNA16 Lung Nodule Challenge: The True Strategic Blueprint

A more strategically relevant precedent for intracranial aneurysm detection is the LUNA16 (LUng Nodule Analysis 2016) challenge.[5] The parallels are striking: both tasks involve the detection of small, focal, and often subtle pathological findings within large 3D volumetric CT scans, against a backdrop of complex anatomical structures. The class imbalance is extreme in both cases, with the target pathology occupying a minuscule fraction of the total voxels.

The dominant and most successful strategy to emerge from LUNA16 was the **two-stage pipeline**.[7] This approach decouples the problem into two distinct, manageable sub-problems:

1. **Candidate Detection:** A high-sensitivity, lower-specificity model is first used to generate a list of potential nodule candidates. This stage is designed to cast a wide net, ensuring that very few, if any, true nodules are missed, at the cost of including many false positives.[7]
2. **False Positive Reduction:** A second, more powerful and computationally intensive model is then used to classify these candidates. This classifier focuses exclusively on the small

3D patches extracted around each candidate, learning the fine-grained features necessary to distinguish true nodules from mimics.[7]

This two-stage paradigm proved immensely successful.[10] Winning teams demonstrated that this approach was more effective than attempting to build a single, monolithic model to perform detection and classification simultaneously. Furthermore, top LUNA16 competitors leveraged auxiliary information to enhance their classifiers. They utilized the richer LIDC-IDRI dataset, from which LUNA16 was derived, to train their models not just to detect nodules but also to predict associated features like radiologist-annotated malignancy scores, which served as a powerful proxy for cancerous features.[13]

The adoption of a two-stage pipeline is not merely a tactical choice; it is a strategic imperative for risk mitigation and performance maximization in the context of problems like aneurysm detection. An end-to-end model trained on an entire 3D volume faces the daunting task of simultaneously learning to locate rare, small objects and classify them, all while being exposed to an overwhelming number of background voxels. This can lead to inefficient training, where the model's capacity is diluted by learning to represent the vast negative space, potentially compromising its ability to discern the subtle features of a true positive.[16]

The two-stage approach elegantly circumvents this dilemma. The initial candidate generation stage acts as an efficient filter, dramatically reducing the search space from billions of voxels to a few hundred candidates per scan. This allows the second-stage classifier to dedicate its full representational power to the critical task of distinguishing true aneurysms from confounding structures like vessel bifurcations or infundibula. This focused learning environment is far more conducive to achieving the high precision and specificity required to top the leaderboard. Therefore, the entire modeling strategy proposed in this report will be built upon the proven foundation of a two-stage detection and classification pipeline, adapting the core principles of the LUNA16 winners to the specific domain of intracranial aneurysm detection.

Architecting the Core Model: A 3D-First, Transformer-Centric Approach

The selection of the core model architecture is the cornerstone of the technical strategy. Informed by the limitations of past approaches and the capabilities of modern networks, the proposed architecture will be fully three-dimensional, leveraging the power of Vision Transformers to build a state-of-the-art system for volumetric analysis.

The Imperative of 3D Volumetric Analysis

Intracranial aneurysms are inherently three-dimensional pathologies. Their morphology—including size, shape, aspect ratio, and relationship to parent vessels—is critical for both detection and clinical risk assessment.[18] A 2D slice-based analysis, by its very nature, loses this crucial spatial context and can be ambiguous. A small vessel viewed in cross-section can mimic a tiny aneurysm, a distinction that becomes clear only when viewed in the full volumetric context.

Therefore, a successful model must operate on the full 3D data. Architectures developed specifically for 3D medical imaging, such as the multi-scale 3D CNN DeepMedic [19] and the ubiquitous 3D U-Net [21], have consistently demonstrated the superiority of volumetric processing for tasks like lesion segmentation. Grad-CAM visualizations from 3D CNNs have confirmed that these models learn to focus on clinically relevant 3D areas, enhancing both diagnostic accuracy and interpretability.[23] A 3D-first approach is not an option; it is a prerequisite for a competitive solution.

Swin UNETR: The State-of-the-Art Backbone

While 3D CNNs represent a significant step up from 2D methods, they are not without limitations. The core operation of a CNN, the convolution, uses a kernel with a fixed, local receptive field. This makes it challenging for standard CNNs to model long-range spatial dependencies within an image [3]—for example, understanding the relationship between a potential aneurysm in the anterior communicating artery and the overall structure of the Circle of Willis.

This is precisely the weakness that Vision Transformers (ViTs) are designed to overcome. The Swin UNETR architecture stands as the current state-of-the-art for 3D medical image analysis, particularly for segmentation and related dense prediction tasks.[3] It masterfully fuses the strengths of two paradigms into a single, powerful hybrid model. The encoder is a hierarchical Swin Transformer, which processes the 3D volume by computing self-attention within non-overlapping, shifted windows. This design allows it to capture global context and long-range dependencies efficiently, with computational complexity that scales linearly with the input size, a crucial advantage for high-resolution 3D data.[25] The decoder is a CNN-based architecture, similar to that of a U-Net, which receives features from the encoder at multiple resolutions via skip connections.[3]

This hybrid design is the key to Swin UNETR's success. A pure Transformer model might excel at global context but can struggle to preserve the fine-grained local details necessary for precise localization. A pure CNN excels at local features but has a limited global view. Swin UNETR gets the best of both worlds. The Transformer encoder builds a rich, multi-scale understanding of the global anatomical context, while the CNN decoder uses this information to reconstruct a precise, voxel-level output, making it exceptionally well-suited for our task. It can understand the overall vascular tree (global context) while precisely delineating the boundaries of a tiny aneurysm (local detail). Its performance on demanding 3D segmentation benchmarks like BTCV and MSD consistently surpasses both pure CNN and other Transformer variants.[24]

