Final Project Report

Veronica Patterson and Caleb Skinner

November 2025

1 Introduction

Intracranial aneurysms are a bulge on the artery wall of the brain, affecting approximately 3% of the human population, with nearly 30% of diagnosed patients presenting with multiple aneurysms [Keedy, 2006]. Most unruptured aneurysms are asymptomatic; however, rupture leads to subarachnoid hemorrhage, a condition with high mortality, emphasizing the clinical importance for early detection [Williams and Brown, 2013].

Detecting intracranial aneurysms by eye is challenging, even for experienced physicians. The main difficultly with manual detection is attributed to the size of the aneurysms. Kim et al. [2021] found that among 105 unruptured intracranial aneurysms detected in asymptomatic adults, the mean size was 3.10 ± 1.62 mm, making them difficult to distinguish from surrounding vascular structure in volumetric imaging. These subtle features increase the risk of oversight, especially under high clinical reading workloads [Pesapane et al., 2024].

Computer-aided detection (CADe) methods have been increasingly explored to support radiologists. In particular, deep learning approaches have demonstrated improvements in reader sensitivity. For example, Pennig et al. [2021] reported that a deep learning model, three dimensional Convolutional Neural Networks (3D CNN), improved radiologist sensitivity for aneurysm detections by 9% on average. Motivated by these findings, this study proposes a deep learning pipeline for aneurysm detection and spatial localization using multimodal brain imaging from the RSNA Intercranial Aneurysm Challenge dataset.

Our specific objective is twofold. First, we aim to determine if the brain scan contains an aneurysm. This binary classification problem is clinically crucial as it enables timely treatment planning and possible avoidance of rupture. Second, conditional on aneurysm being present, we seek to determine the anatomical location(s) of the brain aneurysm(s). The RSNA dataset provides 13 possible sites corresponding to specific intracranial arterial regions.

The first objective is simpler and paramount, while identifying the location of the aneurysm is more challenging, yet highly impactful. Accurate spatial localization can increase clinical trust in CADe since the model provides specific arterial segments to review rather than only a binary decision. Moreover, localization directly influences downstream clinical decisions, specifically in terms of treatment plan and/or surgical removal.

Inspired by the work of Pennig et al. [2021], we deploy a 3D CNN to address these objectives. We implement two architectures, referred to as **one-step** and **two-step**, which differ in whether they unite (one-step) or decouple (two-step) the tasks of aneurysm detection and spatial localization.

This report is divided into 5 sections. In Section 2, we detail the extensive data preprocessing steps. Section 3 describes the Convolutional Neural Network and its training regimen. Lastly, results are found in Section 4 and the discussion in Section 5.

2 Data Description and Preprocessing

We acquire data from a Kaggle competition hosted by the Radiological Society of North America (RSNA) together with the American Society of Neuroradiology, the Society of Neurointerventional Surgery, and the European Society of Neuroradiology [Rudie et al., 2025]. Each subject contributes one brain scan, stored in Digital Imaging and Communications in Medicine (DICOM) format. Labels indicate whether the scan contains an intracranial aneurysm and, if present, which of 13 arterial regions are involved (see Appendix A.1).

To understand the characteristics of the dataset prior to modeling, we perform exploratory data analysis (EDA). The dataset consists of 4,348 subjects, with imaging obtained from two primary modalities: 41.6% Computed Tomography Angiography (CTA) and 58.4% Magnetic Resonance Imaging (MRI). Aneurysms are present in 1,863 subjects (42.8%), while the remaining 2,485 subjects exhibit no aneurysm.

Aneurysm occurrences among the 13 arterial regions are highly imbalanced. The most frequently involved sites are the Anterior Communicating Artery (363 cases), Left Supraclinoid Internal Carotid Artery (330 cases), and Right Middle Cerebral Artery (294 cases). In contrast, far fewer aneurysms occur in the Right Anterior Cerebral Artery (56 cases) and Left Anterior Cerebral Artery (46 cases), each representing only a small fraction of the dataset. Because a single subject can exhibit multiple aneurysms, multiple arterial regions may be labeled positive for the same scan.

