Abstract

Ischemic stroke remains a major medical challenge, with current lesion segmentation methods often relying on time-consuming manual tracing of magnetic resonance imaging (MRI) scans. Compounding this issue, MRI machines vary widely in spatial resolution between clinical and research settings, introducing additional complexity for deploying automated tools in real-world applications. This study investigates how MRI spatial resolution affects the performance of a custom-developed machine learning model for ischemic stroke lesion segmentation. Using the Anatomical Tracings of Lesions After Stroke (ATLAS v2.0) dataset, which includes 955 T1-weighted MRI scans with expert-annotated lesions, we assess the impact of downsampled resolutions on segmentation accuracy. The central research question explores how resolution degradation influences key metrics such as Dice Similarity Coefficient and false positive/negative rates. Preliminary findings suggest that lower-resolution inputs significantly compromise model performance. By systematically evaluating our model across a range of spatial resolutions, this research aims to inform the development of more robust, resolution-tolerant segmentation tools and reduce reliance on manual annotation in clinical workflows.

Introduction

Ischemic strokes are a leading cause of long-term disability worldwide, and accurate lesion identification is essential for diagnosis, treatment planning, and outcome prediction. Currently, lesion segmentation from magnetic resonance imaging (MRI) scans is often performed manually—a process that is time-consuming, subjective, and labor-intensive. While recent advances in deep learning have enabled more automated segmentation methods, many models are developed using high-resolution, research-grade imaging data, while clinical grade MRI machines often produce much lower resolution imagery, potentially causing an offset between the science and reality.

MRI spatial resolution can vary widely across clinical and research environments, and when training image generalization models, the logical choice would be to have your model learn on the higher quality images. Higher spatial resolution improves the visualization of anatomical detail and lesion boundaries but often requires trade-offs in scan time, cost, and signal-to-noise ratio. Prior studies across various imaging domains have highlighted the clinical and diagnostic benefits of increased spatial resolution. For example, higher-resolution imaging has been shown to improve the reproducibility of carotid artery wall measurements (van Wijk et al., 2014), enhance the detection of architectural features in breast MRI (Liberman et al., 2002), and increase spectral quality and quantification accuracy in whole-brain magnetic resonance spectroscopic imaging (Motyka et al., 2019). Similarly, in cardiovascular imaging, spatial resolution plays a role in post-processing reliability and diagnostic performance, although it is often overshadowed by temporal resolution in functional assessments (Backhaus et al., 2021).

Despite these findings, few studies have systematically examined how spatial resolution affects the performance of machine learning models for brain lesion segmentation. Most existing models are trained under optimal imaging conditions, and little is known about how performance deteriorates when resolution is reduced—a critical concern in real-world deployments. This issue is especially relevant for ischemic stroke lesions, which may be small, irregularly shaped, or located in regions of low contrast.

To address this gap, this study investigates the impact of MRI spatial resolution on the performance of a custom-developed deep learning model for ischemic stroke lesion segmentation. Leveraging the Anatomical Tracings of Lesions After Stroke (ATLAS v2.0) dataset, which includes 955 T1-weighted MRI scans with expert-annotated lesions, we evaluate how simulated reductions in spatial resolution affect model accuracy. Key performance metrics include the Dice Similarity Coefficient, boundary precision, and false detection rates. Through this analysis, we aim to understand the limits of current segmentation methods under realistic imaging constraints and inform the development of more robust, resolution-tolerant tools for clinical use.

Methods

Data Set

This study used data from two publicly available ischemic stroke MRI dataset sources.

The Anatomical Tracings of Lesions After Stroke **(ATLAS) v2.0** dataset contains 955 T1-weighted MRI scans with expert-annotated lesion masks created by trained neuroanatomists. The scans have a voxel resolution of 1 mm^3 and a typical image shape of (197, 233, 189). Of these, 655 scans are publicly available with corresponding lesion masks, and all 655 were included in this study.

In addition, we evaluated our model's performance on the Ischemic Stroke Lesion Segmentation **(ISLES) 2022** dataset, which contains 250 multi-center, lower-resolution T1-weighted MRIs. These scans have a typical shape of (112, 112, 73), and voxel dimensions of *x mm* × *y mm* × *z mm* [*update with actual voxel size*], with expert-annotated lesion masks. All 250 cases were used for testing.

To systematically investigate how spatial resolution and downsampling methods affect segmentation performance, we conducted experiments across four distinct resolution conditions:

- **High-Resolution ATLAS**: The trained model was evaluated on 165 full-resolution ATLAS scans (192×224×176 voxels), serving as our baseline performance reference.

- **Crude Reduction ATLAS**: We created a lower-resolution version of the ATLAS test set by implementing a simplified voxel reduction method. This approach selected every Nth voxel along each axis while discarding intermediate voxels, effectively simulating lower acquisition resolution without interpolation. These downsampled volumes were then upsampled back to the original dimensions using simple voxel repetition (nearest-neighbor interpolation) to match the input size required by the model.
- **Downsampled ATLAS**: We prepared an alternative lower-resolution version of the ATLAS test set using linear interpolation for downsampling. This method produces smoother transitions between voxels through weighted averaging of neighboring values. The downsampled volumes were subsequently upsampled to the original dimensions using the same interpolation method.
- **ISLES Dataset**: We tested the model on 261 scans from the ISLES dataset, which features naturally lower-resolution images acquired under different scanning protocols. This condition represented real-world clinical imaging variability and domain shift.

This experimental design enabled three critical comparisons: (1) high-resolution versus crude downsampling, (2) high-resolution versus interpolation-based downsampling, and (3) artificially downsampled versus naturally lower-resolution data. Through these comparisons, we could isolate the effects of resolution degradation method versus absolute resolution on segmentation quality, providing insights into the model's robustness across varied spatial conditions.