Architecture	**Architecture Type**	**Core Mechanism**	**Strengths**	**Weaknesses**	**Suitability for Aneurysm Detection**
3D U-Net	CNN	Multi-scale Convolution, Skip Connections	Excellent for precise localization, widely adopted baseline.[21]	Limited effective receptive field, struggles with long-range context.[3]	Strong candidate for Stage 1 (Candidate Generation).
DeepMedic	CNN	Multi-scale 3D Convolution	Computationally efficient, proven in lesion segmentation competitions.[19]	Less flexible than U-Net, older architecture.	Alternative candidate for Stage 1 (Candidate Generation).
Swin UNETR	Hybrid Transformer-CNN	Hierarchical Shifted-Window Self-Attention	Models global and long-range context effectively, linear	Data-hungry, can be difficult to train from	State-of-the-art choice for Stage 2 (Classification).

		(Encoder), CNN (Decoder)	complexity, SOTA performance.[24]	scratch without pre-training.[26]	

The Self-Supervision Mandate: Training a World-Class Model from Scratch

A well-documented challenge of ViT architectures is their significant data appetite and a relative lack of inductive bias compared to CNNs. This can make them difficult to train effectively from scratch, especially on medical datasets which are often limited in size.[26] The most effective strategy to overcome this is not to abandon Transformers, but to pre-train them using self-supervised learning on a large corpus of unlabeled data.

The proposed strategy involves creating a powerful, pre-trained Swin UNETR encoder that will serve as the backbone for our high-performance classifier. This process follows the framework laid out in the literature [25]:

1. **Large-Scale Data Curation:** The first step is to assemble a large, diverse, and unlabeled dataset of 3D head scans. This can include data from previous RSNA challenges (e.g., the ICH dataset [28]), the LUNA16 dataset [6], and any other publicly available head CT, CTA, or MRA datasets. The goal is to collect thousands of scans to expose the model to a wide variety of anatomies, scanners, and acquisition protocols.[25]
2. **Self-Supervised Pre-training via Proxy Tasks:** The Swin UNETR encoder will be pre-trained on this large, unlabeled dataset. Instead of predicting human-provided labels, the model is trained to solve a set of "proxy" or "pretext" tasks for which labels can be generated automatically from the data itself. A powerful combination of tasks includes [25]:
    - **3D Image Inpainting:** Randomly masking or cutting out a sub-volume of the input scan and training the model to predict the missing content. This forces the model to learn contextual information and local textures.
    - **3D Rotation Prediction:** Applying a random rotation (e.g., 0, 90, 180, or 270 degrees) to the input volume and training the model to classify which rotation was applied. This teaches the model to understand the canonical orientation and structural content of the anatomy.

- **Contrastive Learning:** Creating two or more augmented views of the same input scan (e.g., through random cropping, flipping, or intensity shifts). The model is trained to produce similar feature embeddings for views from the same scan (positive pairs) and dissimilar embeddings for views from different scans (negative pairs).

This pre-training phase is far more than a simple initialization technique to improve convergence. It is a process of creating a **foundational model of general neurovascular anatomy**. By learning to solve these challenging proxy tasks on thousands of diverse scans, the encoder develops a rich, hierarchical understanding of what constitutes a "normal" brain and vascular system. It learns the typical textures of bone, gray matter, and white matter; the expected shape and location of the ventricles; and the complex branching patterns of the cerebral arteries.

When this pre-trained encoder is later fine-tuned on the much smaller, expert-labeled aneurysm dataset, the learning task is fundamentally transformed. The model is no longer learning anatomy from scratch. Instead, it is fine-tuning its robust anatomical prior to recognize a specific, rare pathological deviation—the aneurysm. This approach dramatically reduces the amount of labeled data required to achieve high performance, accelerates training convergence, and results in a model that is significantly more robust and generalizable to unseen data from different clinical sites—a crucial advantage in a competition setting.

The High-Performance Two-Stage Detection and Classification Pipeline

Operationalizing the strategic framework derived from the LUNA16 challenge requires a carefully designed two-stage pipeline. Each stage will have a distinct objective, a tailored model architecture, and a specific evaluation focus, working in concert to maximize the final competition score.

Stage 1: High-Recall Candidate Generation

The singular objective of the first stage is to identify every potential aneurysm with the highest possible sensitivity. At this point, precision is a secondary concern; the primary goal is to minimize false negatives, ensuring that true aneurysms are passed to the next stage for scrutiny. The evaluation metric for this internal stage is recall.

For this task, a lightweight and fast 3D segmentation model is ideal. The architecture does not need the full representational power of the final classifier, but it must be efficient enough to process entire 3D volumes quickly. Two excellent candidates are:

- **3D U-Net:** The standard 3D U-Net architecture is a proven workhorse in medical image segmentation.[21] Its symmetric encoder-decoder structure with skip connections is effective at producing voxel-wise predictions. Numerous open-source implementations are available, making it a reliable and straightforward choice.
- **DeepMedic:** This multi-scale 3D CNN was a winning architecture in several lesion segmentation challenges.[19] Its dual-pathway design processes the input at different resolutions simultaneously, allowing it to capture context efficiently. It is another strong, battle-tested option for this stage.

The chosen model will be trained on the competition data to produce a binary segmentation mask, where voxels belonging to potential aneurysms are labeled as '1' and all others as '0'. The output probability map from the model will be thresholded at a very low value to favor high recall. Following inference, a connected-components analysis will be performed on the resulting binary mask. The centroid coordinates of each distinct connected component will be extracted, forming the list of candidates to be fed into Stage 2. This process mirrors the candidate generation step that was a key component of the LUNA16 challenge workflow.[7]

Stage 2: High-Precision Classification and False Positive Reduction

The objective of the second stage is to take the list of candidate coordinates generated by Stage 1 and classify each one with the highest possible precision and specificity, thereby eliminating the false positives. The final output of this stage will be the probability score for each candidate that is submitted to the competition leaderboard. The evaluation metric here is the competition's weighted Area Under the Receiver Operating Characteristic Curve (AUC).