The number of slices per scan ranges from 1 to 1,441 slices, reinforcing the need for consistent volume formation when building 3D models. Visual summaries of modality distribution, aneurysm prevalence, and regional imbalance are provided in Appendix A.2. Together, these trends motivate the choice of intensity normalization and mini-volume construction, described in detail below.

2.1 Preprocessing Pipeline and Model Specific Input Construction

This work evaluates two related modeling strategies, both predicting (1) whether an aneurysm is present and (2) its arterial location if present. The strategies differ only in how inputs are constructed and labels are assigned.

The **one-step model** performs global detection and localization simultaneously. All slices for each scan are retained and later partitioned into multiple 32-slice mini-volumes. When a slice is multi-frame or RGB, frames or channels are averaged. The resulting 32-slice mini-volumes are each paired with the subject-level aneurysm label and the 13 regional labels.

The **two-step model** decouples detections and localization. All slices for each scan are retained and later partitioned into multiple 32-slice mini-volumes. Stage one is a binary classifier: every mini-volume from a subject with an aneurysm receives a positive global label, exactly as in the one-step model. Stage two localizes aneurysms, but only using mini-volumes that are guaranteed to contain an aneurysm. To achieve this, we map the dataset's slice-level annotations (each specifying a precise DICOM instance and arterial region) to the ordered slices. A mini-volume is labeled positive only if at least one of its slices corresponds to an annotated aneurysm slice. This reduces label noise by eliminating positive mini-volumes that may not intersect the aneurysm in 3D space.

2.1.1 Slice-Level Preprocessing

Each scan contains a variable number of 2D slices represented by pixel intensities, either in grayscale or in full color, the latter adding an additional channel dimension for red, green, and blue color scales. When properly ordered and stacked, these slices form a 3D brain volume composed of voxels. The scans originate from CTA, which uses X-rays to measure tissue intensity, and MRI, which uses magnetic fields and radio waves to capture tissue-dependent signal intensities [Center, 2022]. Because CTA and MRI produce different brightness and contrast patterns, preprocessing must enforce consistent slice ordering and intensity normalization across all subjects.

To begin, we order slices using metadata from the DICOM files, specifically the `ImagePositionPatient` (IPP) and `ImageOrientationPatient` (IOP) fields. Atkinson [2022] provides a detailed explanation of how these fields define the spatial geometry of medical images. IPP represents the 3D coordinates of the center of the top-left pixel, and IOP specifies two unit vectors defining the image row and column directions. We compute the cross product of these vectors to obtain the slice normal (the positive through-slice direction) and use the dot product between the slice normal and IPP to order slices along that direction [Atkinson, 2022].

After obtaining the slice set appropriate for each model (one-step or two-step), we preprocess each slice individually. First, we resize all slices to a uniform image size of 256×256 pixels, ensuring consistent height and width across slices and subjects. Next, we apply the intensity scaling parameters stored in the DICOM metadata to adjust pixel values. This step applies only to CTA-based scans, for which pixel values are converted to Hounsfield units (HU); MRI intensities are left in their native scanner units. Before clipping and normalization, we ensure that slices labeled as inverted grayscale are corrected so that higher pixel values correspond to brighter intensities.

Lastly, we clip each slice using modality-appropriate windowing levels and normalize the intensities to the range $[0, 1]$. To satisfy memory constraints, we then reduce the image size of each slice to 128×128 pixels using a max pooling layer. This preprocessing pipeline yields 3,726 subjects that make up a balanced training set, balanced by global aneurysm presence rather than by localized arterial labels.

3   Model

To perform multi-label classification of the brain scan images, we employ a 3D CNN. CNNs are a type of neural network that learn features through a series of convolutional filters and pooling layers. These networks are specifically engineered for image and video detection tasks.