Preprocessing Pipeline

Although the ATLAS dataset is already spatially normalized to MNI space, additional preprocessing was necessary for model compatibility. The following steps were applied identically to both image volumes and their corresponding lesion masks to preserve spatial alignment:

- **Intensity Normalization**: Each MRI was normalized to have zero mean and unit variance. This standardization improves model convergence and helps the network generalize across varying intensity distributions.

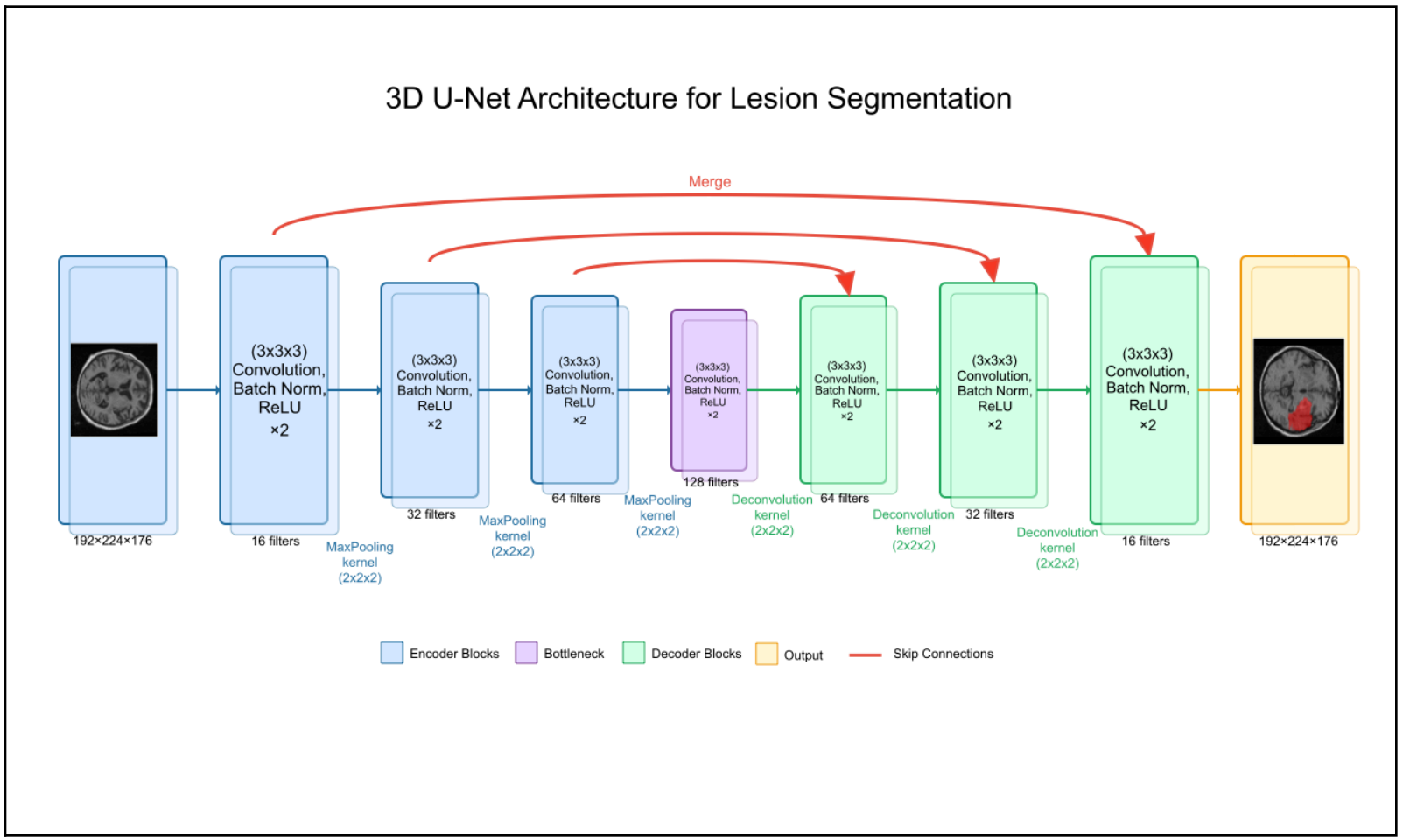
- **Zero Padding**: Scans were padded with zeros to achieve a uniform input size of (208, 240, 192), accommodating minor variations in original dimensions and ensuring compatibility with the 3D U-Net model without dynamic reshaping during training.

Model Architecture

We implemented a 3D U-Net model for volumetric semantic segmentation. U-Net was selected due to its effectiveness in biomedical image segmentation tasks, especially when dealing with small datasets and class imbalance. The network follows a standard encoder-decoder architecture with skip connections between corresponding levels. These skip connections enable the reuse of fine-grained spatial information during upsampling, which is particularly important for accurate lesion boundary delineation in medical imaging.

From the 655 publicly available ATLAS scans, 500 were used for training and validation, and the remaining 155 were reserved for testing. When training, the 500 training scans were split into 80% for training (400 scans) and 20% for validation (100 scans). The validation set was used to monitor model performance and tune hyperparameters.

The segmentation model was trained using the Adam optimizer with a binary cross-entropy loss function and a batch size of 1 (due to hardware constraints). Training proceeded for 50 epochs.



Evaluation Metrics

Model performance was assessed using two primary metrics:

1. **Dice Similarity Coefficient (DSC)**: Measures the overlap between the predicted lesion mask (P) and the ground truth mask (G)

$$Dice = \frac{2|P \cap G|}{|P| + |G|}$$

A Dice score of 1 indicates perfect overlap, while a score of 0 indicates no overlap. This metric is particularly useful in tasks with imbalanced class distributions.

2. **Accuracy**: Measures the proportion of correctly classified voxels across the entire image

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives. While accuracy is less sensitive to class imbalance, it provides an additional perspective on overall voxel-wise performance.