The model for this critical stage will be the powerful, self-supervised pre-trained **Swin UNETR** architecture developed in the previous section. Its ability to model global context while attending to fine-grained local details makes it perfectly suited for the difficult task of distinguishing a true aneurysm from its anatomical mimics.

The implementation workflow for this stage is as follows:

1. For each candidate coordinate received from Stage 1, a 3D patch of a fixed size (e.g., 64x64x64 or 96x96x96 voxels) is extracted from the original, preprocessed 3D scan, centered on the candidate's coordinate.
2. This collection of 3D patches constitutes the training and inference dataset for the Stage 2 classifier.
3. The pre-trained Swin UNETR model is then fine-tuned on these patches. The task is binary classification: to output a single probability score (between 0 and 1) representing the likelihood that the patch contains an aneurysm.

This patch-based classification approach is highly effective and computationally efficient. It allows the most powerful model in the pipeline to focus its capacity exclusively on the small, information-dense regions of interest, rather than wasting computation on the vast, empty background of the full scan. This strategy was central to the success of many top teams in the LUNA16 and Data Science Bowl 2017 lung cancer detection competitions [10] and was validated in a large-scale clinical study for aneurysm detection on CTA, which used a similar cascaded architecture with a local fine-grained network for false positive reduction.[17]

Optimizing for Victory: Mastering the Weighted AUC Metric

Achieving a top rank on the leaderboard requires more than just a powerful model; it demands a training strategy that is explicitly tailored to the competition's evaluation metric and the inherent challenges of the data. For aneurysm detection, this means confronting the dual problems of extreme class imbalance and the need to optimize for a ranking-based metric like weighted AUC.

The Challenge: Extreme Imbalance and a Ranking-Based Metric

The competition dataset will be characterized by profound class imbalance. Intracranial aneurysms are rare findings; the number of positive samples (individual patients, or more granularly, voxels or patches containing an aneurysm) will be dwarfed by the number of negative samples by several orders of magnitude. Standard loss functions, such as Binary Cross-Entropy (BCE), are ill-suited for this scenario. With BCE, the total loss is a simple sum over all samples. In an imbalanced setting, the contribution from the overwhelming number of easy negative samples will dominate the loss signal, biasing the model's training towards simply predicting 'negative' and under-emphasizing the critical gradient updates from the rare positive examples.[30]

Compounding this issue is the nature of the AUC metric itself. AUC is a measure of a classifier's ability to rank a random positive sample higher than a random negative sample.[33] It is insensitive to the absolute predicted probabilities and only depends on their relative order.[34] Therefore, a loss function that solely focuses on making individual predictions more accurate (like BCE) may not be the most direct or effective way to maximize the final ranking score. A winning strategy must employ loss functions that directly address both the class imbalance and the ranking objective.

A Triad of Advanced Loss Functions for Top-Tier Performance

To construct a training objective that is maximally aligned with the competition's goals, a hybrid loss function composed of state-of-the-art components will be employed.

1. **Primary Choice - Asymmetric Loss (ASL):** The foundational component of our loss will be Asymmetric Loss (ASL), which will replace BCE. ASL is a recent and highly effective loss function designed specifically for multi-label classification problems with significant positive-negative imbalance.[35] Its key innovation is the decoupling of the focusing mechanism for positive and negative samples. Standard Focal Loss applies a single focusing parameter, γ, to modulate the loss for both positive and negative examples based on their confidence. In a highly imbalanced setting, a high γ is needed to sufficiently down-weight the millions of easy negatives, but this can inadvertently suppress the already-small gradients from the rare positive samples. ASL solves this by using two separate focusing parameters: γ_- for negative samples and γ_+ for positive samples. This allows for an aggressive down-weighting of easy negatives (by setting a high γ_-, e.g., 4) while maintaining a strong training signal from positives (by setting a low γ_+, e.g., 0 or 1). ASL has become the de-facto standard for achieving state-of-the-art results in this domain.[36]
2. **Direct Metric Optimization - AUC Margin Loss:** To directly optimize for the competition metric, a surrogate loss for AUC will be incorporated. While many such losses exist, a superior modern choice is the **AUC Margin Loss**.[37] This is a min-max surrogate loss function that improves upon older pairwise losses (like the AUC square loss) by being more robust to noisy data and less adversely affected by very easy examples. It is designed for large-scale deep learning and directly encourages the model to produce

scores for positive samples that are higher than those for negative samples by a certain margin, which is precisely the behavior that maximizes the AUC score.[37]

3. The Winning Combination - Hybrid Loss: The most powerful strategy is to combine these two objectives into a single, weighted hybrid loss function. The total loss for a training batch will be calculated as:
   $$ \mathcal{L}{\text{total}} = \alpha \cdot \mathcal{L}{\text{ASL}}(y_{\text{pred}}, y_{\text{true}}) + (1 - \alpha) \cdot \mathcal{L}{\text{AUC-Margin}}(y{\text{pred}}, y_{\text{true}}) $$
   This hybrid approach creates a synergistic training signal. The ASL component ($\mathcal{L}_{\text{ASL}}$) acts as a strong, point-wise regularizer, grounding the model in producing well-calibrated probabilities for each individual sample. The AUC Margin component ($\mathcal{L}_{\text{AUC-Margin}}$) acts as a pairwise or listwise regularizer, fine-tuning the model's decision boundary to explicitly optimize the ranking of scores across the entire batch. This combination leverages the strengths of both classification and ranking objectives.31 The weighting hyperparameter α can be tuned, perhaps even scheduled during training, to balance the two goals. For instance, training could start with a higher weight on ASL to establish a stable baseline and gradually increase the weight on the AUC loss to focus on metric maximization as the model converges.