While each slice of a brain scan is a two dimensional image, stacking the entire series creates a three dimensional depiction of the brain. Within the CNN framework, there are two distinct approaches to leverage this third dimension. First, one could coerce the series of images into different channels of a 2D CNN similar to how the RGB components of a colorful image are divided into three channels of a 2D CNN. This approach is more computationally feasible, but it does not allow the model to learn patterns across the third dimension. Alternatively, 3D CNN incorporates the brain's height like any other dimension, equipping it to learn patterns across the third dimension. This comes at a steep computational and memory cost, so we take creative steps to conserve these resources. We adopted the 2D CNN for our initial submissions to Kaggle, but we updated our model to a 3D CNN for this report.

3.1   One-Step Approach

The architecture of our convolutional neural network relies on a popular methodology: networks using blocks. This concept builds on the pioneering work of Simonyan and Zisserman [2014], who organized CNN into a sequence of VGG blocks. Our proposed CNN is composed of three blocks and three feedforward layers. Each block consists of two sets of the following: convolutional layer, a batch normalization layer [Ioffe and Szegedy, 2015], and leaky ReLU activation layer Maas et al. [2013]. The block is completed by a single max-pooling layer. The convolutional layers are all $3 \times 3 \times 3$ filters with a stride of 1 and padding of 1. Each convolutional layer doubles the number of channels in the preceding layer. A block is visualized in Figure 1.

The convolutional blocks and feedforward layers are bridged by a single global average pooling layer to flatten the 4 dimensional convolutional output into a vector that can be more easily managed by the feedforward layers. The feedforward layers are each separated by a Leaky ReLU activation layer with a dropout probability of 0.5 [Hinton et al., 2012]. The output layer is a 14×1 vector. The probability of each class can be inferred by applying the sigmoid function to this output layer.

As the images are fed through the network, the convolutional layers produce increasingly many channels, but the length, width, and depth of the images are slowly reduced by the max-pooling. Eventually, the channels are averaged in a global average pooling layer and flattened to satisfy the dimensionality requirements of the feedforward layers. The net effect is to gradually funnel the information towards the smaller output layer of length 14. Even still, the CNN is a large model. The data set begins with dimensions $32 \times 128 \times 128$, and after the three batches, it is $128 \times 4 \times 16 \times 16$. This amounts to roughly 290,000 parameters in the three convolutional layers and 67,000 parameters in the three feedforward layers.
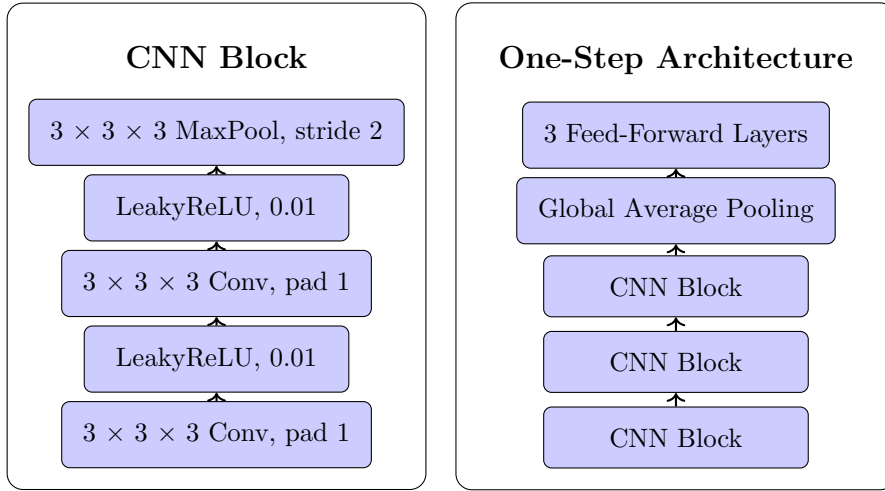

Figure 1: One-Step Approach Architecture

We train the model using backpropagation with a weighted multi-label classification cross entropy loss function. This weighted cross entropy loss function mirrors the weighted AUC metric used to evaluate the results, while maintaining computational feasibility. The weighted loss function assigns half of the loss to the detection of an aneurysm and half of the loss to the 13 arterial location predictions.