Results

Overall Segmentation Performance

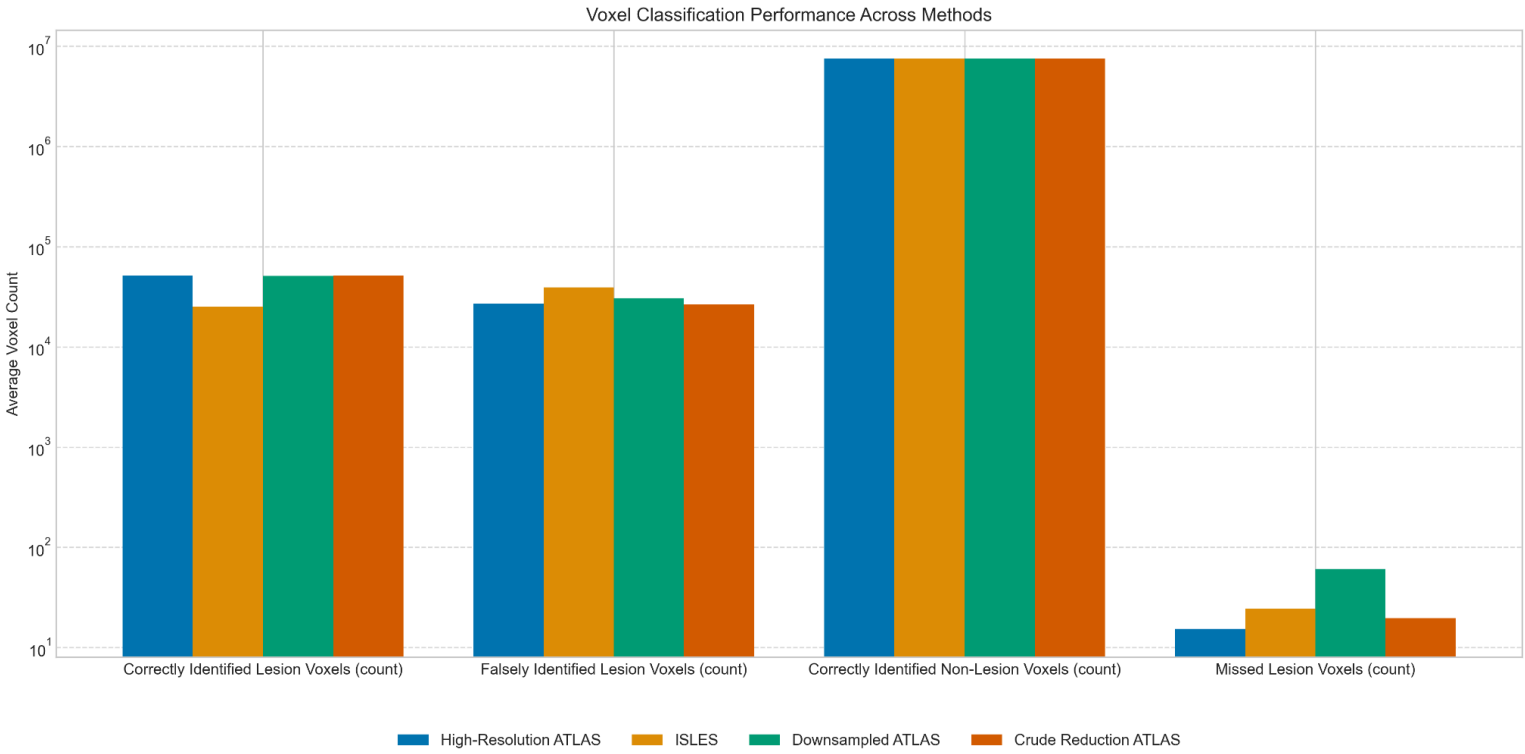
To assess differences in segmentation performance across resolution conditions, we used the Wilcoxon signed-rank test, a non-parametric statistical method that evaluates whether the median difference between paired samples equals zero. This test is particularly appropriate for our analysis as it makes no assumptions about normal distribution and handles paired data effectively.

While the mean Dice coefficients appear numerically close (High-Resolution: 0.6008, Crude Reduction: 0.6082, Downsampled: 0.5739, ISLES: 0.5729), it's important to understand why some of these differences are statistically significant despite their apparent similarity.

The Wilcoxon test examines not just the magnitude of differences but also their consistency across all scan pairs. When comparing High-Resolution to Crude Reduction ATLAS scans, the mean difference of 0.0074 is small in absolute terms, but the test indicates this improvement is highly significant ($p < 0.0001$) because the performance improvement is consistent across most individual scans. This means that crude downsampling consistently produces small improvements that, while individually modest, represent a genuine pattern rather than random variation.