Loss Function	**Mechanism**	**Key Hyperparameters**	**Pros**	**Cons**
Binary Cross-Entropy (BCE)	Point-wise log loss	None	Simple, standard baseline.	Performs poorly on imbalanced data, biased to majority class.[30]
Focal Loss	Point-wise, symmetric focusing	γ (focusing factor)	Focuses training on hard examples, better for	Symmetric focusing can suppress gradients from

			imbalance than BCE.[30]	rare positive samples.[35]
Asymmetric Loss (ASL)	Point-wise, asymmetric focusing	γ_+, γ_-, clip	Decouples focusing for +/- samples, SOTA for multi-label imbalance.[35]	Does not directly optimize the ranking metric (AUC).
AUC Margin Loss	Pairwise ranking loss	margin	Directly optimizes a surrogate for the AUC metric, robust to noise.[37]	Can have unstable gradients if used alone, not a calibration loss.

Strategic Sampling: Hard Negative Mining

Even with a sophisticated loss function, the model's performance can be further enhanced by carefully curating the data it sees during training. The set of false positives generated by Stage 1 will not be random noise; they will be **hard negatives**—anatomical structures that are visually similar to aneurysms, such as prominent vessel bifurcations, infundibula, or tortuous vessel segments. These are the examples that are most likely to confuse the classifier.

To address this, a hard negative mining strategy will be implemented during the training of the Stage 2 classifier.[40] The process is as follows:

1. Train the Stage 2 classifier for one or more initial epochs on a balanced set of positive patches and randomly sampled negative patches.
2. After this initial phase, perform inference with the current model on the entire training set of negative patches.
3. Identify the negative patches that receive the highest prediction scores—these are the "hardest" negatives, the ones the model is most confident are aneurysms but are, in fact, not.

4. In all subsequent training epochs, construct batches by continuing to use all positive samples but oversampling from this pool of identified hard negatives.

This technique forces the model to dedicate more of its learning capacity to the most challenging and ambiguous examples. It explicitly trains the model to distinguish between the fine-grained features of a true aneurysm and its closest mimics, thereby sharpening the decision boundary and significantly improving the classifier's discriminative power and precision.[13]

Data-Centric Strategies for a Decisive Competitive Edge

While model architecture and loss function design are critical, a truly decisive competitive advantage often comes from data-centric strategies. The following approaches are designed to provide the model with richer information and make it more robust than competitors who rely solely on the raw image data as provided.

Advanced 3D Data Augmentation: Simulating the Real World

Data augmentation is an essential technique for improving model generalization and preventing overfitting, particularly in medical imaging where datasets are often limited.[42] A robust augmentation strategy goes far beyond simple geometric transformations. To build a model that can perform well on a hidden test set—which will inevitably contain data from different hospitals, scanners, and acquisition protocols—it is necessary to simulate the variations and artifacts encountered in real-world clinical practice.

The proposed augmentation pipeline, to be implemented using the MONAI framework, will consist of three tiers of transformations applied on-the-fly during training [44]:

1. **Geometric Transformations:** These are the standard augmentations that teach the model spatial invariance. They include random flips along any axis, random rotations, scaling (zooming in and out), and elastic deformations. Elastic deformations are particularly valuable as they apply smooth, random distortions to the volume, mimicking the natural anatomical variability between patients.[42]
2. **Intensity Transformations:** These augmentations teach the model to be robust to variations in image contrast and noise levels. They include adding random Gaussian noise, applying Gaussian blurring, and randomly adjusting brightness and contrast.[45]

3. **Scanner Artifact Simulation:** This is the key differentiating tier of the augmentation strategy. It involves simulating common artifacts that arise during the MRI/CTA acquisition process. This is a form of domain generalization; by exposing the model to these simulated imperfections during training, it becomes more robust when it encounters them in the real, unseen test data. Key artifact simulations include:

- **Motion Artifacts:** Patient movement during a scan introduces characteristic artifacts. These can be realistically simulated by applying small, random translations and rotations to the data in k-space (the frequency domain) before applying the inverse Fourier transform to get back to the image domain.[47]
- **Pulsation Artifacts:** Particularly relevant for angiography, artifacts from pulsatile blood flow can create ghosting. These can be simulated by adding periodic errors to the k-space magnitude data along the phase-encoding direction before reconstruction.[48]
- **Signal Loss and Field Inhomogeneity:** These can be simulated by creating masks that locally reduce signal intensity or by applying a smooth, spatially varying multiplicative field to the image, respectively.[47]

This comprehensive augmentation strategy does more than just artificially increase the dataset size. It actively trains the model to be invariant to a wide range of common, non-pathological sources of variation. A model that has learned to ignore simulated motion artifacts during training will be less likely to be confused by a real scan from a slightly restless patient. This proactive engineering for robustness is a hallmark of competition-winning solutions.