3.2   Two-Step Approach

One of the primary limitations of the one-step approach is its underemphasis on localizing the aneurysm. Since some particular locations appear infrequently in the training data, it is challenging for the model to sufficiently learn the important features. To mitigate these challenges, we propose a two-step approach as an alternative to the preceding single model. This new approach identifies the presence of the aneurysm, and then locates the aneurysm in two models. This division permits the second model to train on a more balanced data set

and focus specifically on the challenging task of locating the aneurysm. In the one-step approach, the weighted cross-entropy loss encourages the model to focus on improvements to identifying an aneurysm. Splitting the model into two steps, divides the loss function into two manageable functions.

Both the first and second steps are 3D CNN following the same procedure detailed before, but the three feedforward layers are adapted to a single output and location vector, respectively. We leverage this similar structure by initializing the convolutional parameters of the second step with the trained values of the first step and freezing all but the last convolutional layer during training. This permits the second model to utilize the underlying features learned in the first model without retraining.

Step 1: Identification

FC (1)

FC (32)

Dropout (0.5)

FC (64)

3 Conv Blocks

Step 2: Localization

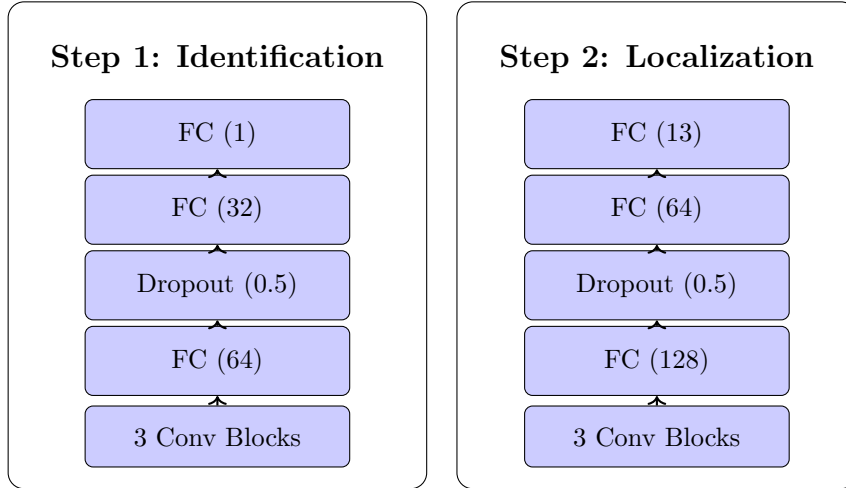FC (13)

FC (64)

Dropout (0.5)

FC (128)

3 Conv Blocks

Figure 2: Two-Step Approach Architecture

In sum, the one-step and two-step approaches have different strengths that will be evaluated in Section 4. The one-step approach requires one model and one training period, while the two-step approach requiring two, roughly doubling the computational and memory burden. However, the two-step approach offers a stronger chance of recovering the aneurysm's location.