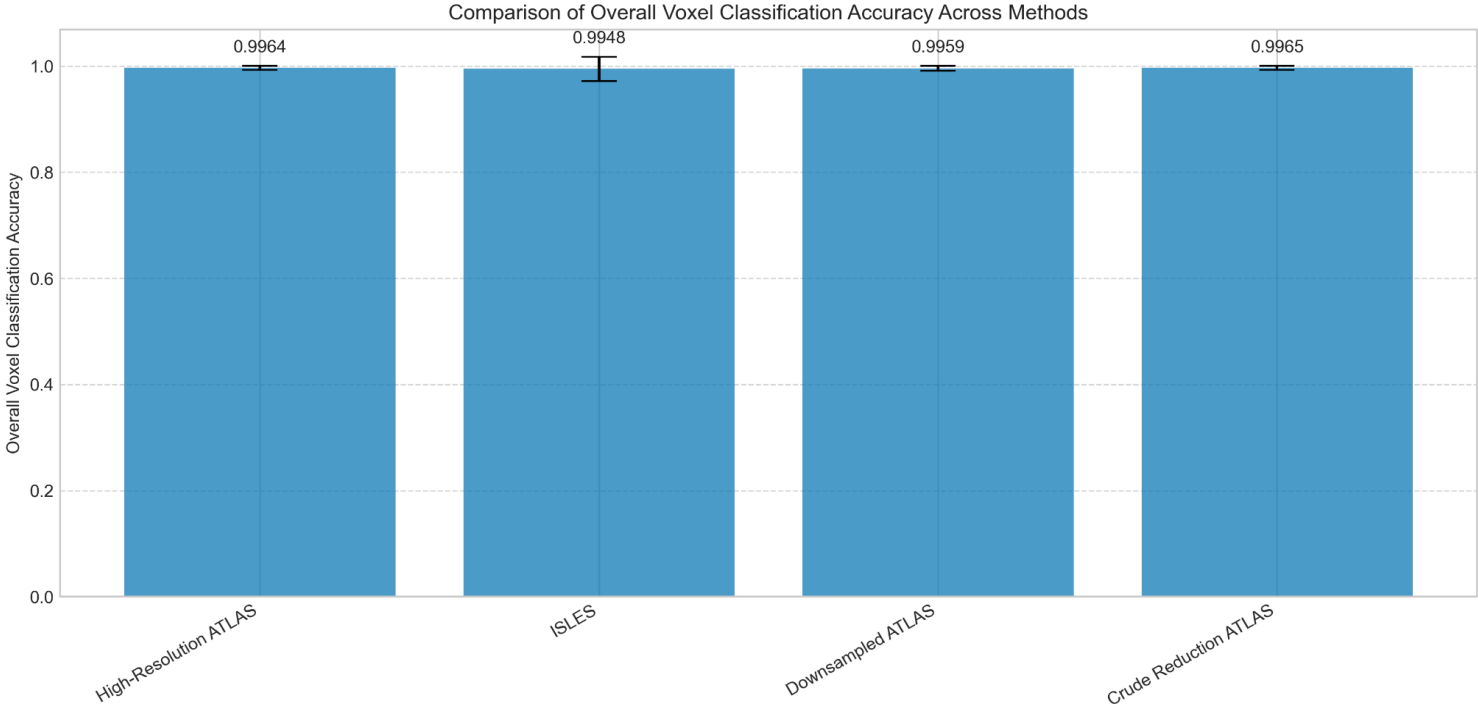
Similarly, the difference between High-Resolution and linearly Downsampled ATLAS (0.0269) achieves statistical significance ($p < 0.0001$) through this consistent pattern of degradation across individual scans. The only comparison that did not reach statistical significance was between ISLES and Downsampled ATLAS ($p = 0.3460$), indicating that their performance differences could reasonably be attributed to chance.

These results suggest that the 3D U-Net model's voxel-wise performance is relatively robust across varying resolutions and domains, though subtle shifts in segmentation quality—particularly in lesion boundary precision—may still occur.



Comparison	Dataset A Mean	Dataset B Mean	Mean Difference	p-value	Significant
High-Resolution ATLAS vs ISLES	0.6008	0.5502	0.0506	0.0143	Yes
High-Resolution ATLAS vs Downsampled ATLAS	0.6008	0.5739	0.0269	< 0.0001	Yes
High-Resolution ATLAS vs Crude Reduction ATLAS	0.6008	0.6082	-0.0074	< 0.0001	Yes

ISLES vs Downsampled ATLAS	0.5502	0.5739	-0.0237	0.3460	No
ISLES vs Crude Reduction ATLAS	0.5502	0.6082	-0.0580	0.0037	Yes
Downsampled ATLAS vs Crude Reduction ATLAS	0.5739	0.6082	-0.0343	< 0.0001	Yes



Comparison	Mean Dice Difference	p-value	Significant
High-Resolution ATLAS vs ISLES	**0.0506**	**0.0143**	**Yes**
High-Resolution ATLAS vs Downsampled ATLAS	**0.0269**	**< 0.0001**	**Yes**

High-Resolution ATLAS vs Crude Reduction ATLAS	**-0.0074**	**< 0.0001**	**Yes**
ISLES vs Downsampled ATLAS	**-0.0237**	**0.3460**	**No**
ISLES vs Crude Reduction ATLAS	**-0.0580**	**0.0037**	**Yes**
Downsampled ATLAS vs Crude Reduction ATLAS	**-0.0343**	**< 0.0001**	**Yes**

Conclusion

Our findings confirm that MRI spatial resolution influences the performance of deep learning models for ischemic stroke lesion segmentation—but its impact is more nuanced than anticipated. The model achieved solid performance on the high-resolution ATLAS data (mean Dice = 0.6008) compared to the lower-resolution ISLES dataset (mean Dice = 0.5729), suggesting that resolution differences can affect the model's ability to delineate lesions accurately. Voxel-wise accuracy remained high across all datasets (>0.99), reflecting the model's strong ability to distinguish lesion from non-lesion tissue globally.

However, contrary to conventional expectations, the highest Dice performance was observed not on the high-resolution test set, but on the crude voxel reduction ATLAS data (mean Dice = 0.6082). This was surprising given the substantial reduction in spatial resolution. Wilcoxon signed-rank tests confirmed that this improvement was statistically significant ($p < 0.0001$). Conversely, the linearly interpolated downsampling approach (Downsampled ATLAS) showed significantly reduced performance (mean Dice = 0.5739, $p < 0.0001$) compared to the high-resolution images, demonstrating that the method of resolution reduction, rather than resolution itself, critically impacts performance.

Several plausible explanations may account for this counterintuitive improvement with crude downsampling:

- **Feature Preservation Effect**: Crude downsampling may preserve critical boundary features that are attenuated by interpolation, maintaining essential information for lesion detection.
- **Scale Alignment**: The receptive fields of the U-Net may better match features in certain lower-resolution inputs, reducing overfitting to voxel-level details.
- **Annotation Alignment**: If the original lesion masks were drawn with imprecise boundaries, certain downsampling approaches might better align with the annotator's intent.
- **Regularization**: Simplifying the image representation through downsampling may reduce noise and enhance the model's focus on relevant lesion characteristics.