Transform	**Key Parameters**	**Purpose**
monai.transforms.RandFlipd	prob, spatial_axis	Basic geometric invariance to orientation.
monai.transforms.RandRotated	prob, range_x, range_y, range_z	Invariance to patient positioning.
monai.transforms.RandAffined	prob, scale_range, shear_range	Invariance to scaling and shearing.

monai.transforms.RandElasticd	prob, sigma_range, magnitude_range	Simulate natural tissue variability and deformation.[44]
monai.transforms.RandGaussianNoised	prob, std	Simulate varying scanner signal-to-noise ratios.[45]
monai.transforms.RandAdjustContrastd	prob, gamma	Invariance to different contrast settings.
SimulateMotionArtifactd (Custom)	prob, translation_range, rotation_range	Simulate patient movement during scan via k-space manipulation.[47]
SimulatePulsationArtifactd (Custom)	prob, frequency, magnitude	Simulate blood flow ghosting artifacts via k-space manipulation.[48]

The Power of Auxiliary Inputs: Vessel Segmentation and Patient Metadata

The most innovative strategies often involve providing the model with additional, context-rich information that is not present in the raw pixel data alone. Two such strategies are proposed to give our model a significant analytical edge.

Strategy 1: Vessel Segmentation as a Structural Prior

Aneurysms are not random occurrences; they are pathologies *of* the vasculature. Providing the model with an explicit map of the blood vessels gives it a powerful anatomical prior, simplifying its learning task.

The implementation involves a multi-step process:

1. **Train a Vessel Segmentation Model:** A separate, lightweight 3D U-Net will be trained specifically for the task of cerebrovascular segmentation. Publicly available datasets with vessel annotations can be used for this, or it can be trained on the competition data if segmentation labels are available or can be semi-automatically generated. The goal is to create a model that can produce a reasonably accurate binary mask of the arteries in any given scan.[49]
2. **Create a Multi-Channel Input:** The Stage 2 classifier's input will be modified. Instead of a single-channel 3D patch, it will receive a **two-channel 3D patch**. Channel 1 will contain the original image intensities from the CTA or MRA scan. Channel 2 will contain the corresponding binary vessel segmentation mask generated in the previous step.[51]

This approach fundamentally reframes the learning problem for the Stage 2 classifier. Without the mask, the model must learn two things simultaneously: "Where are the blood vessels?" and "Is there an anomalous outpouching on one of those vessels?". By providing the vessel mask as a second channel, we are effectively giving the model the answer to the first question. Its task is simplified to: "Given the location of the vessels, identify regions with abnormal morphology." This allows the model to dedicate its full parametric capacity to the more subtle and critical task of classifying aneurysm morphology, which should lead to higher accuracy and better discrimination from vessel-like mimics.

Strategy 2: Fusing Patient Metadata for Enhanced Prediction

Clinical parameters such as patient age and sex are known to be associated with the prevalence and rupture risk of intracranial aneurysms.[18] This non-imaging data provides a valuable and orthogonal source of information that can improve predictive performance.

The most effective way to integrate this tabular metadata with the 3D imaging data is through a **late fusion** architecture [54]:

1. **Image Feature Extraction:** The 3D image patch (containing the raw image and optionally the vessel mask) is processed by the powerful, pre-trained Swin UNETR encoder. The output of the encoder is a high-dimensional feature vector that represents the learned visual features of the patch.
2. **Metadata Embedding:** The tabular patient data (e.g., a vector containing numerical values for age and an encoded value for sex) is processed by a small, separate Multi-Layer Perceptron (MLP). This MLP learns to transform the raw clinical data into a dense, meaningful embedding vector.

3. **Fusion via Concatenation:** The image feature vector from the Swin UNETR and the metadata embedding vector from the MLP are concatenated into a single, longer feature vector.
4. **Final Classification:** This combined vector, which now contains rich information from both modalities, is passed through a final set of fully connected layers that produce the ultimate probability score for the presence of an aneurysm.

This late fusion approach is superior to simpler methods like early fusion (e.g., tiling the metadata values into an extra image channel). Early fusion forces the initial convolutional or attention layers to learn relationships between fundamentally different data types (pixel intensities and clinical numbers), which is an inefficient and difficult task. Late fusion allows each modality to be processed by a specialized encoder best suited for its structure (a 3D Transformer for images, an MLP for tabular data). The fusion occurs at a high level of semantic abstraction, combining "what the model sees in the image" with "what the model knows about the patient" just before the final decision is made. This is a more robust, principled, and typically higher-performing method for multi-modal learning.[54]

An Implementation Blueprint and Final Ensemble Strategy

A winning strategy requires not only advanced conceptual components but also a robust and reproducible implementation plan. This section outlines the choice of framework and the methodology for constructing the final, high-performance ensemble submission.

A Principled Workflow with MONAI and the nnU-Net Philosophy

To translate the preceding strategies into a functional pipeline, the **MONAI (Medical Open Network for AI)** framework will be used as the primary development environment.[56] MONAI is a PyTorch-based, open-source ecosystem designed specifically for deep learning in medical imaging. Its adoption provides several key advantages:

- **Domain-Specific Tools:** MONAI offers a rich library of pre-built components tailored for medical imaging, including loaders for NIfTI and DICOM formats, an extensive set of 3D-aware data augmentation transforms, and implementations of state-of-the-art network architectures like 3D U-Net and Swin UNETR.[57]

- **Specialized Components:** It includes implementations of medical-centric loss functions (e.g., DiceLoss, TverskyLoss) and evaluation metrics, which are essential for building and validating models in this domain.[57]
- **Reproducibility:** MONAI's "Bundle" system allows for the packaging of models, data, and configurations, promoting reproducible research and making it easier to manage complex experiments.[58]

While a custom pipeline will be built, the overarching workflow will be guided by the principles of the highly successful **nnU-Net** framework.[59] The nnU-Net has dominated numerous medical segmentation challenges by automating the process of configuring and training U-Net models. Adopting its core philosophy brings a level of rigor and robustness that is critical for competitive success:

- **Systematic Cross-Validation:** All models will be trained and evaluated using a strict k-fold cross-validation scheme (e.g., 5-fold or 10-fold). This is non-negotiable. It provides a much more reliable estimate of the model's generalization performance than a single train/validation split and is the foundation for effective ensembling.
- **Automated Configuration and Preprocessing:** The nnU-Net's strength lies in its ability to automatically determine optimal parameters like patch size, batch size, and normalization schemes based on the dataset's properties (e.g., voxel spacing, image dimensions) and the available hardware (GPU memory).[59] This logic will be scripted to ensure that the pipeline is optimally configured for the competition data and the training environment.
- **Post-Inference Ensembling:** A core feature of nnU-Net is the automatic ensembling of the models trained on each of the k cross-validation folds for final prediction. This simple averaging of outputs from models trained on different subsets of the data is a powerful technique for reducing variance and improving performance. This will be the default inference strategy.