4 Results

We perform a 70-10-20 data split across the subjects into train, validation, and test sets. Each subject is split into several mini-volumes, multiplying the number observations of each image and dramatically increasing the overall observations on which our models are trained. In total, we have roughly 18,000 mini-volumes in our training set, 2500 in our validation set, and 5000 in our test set. While the one-step and two-step approach are similar in structure, they incur different computational burdens, so we adjust the model tuning parameters accordingly.
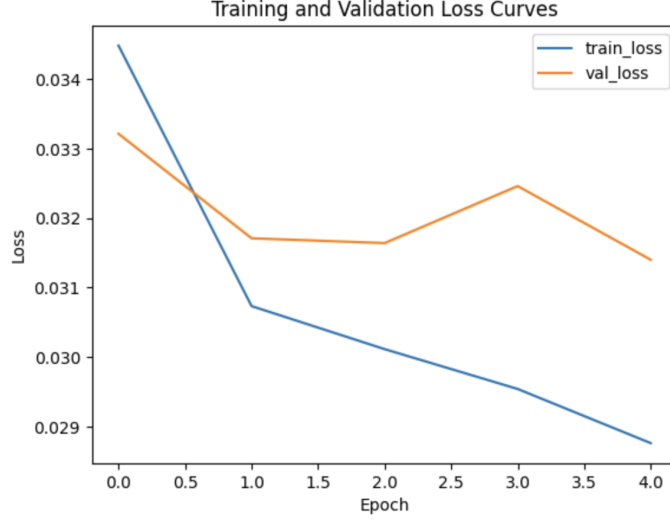
Figure 3: One-Step Approach: Training-Validation Loss Curves

The training and validation loss over 5 epochs with learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. We select the highest performing model in terms of the validation set.

To expedite stochastic gradient descent, we divide the data into batches of 64 observations for the single step approach and 32 observations for the two-step approach. Over several training sessions, we toggle the learning rate, weight decay, and number of epochs. For both models, we find that a learning rate of 1×10^{-4} and weight decay of 5×10^{-5} yield the best results.

As the model trains, we record the training and validation loss at each epoch. If the validation loss improves upon the best model, we update the best model and save the parameters. We document the training and validation loss curves of the single model approach in Figure 3 and the two-step approach in Figure 4.

The effectiveness of the model is evaluated on a test set using weighed AUC. For each subject, we aggregate the results from all of its mini-volumes, capturing information from across the brain scan. For a subject with a brain aneurysm, it is likely that some mini-volumes do possess a slice clearly containing the brain aneurysm. For that reason, we want to be sensitive to mini-volumes that report strong probabilities of the presence of a brain aneurysm or a specific location. Thus, we follow a simple decision rule for aggregating the predicted probabilities p_j for the subject's J mini-volumes:

$$p_{\text{agg}} = \begin{cases} p_{(J)} & \text{if } p_{(J)} > t \\ \bar{p} & \text{otherwise,} \end{cases}$$

where $p_{(J)} = \max_{j=1,\ldots,J}(p_j)$, $\bar{p} = \frac{1}{J}\sum_{j=1}^{J} p_j$, and $t \in (0,1)$ is some threshold.
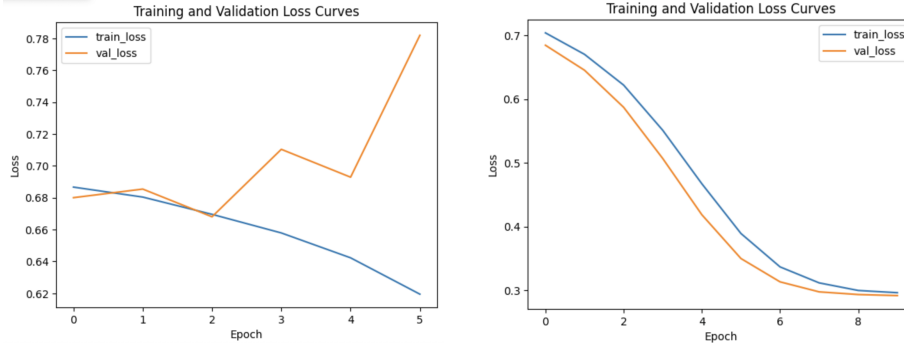
7

Figure 4: Two-Step Approach: Training-Validation Loss Curves

The training and validation loss over 6 and 10 epochs, respectively, with learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. We select the highest performing model in terms of the validation set. The step 1 model begins to overfit in epoch 3. The proximity of the training and validation losses in the step 2 model may indicate an under-parameterized model.