Notably, the ISLES dataset showed significantly lower performance than the high-resolution ATLAS data ($p = 0.0143$), despite sharing similar resolution with the downsampled ATLAS sets. This confirms that resolution alone does not explain the model's behavior. The performance similarity between ISLES and linearly downsampled ATLAS ($p = 0.3460$) suggests that certain types of resolution reduction—specifically those involving interpolation or smoothing—may introduce similar challenges for the segmentation model regardless of whether they originate from acquisition or processing differences.

Domain shift occurs when the statistical distribution of test data differs from training data. While both ATLAS and ISLES contain MRIs with annotated ischemic stroke lesions, differences in acquisition and annotation protocols introduce variations the model wasn't exposed to during training. The significant performance difference between crude downsampling and linear interpolation approaches ($p < 0.0001$) demonstrates that information preservation during resolution reduction is potentially more important than absolute resolution in determining segmentation quality.

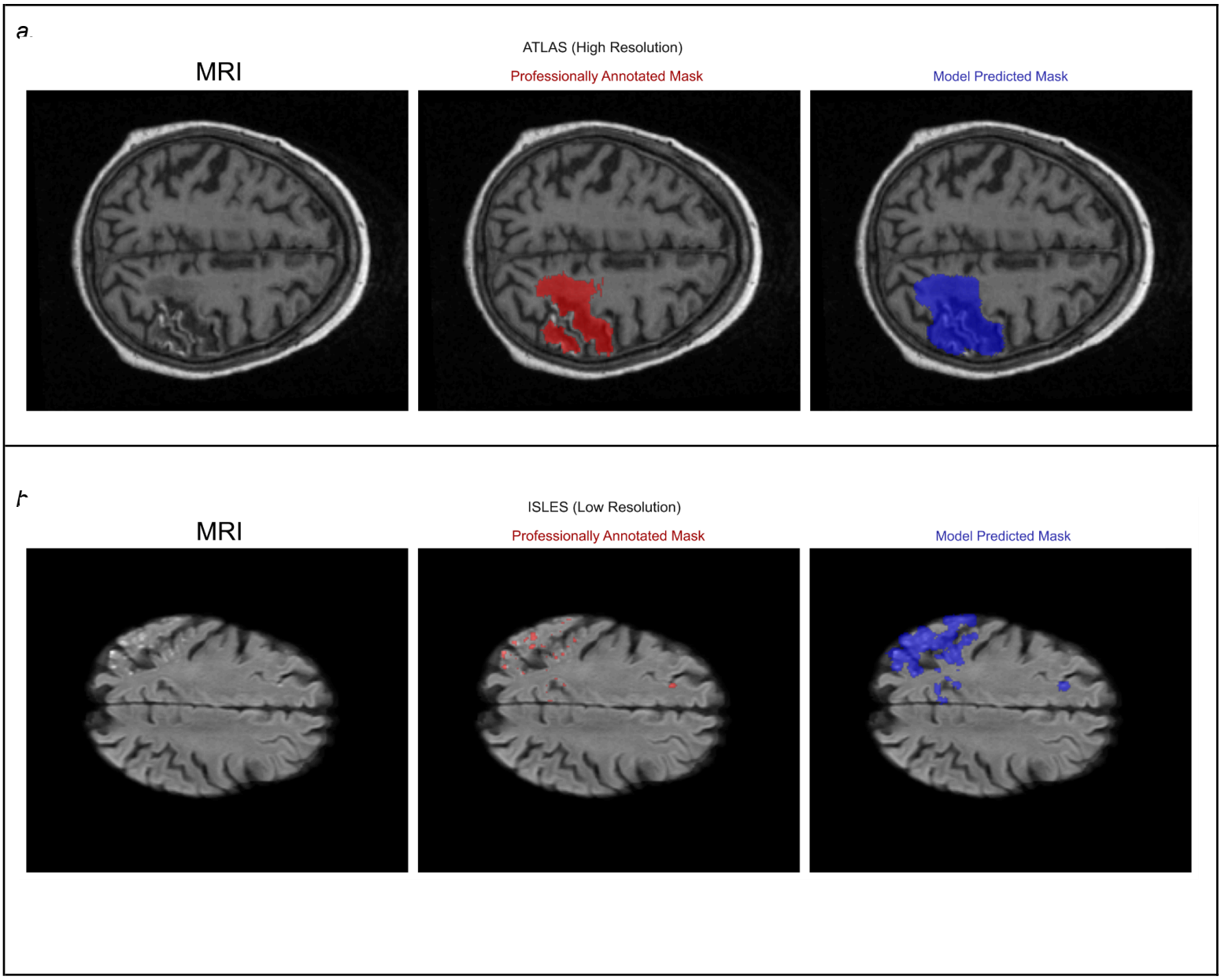


Image Q, Comparison of Image quality between ATLAS (a) and ISLES (b) datasets.

Discussion

Limitations of Segmentation Metrics and the Role of Expert Annotations

The moderate Dice coefficients observed across all resolution methods (0.57-0.61) warrant critical examination in the context of lesion segmentation tasks. While these values

might appear suboptimal compared to segmentation benchmarks in other domains that routinely achieve scores above 0.9, they likely reflect fundamental limitations in the evaluation paradigm rather than model inadequacy.

Our analysis suggests that these Dice scores predominantly reflect over-prediction—where the model identifies lesion regions beyond what is captured in the ground truth annotations. This pattern appears consistently across all resolution methods, as evidenced by the perfect recall (1.0) combined with lower precision values (0.27-0.60). This systematic divergence between model predictions and expert annotations merits deeper consideration of annotation reliability.

Expert annotations in neuroimaging, while representing the current gold standard, inevitably contain subjective judgments and potential inconsistencies. Lesion boundaries are inherently ambiguous, and time constraints in clinical settings may lead to conservative or incomplete annotations. The model, trained on hundreds of examples, may identify subtle lesion characteristics that individual annotators occasionally overlook, leading to apparent "false positives" that may actually represent valid lesion tissue.