Constructing the Winning Ensemble: The Power of Diversity

The final step to maximize performance on the private leaderboard is to construct a powerful ensemble. The history of Kaggle competitions is clear: ensembles of diverse models consistently outperform any single model.[2] The key to a successful ensemble is

diversity. Averaging the predictions of five identical models trained with different random seeds yields marginal gains. Averaging the predictions of five models that have learned different aspects of the problem space can yield substantial improvements.

The proposed strategy is to build an ensemble from the most promising and diverse models developed throughout this project. The final submission will be a weighted average of the predictions from the following models, each trained across all k cross-validation folds:

1. **Model 1 (The Core Model):** The self-supervised pre-trained Swin UNETR classifier, fine-tuned on 3D patches using the hybrid ASL + AUC Margin loss. This is the primary, state-of-the-art workhorse.
2. **Model 2 (The Structurally-Aware Model):** The Swin UNETR classifier trained with the two-channel input: the raw image patch and the corresponding vessel segmentation mask. This model has access to explicit anatomical context that Model 1 does not.
3. **Model 3 (The Clinically-Informed Model):** The Swin UNETR classifier with the late-fusion architecture that integrates patient metadata (age, sex). This model incorporates non-imaging clinical risk factors.
4. **Model 4 (The Architecturally-Diverse Model - Optional):** To introduce maximum architectural diversity, a completely different high-performance 3D CNN could be trained as the Stage 2 classifier. A modern variant of a ResNet or DenseNet, adapted for 3D inputs, would be an excellent choice.[21] This model would learn different feature representations from the Transformer-based models, making its errors less likely to be correlated with theirs.

The predictions from each of these model families (each already an ensemble of its k-fold versions) will be combined using a simple weighted average. The weights can be optimized on a held-out validation set to maximize the final AUC score. This multi-faceted ensemble, built on models with diverse inputs, loss functions, and potentially architectures, represents the most robust and highest-potential strategy for achieving a winning rank.

Concluding Strategic Recommendations

This report outlines a comprehensive, multi-faceted strategy for the RSNA 2025 Intracranial Aneurysm Detection Challenge. The approach is grounded in the analysis of past competitions, built upon state-of-the-art architectures, and enhanced with innovative data-centric techniques. The success of this endeavor hinges on the disciplined execution of five key strategic pillars:

- **Pillar 1: Adopt the Two-Stage Pipeline.** The problem must be decomposed into two manageable stages: a high-recall candidate generation stage to ensure no aneurysms are missed, followed by a high-precision classification stage to eliminate false positives. This proven paradigm from analogous challenges is the most effective way to tackle the small object detection problem in large volumetric scans.[7]
- **Pillar 2: Build on a Pre-Trained Swin UNETR.** The core classifier must be a state-of-the-art 3D Vision Transformer. The Swin UNETR architecture, with its hybrid Transformer-CNN design, is the ideal choice.[24] Its data requirements will be met and its performance enhanced through a rigorous self-supervised pre-training phase on a large corpus of unlabeled head scans, creating a powerful foundational model of neurovascular anatomy.
- **Pillar 3: Optimize Directly for the Metric.** Standard training objectives are insufficient. The training process must be explicitly tailored to the competition's weighted AUC metric and the extreme class imbalance of the data. This will be achieved by using a sophisticated hybrid loss function that combines the imbalance-handling capabilities of Asymmetric Loss (ASL) with the direct ranking optimization of an AUC Margin Loss [35], further sharpened by a hard negative mining strategy.
- **Pillar 4: Win with Data-Centric Innovation.** A decisive competitive edge will be gained by moving beyond the provided image data. This includes implementing an advanced 3D data augmentation pipeline that simulates realistic scanner artifacts to build a more robust model [47], and engineering models that leverage auxiliary inputs. Providing the classifier with a vessel segmentation mask as an anatomical prior and fusing patient metadata will give it access to contextual information unavailable to competing models.
- **Pillar 5: Engineer for Robustness and Diversity.** The entire workflow must be built with rigor and reproducibility using the MONAI framework and adhering to the principles of the nnU-Net (automated cross-validation and configuration).[57] The final submission must be a powerful ensemble that derives its strength from diversity—combining models trained with different inputs, different objectives, and potentially different underlying architectures to produce a final prediction that is more accurate and robust than any of its individual components.