Intuitively, if a single mini-volume catches the aneurysm and reports a large probability, then the aggregate probability remains high in spite of the neighboring mini-volumes. We typically set $t = 0.5$ for the presence of aneurysm and $t = 0.2$ for the location of the aneurysm. From here, it is straightforward to evaluate the single model on the test set, but the two-step approach requires deeper consideration. In this case, we apply the out-of-sample test set to the first model. If an aneurysm is identified, we proceed to locate the aneurysm in the second model. This two-step evaluation leverages context-specific information that is otherwise missed in the single model: the aneurysm's location is conditional on the presence of aneurysm.

After this, the AUC is computed using the aggregated predictions for all fourteen classes, and the results are combined with the aneurysm present class designated half the weight. On our out-of-sample test set, we achieve a weighted AUC of 0.6317 for the single model approach and 0.6224 for the two-step approach. This is comparable to the official Kaggle competition, where we scored a weighted AUC of 0.6299. Despite our tireless efforts after the Kaggle submission deadline, our 2D CNN from October 14 is still competitive with our top current model in terms of out-of-sample weighted AUC (Table 1).

It is clear that the model provides mediocre detection of the aneurysm, but essentially no meaningful inference on its' location. For example, in the two-step approach, the F1-Score of the aneurysm detection class is 0.68, while the F1-Score of the 13 location classes is 0.00. See Appendix Tables 3 and 4 for the full results across all output classes.

8

Date	Submission	Model	Weighted AUC
Oct 7	1	Logistic Regression	0.5074
Oct 10	2	Logistic Regression	0.5532
Oct 14	3	2D CNN	0.6218
Oct 14	4	2D CNN	0.5560
Oct 14	5	2D CNN	0.6299

Table 1: Successful Submissions to Kaggle Competition
We report the weighted AUC on from all successful submissions to the Kaggle Competition.

5 Conclusion

In this paper, we develop a deep three dimensional convolutional neural network to identify and locate brain aneurysms from brain scans. The model proves to be helpful for identifying the presence of brain aneurysms, but useless for locating them. An effective model would prove invaluable for expediting and assisting the decisions of medical professionals. Without these machine learning tools, the cumbersome task of reviewing large swaths of brain imaging scans would fall to medical technicians. An effective prediction would at least give them a strong starting point and reduce missed aneurysms. Clearly, this is an incredibly complex problem, so it is not surprising that our method falls short.

Throughout the project, we faced several computational and memory constraints that limited our progress and forced us to search for creative solutions. First, the size of the brain scan for each subject is enormous and unwieldy. Under our constraints, it is impractical to use the entire scan for each prediction. While dividing the scans into mini-volumes increases the training data, it may also lose information. This - and the variety of types and sizes of scans - impedes the model's ability to generalize to unseen data. More substantially, the memory constraint dramatically limited the number of parameters in the CNN and forced us to reduce the size of the raw slices. Our model is likely underpowered and would certainly benefit from more parameters and layers to identify nuanced features in the brain scans. It is also possible that some of this nuanced information is lost in down-scaling the image. Given unlimited resources, we would certainly choose to increase the number of layers and size of the 3D CNN, while maintaining the full size of the image. To conserve memory, we stored the pre-processed images to preserved memory at the expense of the computational burden, noticeably increasing the runtime. Kaggle resources limit the runtime. Hence, the models are trained for fewer epochs than we would have preferred. That said, we have several ideas for future improvement.

Improvements may come from implementing Maximum Intensity Projection (MIP) in the preprocessing step. This may be utilized to reduce the memory burden (in place of the mini-volumes) or as a separate feature to support the mini-volumes. Moreover, incorporating the DICOM metadata attribute `PixelSpacing` when resizing slices may preserve the original spatial resolution and improve image image quality as a whole.