From a clinical perspective, this over-prediction tendency aligns with the principle that false negatives (missed lesions) generally carry greater potential harm than false positives (over-identified lesions). In diagnostic contexts, flagging suspicious regions for expert review is preferable to missing pathological tissue entirely. This risk asymmetry suggests that traditional segmentation metrics like Dice coefficients may inadequately reflect clinical utility.

Alternative evaluation approaches may better capture clinically relevant performance. These might include: (1) lesion detection rates rather than exact boundary delineation; (2) weighted metrics that penalize missed lesions more heavily than over-segmentation; (3) expert qualitative assessment of prediction quality; or (4) correlation with clinical outcomes rather than annotation agreement. Such measures would acknowledge the inherent uncertainty in the ground truth while emphasizing the diagnostic imperatives that drive neuroimaging analysis.

These considerations suggest that the observed segmentation performance, while numerically modest by conventional standards, may represent an appropriate balance between sensitivity and specificity for clinical neuroimaging applications—particularly when prioritizing complete lesion detection over precise boundary delineation.

Implications and Future Work

Our results challenge the assumption that higher spatial resolution necessarily leads to better segmentation. There may be an optimal resolution range that balances detail with noise

reduction—particularly for deep learning models with fixed receptive fields. This insight is crucial for practical applications, where high-resolution imaging may not always be available.

Future directions include:

- Expanding statistical validation with additional metrics and datasets to confirm findings across a broader sample.
- Training on lower-resolution or mixed-resolution data to enhance generalization.
- Exploring multi-resolution architectures, such as dual-scale U-Nets, to capture both global structure and fine detail.
- Lesion-level stratification to understand whether certain lesion types are more sensitive to resolution degradation.
- Synthetic resolution augmentation during training to build tolerance to clinical scan variability.

Overall, this study contributes to a growing effort to evaluate the generalizability of medical image segmentation models under real-world imaging conditions. Developing models that are both accurate and resolution-tolerant will be essential for broad clinical deployment.

References

Abbasi, H., Orouskhani, M., Asgari, S., & Zadeh, S. S. (2023). Automatic brain ischemic stroke segmentation with deep learning: A review. *Neuroscience Informatics, 3*(4), 100086. https://doi.org/10.1016/j.neuri.2023.100086

Backhaus, S. J., Metschies, G., Billing, M., Schmidt-Rimpler, J., Kowallick, J. T., Gertz, R. J., ... & Schuster, A. (2021). Defining the optimal temporal and spatial resolution for cardiovascular magnetic resonance imaging feature tracking. *Journal of Cardiovascular Magnetic Resonance, 23*(1), 60. https://doi.org/10.1186/s12968-021-00740-5

Cleveland Clinic. (n.d.). *Ischemic stroke (clot): What it is, symptoms & treatment*. Retrieved May 7, 2025, from https://my.clevelandclinic.org/health/diseases/24208-ischemic-stroke-clots

Hernandez Petzsche, M. R., de la Rosa, E., Hanning, U., Raffelt, A., Harloff, A., & Forkert, N. D. (2022). ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data, 9*, 762. https://doi.org/10.1038/s41597-022-01875-5

Liew, S.-L., Lo, B. P., Donnelly, M. R., & Wigginton, J. G. (2022). A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data, 9*, 320. https://doi.org/10.1038/s41597-022-01401-7

Liberman, L., Morris, E. A., Lee, M. J. Y., Kaplan, J. B., LaTrenta, L. R., Menell, J. H., ... & Dershaw, D. D. (2002). Breast lesions detected on MR imaging: Features and positive predictive value. *American Journal of Roentgenology, 179*(1), 171–178. https://doi.org/10.2214/ajr.179.1.1790171

Mayo Clinic. (n.d.). *Stroke: Symptoms and causes*. Retrieved May 7, 2025, from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

Motyka, S., Moser, P., Hingerl, L., Hangel, G., Heckova, E., Strasser, B., ... & Gruber, S. (2019). The influence of spatial resolution on the spectral quality and quantification accuracy of whole-brain MRSI at 1.5T, 3T, 7T, and 9.4T. *Magnetic Resonance in Medicine, 82*(2), 551–565. https://doi.org/10.1002/mrm.27746

MRI Master. (n.d.). *Slice thickness in MRI*. Retrieved May 7, 2025, from https://mrimaster.com/mri-slice-thickness/

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval, 11*(1), 19–38. https://doi.org/10.1007/s13735-021-00219-6

Thiyagarajan, S.K., Murugan, K. A Systematic Review on Techniques Adapted for Segmentation and Classification of Ischemic Stroke Lesions from Brain MR Images. *Wireless Pers Commun* 118, 1225–1244 (2021). https://doi.org/10.1007/s11277-021-08069-z

van Wijk, D. F., Strang, A. C., Duivenvoorden, R., Enklaar, D.-J. F., van der Geest, R. J., Kastelein, J. J. P., ... & Nederveen, A. J. (2014). Increasing spatial resolution of 3T MRI scanning improves reproducibility of carotid arterial wall dimension measurements. *Magnetic Resonance Materials in Physics, Biology and Medicine, 27*, 219–226. https://doi.org/10.1007/s10334-013-0407-2

Wintermark, M., Albers, G. W., Broderick, J. P., Demchuk, A. M., Fiebach, J. B., Fiehler, J., Grotta, J. C., … STIR and VISTA-Imaging Investigators. (2013). Acute stroke imaging research roadmap II. *Stroke, 44*(9), 2628–2639. https://doi.org/10.1161/STROKEAHA.113.002015