Works cited

1. RSNA Intracranial Hemorrhage Detection - Kaggle, accessed August 10, 2025, https://www.kaggle.com/competitions/rsna-intracranial-hemorrhage-detection

2. An Effective Transformer-based Solution for RSNA Intracranial Hemorrhage Detection Competition - Proceedings of Machine Learning Research, accessed August 10, 2025, https://proceedings.mlr.press/v281/shang25a.html
3. Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images - arXiv, accessed August 10, 2025, https://arxiv.org/abs/2201.01266
4. Vision Transformers vs. Convolutional Neural Networks | by Fahim Rustamy, PhD | Medium, accessed August 10, 2025, https://medium.com/@faheemrustamy/vision-transformers-vs-convolutional-neural-networks-5fe8f9e18efc
5. Evaluation - LUNA16 - Grand Challenge, accessed August 10, 2025, https://luna16.grand-challenge.org/Evaluation/
6. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge | Request PDF - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/publication/311900928_Validation_comparison_and_combination_of_algorithms_for_automatic_detection_of_pulmonary_nodules_in_computed_tomography_images_The_LUNA16_challenge
7. arXiv:1612.08012v4 [cs.CV] 15 Jul 2017, accessed August 10, 2025, https://arxiv.org/pdf/1612.08012
8. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge - Vrije Universiteit Brussel, accessed August 10, 2025, https://researchportal.vub.be/en/publications/validation-comparison-and-combination-of-algorithms-for-automatic
9. gzuidhof/luna16: LUNA16 Lung Nodule Analysis - NWI-IMC037 Final Project - GitHub, accessed August 10, 2025, https://github.com/gzuidhof/luna16
10. How Two-Stage Detector Machine Vision Systems Improve Accuracy - UnitX, accessed August 10, 2025, https://www.unitxlabs.com/resources/two-stage-detector-machine-vision-system-accuracy-benefits/
11. Single Stage Detector vs 2 Stage Detector | by Abhishek Jain - Medium, accessed August 10, 2025, https://medium.com/@abhishekjainindore24/single-stage-detector-vs-2-stage-detector-3e540ea81213

12. How Two-stage Object Detection Enhances Machine Vision Applications - UnitX, accessed August 10, 2025, https://www.unitxlabs.com/resources/two-stage-object-detection-machine-vision/

13. Very quick 1st summary of julian's part of 2nd place solution. - Kaggle, accessed August 10, 2025, https://www.kaggle.com/c/data-science-bowl-2017/discussion/31551

14. 2017 Data Science Bowl, Predicting Lung Cancer: 2nd Place Solution Write-up, Daniel Hammack and Julian de Wit | by Kaggle Team - Medium, accessed August 10, 2025, https://medium.com/kaggle-blog/2017-data-science-bowl-predicting-lung-cancer-2nd-place-solution-write-up-daniel-hammack-and-79dc345d4541

15. 2nd place solution for the 2017 national datascience bowl - Julian de Wit, accessed August 10, 2025, https://juliandewit.github.io/kaggle-ndsb2017/

16. Enhancing Intracranial Aneurysm Detection with Artificial ..., accessed August 10, 2025, https://www.jneurology.com/articles/enhancing-intracranial-aneurysm-detection-with-artificial-intelligence-in-radiology.html

17. A deep-learning model for intracranial aneurysm detection on CT ..., accessed August 10, 2025, https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00268-6/fulltext

18. Prediction of cerebral aneurysm rupture risk by machine learning ..., accessed August 10, 2025, https://pubmed.ncbi.nlm.nih.gov/38183490/

19. DeepMedic, accessed August 10, 2025, https://deepmedic.org/

20. DeepMedic for Brain Tumor Segmentation - Christian Ledig, accessed August 10, 2025, http://www.christianledig.com/Publications/pdf/kamnitsas2016brats.pdf

21. Block designs for U-Net (left), ResNet (center), and DenseNet (right). - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/figure/Block-designs-for-U-Net-left-ResNet-center-and-DenseNet-right_fig2_365760325

22. How to Choose a Neural Net Architecture for Medical Image Segmentation - Innolitics, accessed August 10, 2025, https://innolitics.com/articles/medical-image-segmentation-overview/

23. Leveraging 3D Convolutional Neural Networks for Accurate Recognition a | TCRM, accessed August 10, 2025, https://www.dovepress.com/leveraging-3d-convolutional-neural-networks-for-accurate-recognition-a-peer-reviewed-fulltext-article-TCRM

24. Review — UNETR: Transformers for 3D Medical Image Segmentation | by Sik-Ho Tsang, accessed August 10, 2025, https://sh-tsang.medium.com/review-unetr-transformers-for-3d-medical-image-segmentation-913f497dc90c
25. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis - CVF Open Access, accessed August 10, 2025, https://openaccess.thecvf.com/content/CVPR2022/papers/Tang_Self-Supervised_Pre-Training_of_Swin_Transformers_for_3D_Medical_Image_Analysis_CVPR_2022_paper.pdf
26. SwinUNETR-V2: Stronger Swin Transformers with Stagewise Convolutions for 3D Medical Image Segmentation | MICCAI 2023 - Accepted Papers, Reviews, Author Feedback, accessed August 10, 2025, https://conferences.miccai.org/2023/papers/634-Paper1667.html
27. SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation - arXiv, accessed August 10, 2025, https://arxiv.org/html/2404.10156v2
28. AWS Marketplace: RSNA Intracranial Hemorrhage Detection - Amazon.com, accessed August 10, 2025, https://aws.amazon.com/marketplace/pp/prodview-hofnashho7th2
29. How to build the best medical image segmentation algorithm using foundation models: a comprehensive empirical study with Segment Anything Model - Melba Journal, accessed August 10, 2025, https://www.melba-journal.org/papers/2025:006.html
30. Novel loss functions for ensemble-based medical image classification | PLOS One, accessed August 10, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261307
31. Imbalance-aware loss functions improve medical image classification - Proceedings of Machine Learning Research, accessed August 10, 2025, https://proceedings.mlr.press/v250/scholz24a.html
32. Imbalanced Dataset: Strategies to Fix Skewed Class Distributions - Label Your Data, accessed August 10, 2025, https://labelyourdata.com/articles/imbalanced-dataset
33. AUC Maximization in the Era of Big Data and AI: A Survey - arXiv, accessed August 10, 2025, https://arxiv.org/pdf/2203.15046
34. Has anyone successfully implemented AUROC as a loss function for Theano/Lasagne/Keras? : r/MachineLearning - Reddit, accessed August 10, 2025, https://www.reddit.com/r/MachineLearning/comments/3zksod/has_anyone_successfully_implemented_auroc_as_a/