9

The model itself is limited by the CNN infrastructure. In modern image classification tasks, the state-of-the-art methods employ vision transformers (ViT). Vision Transformers utilize the popular self-attention architecture that powers large language models. ViT divide the images into patches and learn low dimensional embeddings of each patch. These embeddings are fed into a transformer encoder layer and finally a classification head. The attention components of the transformer layers are effective for processing sequential images, making it particularly appealing given the three-dimensional nature of this task. In fact, the method scales well to large data sets, so it may alleviate some of the computational difficulties we have faced.

References

David Atkinson. Geometry in medical imaging: DICOM and NIfTI formats. Technical report, University College London, 2022.

MD Anderson Cancer Center. CT scan vs. MRI: What is the difference?, 2022. URL `https://www.mdanderson.org/cancerwise/ct-scan-vs-mri--what-is-the-difference.h00-159616278.html`.

G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012. URL `https://arxiv.org/abs/1207.0580`.

S Ioffe and C Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv*, 2015.

Alexander Keedy. An overview of intracranial aneurysms. *McGill Journal of Medicine*, 2006.

Taewoo Kim, Jin-Woo Jeon, Jiyoon Kim, Jeong-Ho Kim, Jun-Young Chung, Jeong-Hoon Ahn, Jae-Min Park, and Sang-Won Suh. Prevalence of unruptured intracranial aneurysms in healthy asymptomatic adults: magnetic resonance angiography study. *Neurointervention*, 2021.

A. Maas, A. Hannun, and A. Ng. Rectifier nonlinearities improve neural network acoustic models. 2013. URL `https://api.semanticscholar.org/CorpusID:16489696`.

Lenhard Pennig, Ulrike C. I. Hoyer, Alexandra Krauskopf, Rahil Shahzad, Stephanie T. Jünger, Frank Thiele, Kai R. Laukamp, Jan-Peter Grunz, Michael Perkuhn, Marc Schlamann, Christoph Kabbasch, Jan Borggrefe, and Lukas Goertz. Deep learning assistance increases the detection sensitivity of radiologists for secondary intracranial aneurysms in subarachnoid hemorrhage. *Neuroradiology*, 2021.

Filippo Pesapane, Giulia Gnocchi, Cettina Quarrella, Adriana Sorce, Luca Nicosia, Luciano Mariano, Anna Carla Bozzini, Irene Marinucci, Francesca

Priolo, Francesca Abbate, Gianpaolo Carraffiello, and Enrico Cassano. Errors in radiology: A standard review. *Journal of Clinical Medicine*, 2024.

Jeff Rudie, Evan Calabrese, Robyn Ball, Peter Chang, Rennie Chen, Errol Colak, Maria Correia de Verdier, Luciano Prevedello, Tyler Richards, Rachit Saluja, Greg Zaharchuk, Jason Sho, and Maryam Vazirabad. Rsna intracranial aneurysm detection. `https://kaggle.com/competitions/rsna-intracranial-aneurysm-detection`, 2025. Kaggle.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

Lindsy N. Williams and Robert D. Brown. Management of unruptured intracranial aneurysms. *Neurology: Clinical Practice*, 2013.

A Appendix

A.1 Arterial Label Definitions

Table 2 lists the 13 arterial regions used as localization targets in the RSNA Intracranial Aneurysm Dataset.

Table 2: Thirteen labeled arterial regions in the RSNA dataset.