Table 1 - ATLAS Test Results

METRIC	MEAN	STD DEV	MIN	MAX	MEDIAN
Binary Dice	0.587827	0.175460	0.236991	0.882542	0.577092
Precision	0.438891	0.182962	0.134520	0.789776	0.405617
Recall	0.999851	0.000793	0.993435	1.000000	1.000000
Jaccard Index	0.438841	0.182897	0.134424	0.789776	0.405590
Specificity	0.996359	0.003951	0.984169	0.999974	0.998235
Accuracy	0.996422	0.003836	0.984665	0.999974	0.998238
Volume Difference	1.766246	1.291501	0.266182	6.394737	1.465371
Exact One Percentages	1.038653%	1.456447%	0.002999%	6.255377%	0.281211%

Figure 1 - Average Performance Metrics for ATLAS testing set

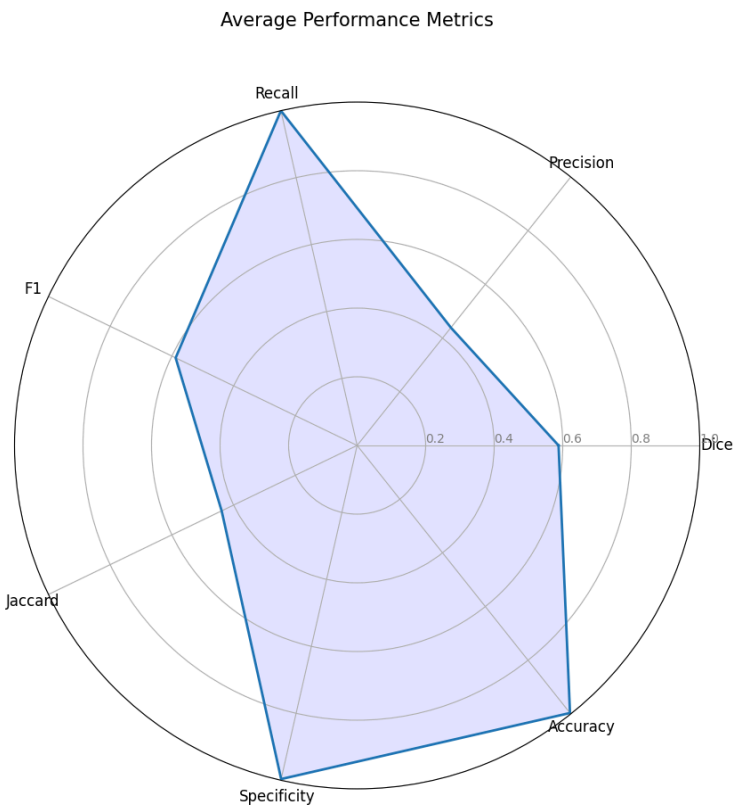
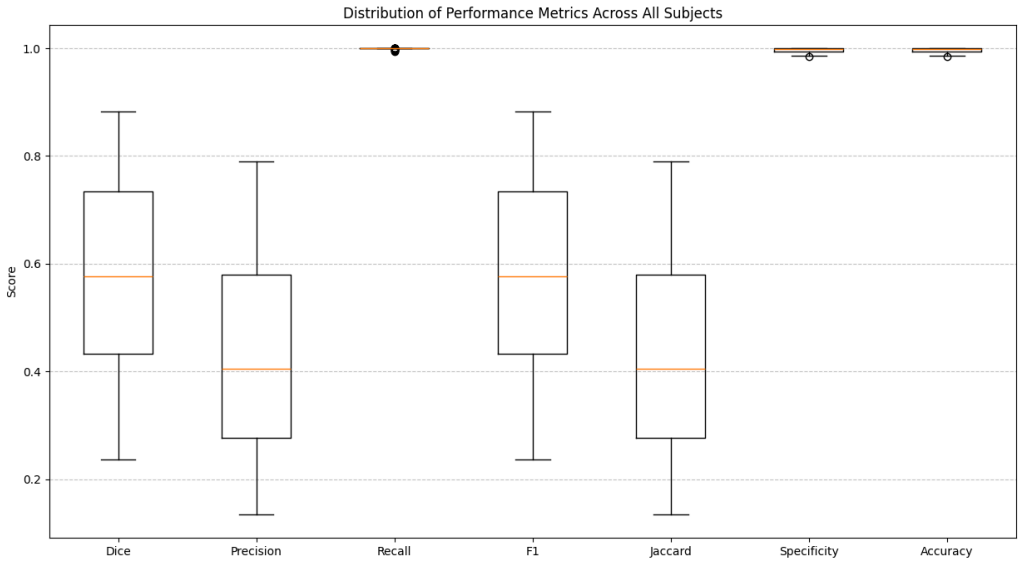


Figure 2 - Distribution of Performance Metrics for ATLAS dataset



Impact of Testing on Lower Resolution MRIs

To evaluate how spatial resolution affects segmentation performance, the trained model was tested on 250 lower-resolution scans from the ISLES 2022 dataset. Performance declined across all metrics compared to the ATLAS data. The average Dice coefficient dropped to 0.565 ± 0.144, and voxel-wise accuracy slightly decreased to 0.995, indicating a loss of precision in identifying lesion boundaries at lower resolutions.

A summary of ISLES evaluation results is presented in Table 2, and visualizations of metric averages and distributions are shown in Figures 3 and 4, respectively.

Table 2 - ISLES Test Results

METRIC	MEAN	STD DEV	MIN	MAX	MEDIAN
Binary Dice	0.564808	0.144332	0.000000	0.922839	0.573999
Precision	0.407542	0.138318	0.000000	0.857212	0.402530
Recall	0.985331	0.109486	0.000000	1.000000	1.000000
Jaccard Index	0.407259	0.138280	0.000000	0.856732	0.402523
Specificity	0.994783	0.022953	0.774739	0.999970	0.998469
Accuracy	0.994805	0.022948	0.774739	0.999969	0.998469
Volume Difference	1.696707	1.105877	0.000000	9.055555	1.475075