35. Asymmetric Loss for Multi-Label Classification - CVF Open Access, accessed August 10, 2025, https://openaccess.thecvf.com/content/ICCV2021/papers/Ridnik_Asymmetric_Loss_for_Multi-Label_Classification_ICCV_2021_paper.pdf

36. Alibaba-MIIL/ASL: Official Pytorch Implementation of: "Asymmetric Loss For Multi-Label Classification"(ICCV, 2021) paper - GitHub, accessed August 10, 2025, https://github.com/Alibaba-MIIL/ASL

37. Large-Scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification - CVF Open Access, accessed August 10, 2025, https://openaccess.thecvf.com/content/ICCV2021/papers/Yuan_Large-Scale_Robust_Deep_AUC_Maximization_A_New_Surrogate_Loss_and_ICCV_2021_paper.pdf

38. L1-penalized AUC-optimization with a surrogate loss, accessed August 10, 2025, http://www.csam.or.kr/journal/view.html?doi=10.29220/CSAM.2024.31.2.203

39. Imbalance-aware loss functions improve medical image classification - OpenReview, accessed August 10, 2025, https://openreview.net/forum?id=5Oiqw76ube

40. A Hard Negatives Mining and Enhancing Method for Multi-Modal Contrastive Learning, accessed August 10, 2025, https://www.mdpi.com/2079-9292/14/4/767

41. Unsupervised Contrastive Learning of Image Representations from Ultrasound Videos with Hard Negative Mining | MICCAI 2022 - Accepted Papers and Reviews, accessed August 10, 2025, https://conferences.miccai.org/2022/papers/538-Paper1638.html

42. What is the best data augmentation for 3D brain tumor segmentation? - DiVA portal, accessed August 10, 2025, https://www.diva-portal.org/smash/get/diva2:1588376/FULLTEXT01.pdf

43. 3D Brain MRI Classification for Alzheimer's Diagnosis Using CNN with Data Augmentation, accessed August 10, 2025, https://arxiv.org/html/2505.04097v1

44. How is 3D data augmentation applied? - Milvus, accessed August 10, 2025, https://milvus.io/ai-quick-reference/how-is-3d-data-augmentation-applied

45. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks - PMC, accessed August 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5977656/

46. How is data augmentation used in medical imaging? - Milvus, accessed August 10, 2025, https://milvus.io/ai-quick-reference/how-is-data-augmentation-used-in-medical-imaging

47. Simulated MRI Artifacts: Testing Machine Learning Failure Modes - PMC, accessed August 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10521705/

48. (PDF) Artifact-robust Deep Learning-based Segmentation of 3D Phase-contrast MR Angiography: A Novel Data Augmentation Approach - ResearchGate, accessed August 10, 2025, https://www.researchgate.net/publication/393413872_Artifact-robust_Deep_Learning-based_Segmentation_of_3D_Phase-contrast_MR_Angiography_A_Novel_Data_Augmentation_Approach

49. 3D vessel segmentation - DLMA: Deep Learning for Medical Applications - BayernCollab, accessed August 10, 2025, https://collab.dvb.bayern/spaces/TUMdlma/pages/73379955/3D+vessel+segmentation

50. tUbe net: a generalisable deep learning tool for 3D vessel segmentation - bioRxiv, accessed August 10, 2025, https://www.biorxiv.org/content/10.1101/2023.07.24.550334.full

51. Cerebrovascular Segmentation Model Based on Spatial Attention-Guided 3D Inception U-Net with Multi-Directional MIPs - MDPI, accessed August 10, 2025, https://www.mdpi.com/2076-3417/12/5/2288

52. Brain age estimation based on 3D MRI images using 3D convolutional neural network, accessed August 10, 2025, https://www.researchgate.net/publication/342598914_Brain_age_estimation_based_on_3D_MRI_images_using_3D_convolutional_neural_network

53. Using Advanced Convolutional Neural Network Approaches to Reveal Patient Age, Gender, and Weight Based on Tongue Images, accessed August 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11309814/

54. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review - PubMed Central, accessed August 10, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12239537/

55. Predicting brain-age from raw T1-weighted Magnetic Resonance Imaging data using 3D Convolutional Neural Networks - arXiv, accessed August 10, 2025, https://arxiv.org/pdf/2103.11695

56. Guide on 3D Medical Image Segmentation with Monai & UNET - Analytics Vidhya, accessed August 10, 2025, https://www.analyticsvidhya.com/blog/2024/03/guide-on-3d-medical-image-segmentation-with-monai-unet/

57. MONAI: The Definitive Framework for Medical Imaging Powered by PyTorch - LearnOpenCV, accessed August 10, 2025, https://learnopencv.com/monai-medical-imaging-pytorch/

58. AI Maker Case Study - MONAI 1.0 Train 3D Segmentation Model Using CT Data | TWS, accessed August 10, 2025, https://docs.twcc.ai/en/docs/concepts-tutorials/twcc/oneai/tutorials/aimaker-public-template/computer-vision/monai1.0/

59. MIC-DKFZ/nnUNet - GitHub, accessed August 10, 2025, https://github.com/MIC-DKFZ/nnUNet

60. nnU-Net - Helmholtz Imaging CONNECT, accessed August 10, 2025, https://connect.helmholtz-imaging.de/solution/3

61. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation - NMMItools, accessed August 10, 2025, https://nmmitools.org/2024/01/01/nnu-net-a-self-configuring-method-for-deep-learning-based-biomedical-image-segmentation/