Label Index	Arterial Region
1	Left Infraclinoid Internal Carotid Artery
2	Right Infraclinoid Internal Carotid Artery
3	Left Supraclinoid Internal Carotid Artery
4	Right Supraclinoid Internal Carotid Artery
5	Left Middle Cerebral Artery
6	Right Middle Cerebral Artery
7	Anterior Communicating Artery
8	Left Anterior Cerebral Artery
9	Right Anterior Cerebral Artery
10	Left Posterior Communicating Artery
11	Right Posterior Communicating Artery
12	Basilar Tip
13	Other Posterior Circulation

A.2 Exploratory Data Analysis Visualizations

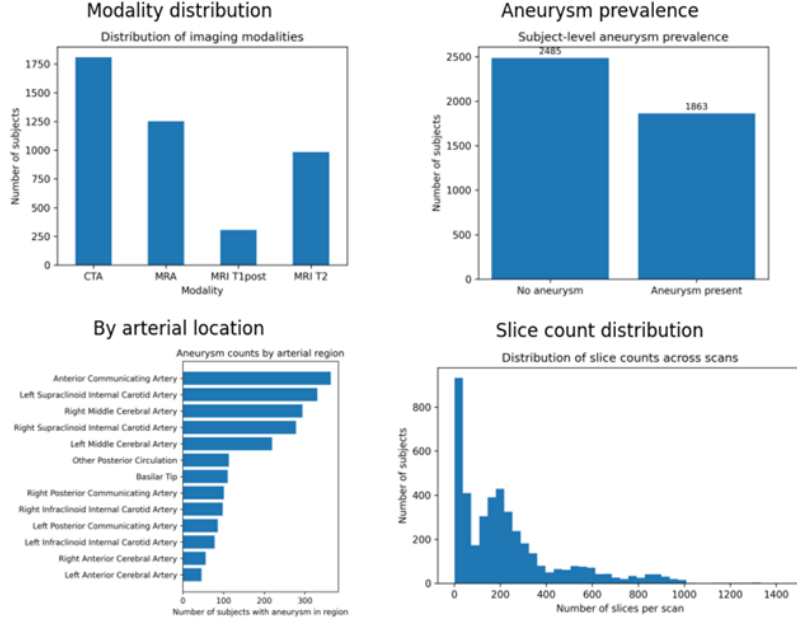Figure 5 provides visual summaries supporting the exploratory analysis in Section 2.

Figure 5: Summary visualizations from exploratory data analysis.

A.3   Model Evaluations

Output	Support	Precision	Recall	F1-Score
Presence	276	0.64	0.88	0.74
Loc - 1	13	0.00	0.00	0.00
Loc - 2	21	0.00	0.00	0.00
Loc - 3	59	0.00	0.00	0.00
Loc - 4	48	0.00	0.00	0.00
Loc - 5	37	0.00	0.00	0.00
Loc - 6	41	0.00	0.00	0.00
Loc - 7	71	0.00	0.00	0.00
Loc - 8	9	0.00	0.00	0.00
Loc - 9	9	0.00	0.00	0.00
Loc - 10	13	0.00	0.00	0.00
Loc - 11	24	0.00	0.00	0.00
Loc - 12	15	0.00	0.00	0.00
Loc - 13	15	0.00	0.00	0.00

Table 3: One-Step Approach Evaluation
We report the precision, recall, and F1-Score for all fourteen output categories.

Output	Support	Precision	Recall	F1-Score
Presence	288	0.72	0.64	0.68
Loc - 1	27	0.00	0.00	0.00
Loc - 2	15	0.00	0.00	0.00
Loc - 3	18	0.00	0.00	0.00
Loc - 4	11	0.00	0.00	0.00
Loc - 5	20	0.00	0.00	0.00
Loc - 6	12	0.00	0.00	0.00
Loc - 7	41	0.00	0.00	0.00
Loc - 8	54	0.00	0.00	0.00
Loc - 9	43	0.00	0.00	0.00
Loc - 10	28	0.00	0.00	0.00
Loc - 11	9	0.00	0.00	0.00
Loc - 12	3	0.00	0.00	0.00
Loc - 13	73	0.00	0.00	0.00

Table 4: Two-Step Approach Evaluation

We report the precision, recall, and F1-Score for all fourteen output categories.