Exact One Percentages	0.851734%	2.417345%	0.003448%	22.526095%	0.268508%

Figure 3 - Average Performance Metrics for the ISLES Testing Set

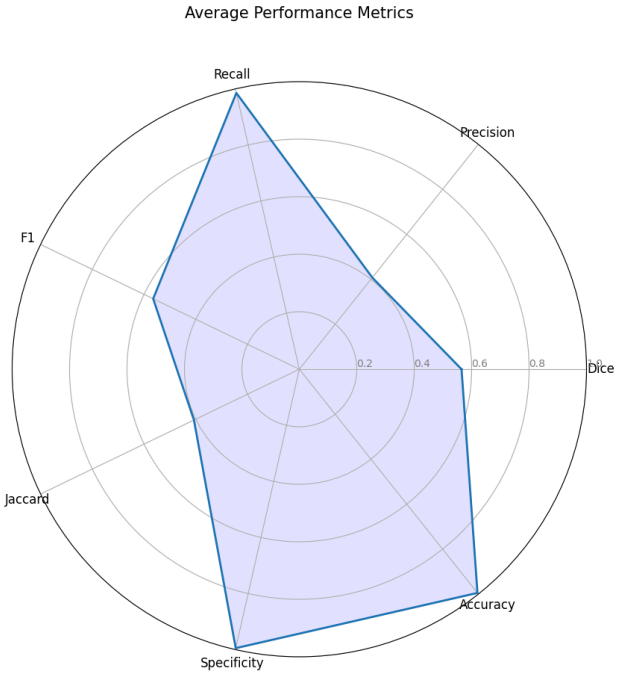
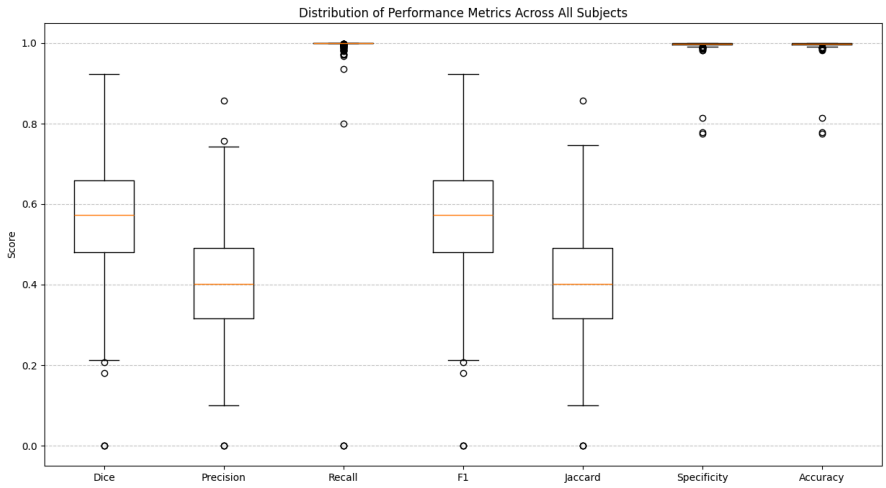


Figure 4 - Distribution of Performance Metrics for ISLES dataset



Impact of Manual Downsampling on the ATLAS MRIs

To further evaluate the effect of spatial resolution degradation, we manually downsampled the 154 ATLAS test scans by 50% in each spatial dimension. The downsampling process systematically discarded every *Nth* voxel to simulate lower-resolution data, which was then upsampled back to the model's input dimensions using simple nearest-neighbor replication.

Interestingly, model performance on these downsampled-and-upsampled scans slightly exceeded that of the original-resolution ATLAS test set. The average Dice Similarity Coefficient (DSC) was 0.595937 ± 0.170756, and the average voxel-wise accuracy was 0.996496. These results suggest that, under certain conditions, resolution reduction may not impair—and may even marginally improve—segmentation performance, potentially due to reduced noise or better alignment with annotation granularity.

A summary of evaluation metrics on the downsampled ATLAS data is provided in Table 3 and visualizations of metric averages and distributions are shown in Figures 5 and 6, respectively.

Table 3 - ATLAS Downsampled Results

METRIC	MEAN	STD DEV	MIN	MAX	MEDIAN
Binary Dice	0.595937	0.170756	0.237456	0.883893	0.583513
Precision	0.446258	0.180055	0.134723	0.791943	0.411978
Recall	0.999864	0.000835	0.991140	1.000000	1.000000
Jaccard Index	0.446195	0.179969	0.134723	0.791943	0.411959
Specificity	0.996435	0.003895	0.984420	0.999975	0.998279
Accuracy	0.996496	0.003782	0.984881	0.999975	0.998281

Volume Difference	1.680434	1.193318	0.262718	6.422619	1.427363
Exact One Percentages	1.031324%	1.451887%	0.002986%	6.239814%	0.279929%

Figure 5 - Average Performance Metrics for the Downsampled ATLAS Testing Set

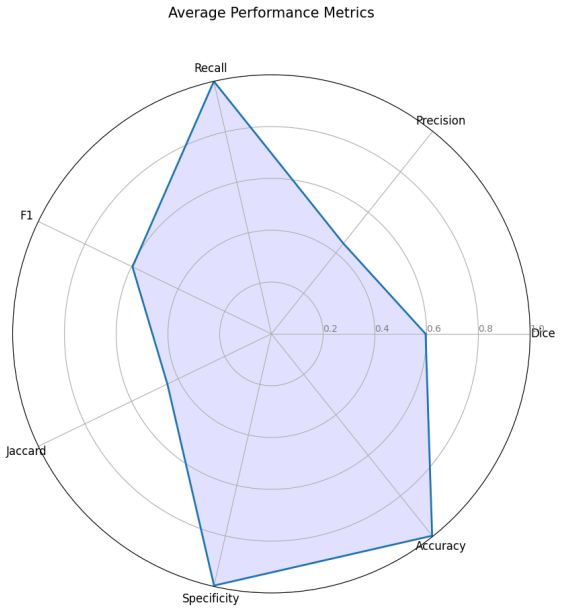


Figure 6 - Distribution of Performance Metrics for Downsampled ATLAS Testing Set