A Reproducible Study of Learning-Based Medical Image Analysis

Anonymous Author(s)

Anonymous Institution(s)

Abstract. Learning-based medical image analysis systems must be evaluated under a protocol that separates genuine generalization from improvements caused by data leakage, favorable case selection, or post-hoc tuning. This manuscript starter organizes a MICCAI study around that principle. It specifies the clinical task before introducing the model, formalizes patient-level data splits and the learning objective, and connects every scientific claim to a predefined experiment. It also demonstrates how to report data provenance, implementation details, uncertainty, statistical comparisons, and failure analysis without relying on supplementary text. The resulting structure is intended to make a submission concise, auditable, and reproducible while remaining compatible with the Springer Lecture Notes in Computer Science format. Authors should replace the illustrative prose and values with their study-specific evidence; no experimental result should be retained unless it is supported by the submitted work.

Keywords: Medical image computing · Reproducibility · Deep learning · Evaluation methodology

1 Introduction

Medical image computing methods are commonly developed from heterogeneous clinical data collected across patients, scanners, sites, and time. These sources of variation make evaluation design inseparable from model design: a high score is not evidence of clinical utility when related observations occur in both training and test sets, when model selection uses the test set, or when important patient subgroups are absent from the evaluation. Strong segmentation architectures such as U-Net [6], 3D U-Net [1], and nnU-Net [3] illustrate the value of carefully matched architectures and training pipelines, but fair comparison still depends on a shared and transparent protocol.

A clear MICCAI paper should answer three questions early. First, what clinically or scientifically meaningful problem is unresolved? Second, why do existing

methods fail to resolve it? Third, which evidence would distinguish the proposed explanation or method from plausible alternatives? The final paragraph of the introduction should state contributions as testable claims, not as a list of implementation components. A suitable pattern is:

- a methodological contribution linked to a specific limitation of prior work;
- an evaluation design that tests generalization under clinically relevant variation; and
- an analysis that explains where the method succeeds, where it fails, and how certain those conclusions are.

The remainder of the paper should then provide evidence for these claims in the same order. Related work establishes the gap, the method defines the intervention, and the experiments isolate its effect.

2 Related Work

Organize related work by technical idea or unresolved limitation rather than by listing papers chronologically. For segmentation, encoder–decoder models with skip connections [6] and volumetric extensions [1,5] are natural reference points. Self-configuring pipelines such as nnU-Net [3] are important comparators because they reduce gains caused only by manual configuration. When confidence estimates are part of the contribution, distinguish predictive uncertainty [4] from empirical calibration quality [2]. End the section with the precise gap left by these approaches and explain why the proposed study is needed.

Avoid claims that a method is "the first" unless the literature search makes that claim defensible. Cite original methods, datasets, and evaluation measures rather than secondary summaries. During double-blind review, self-citations should use the same grammatical form as other citations and must not reveal identity.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i, p_i, s_i)\}_{i=1}^{N}$ denote a dataset of N observations, where $x_i \in \mathbb{R}^{H \times W \times D \times C}$ is an image, $y_i \in \{0, 1, \ldots, K\}^{H \times W \times D}$ is its reference annotation, p_i identifies the patient, and s_i identifies the acquisition site. A model f_θ produces class probabilities $q_i = f_\theta(x_i)$, with q_{ivk} denoting the probability of class k at voxel v. Defining patient and site identifiers explicitly makes the unit of independence clear and prevents image-level splitting from leaking information between partitions.

For a segmentation example, a combined cross-entropy and soft Dice objective can be written as

$$\begin{aligned}\mathcal{L}(\theta) = & -\frac{1}{|V|}\sum_{v\in V}\sum_{k=1}^{K} y_{vk}\log q_{vk} \\ & +\lambda\left(1-\frac{2\sum_{v\in V}\sum_{k=1}^{K} q_{vk}y_{vk}+\epsilon}{\sum_{v\in V}\sum_{k=1}^{K} q_{vk}^2+\sum_{v\in V}\sum_{k=1}^{K} y_{vk}^2+\epsilon}\right),\end{aligned} \tag{1}$$

where V is the set of voxels, λ controls the contribution of the Dice term, and $\epsilon > 0$ provides numerical stability. Every symbol should be defined on first use, and the manuscript should state how multiclass and empty reference cases are handled.

3.2 Model and Inference

Describe the model at a level from which an independent researcher could reimplement it: input representation, dimensionality, encoder and decoder, normalization, nonlinearities, skip connections, output parameterization, and parameter count. A schematic should add information beyond the prose. The compact example in Fig. 1 emphasizes the boundary between training-only operations and evaluation.

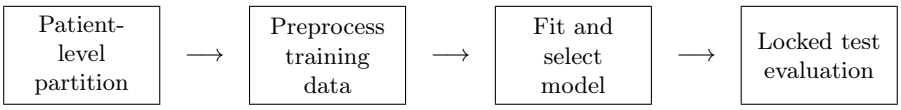

Fig. 1. A leakage-aware experimental workflow. Statistics used for preprocessing and model selection are estimated without access to the test set; the locked model is evaluated once on each predefined test cohort.

Report all inference-time operations, including resampling, sliding-window overlap, ensembling, test-time augmentation, thresholding, and connected component processing. If these choices were tuned, identify the partition on which tuning occurred.

3.3 Training and Reproducibility

State the software framework and version, hardware, optimizer, learning-rate schedule, batch size, epoch or iteration budget, augmentation distribution, initialization, checkpoint-selection rule, and random seeds. Report whether mixed precision, pretrained weights, or external data were used. A public or anonymous repository should include training and evaluation commands, environment specifications, model weights where permissible, and code that recreates every reported table.

4 Experiments

4.1 Data and Study Cohorts

For every dataset, declare its origin, license or data-use agreement, cohort construction, inclusion and exclusion criteria, acquisition equipment and protocols, annotation procedure, quality control, and ethics approval or the reason approval was not required. Report counts at the patient level and explain missing data. Table 1 is a compact reporting pattern; all values in an actual submission must be computed from the final cohort.

Table 1. Cohort information that should be reported for each partition. The entries describe the statistic to provide rather than fabricated study data.

Partition	Patients	Images	Sites	Intended role
Training	final count	final count	site count	parameter fitting
Validation	final count	final count	site count	model selection
Internal test	final count	final count	site count	in-domain estimate
External test	final count	final count	site count	transportability

Partition data by patient before any patch extraction or augmentation. When multiple centers are available, preserve an external center or acquisition period for testing whenever that choice matches the intended use. Fit normalization parameters only on the training set. If cross-validation is used, state whether the same folds and preprocessing are shared by all methods.

4.2 Comparators, Metrics, and Statistical Analysis

Compare against the most relevant published approach, a strong standardized baseline, and ablations that isolate each claimed contribution. Use the same data, augmentation, tuning budget, and evaluation code for all trainable methods whenever possible. Report computational cost when the contribution changes memory, latency, or parameter count.

Choose a primary endpoint before inspecting the test results. For segmentation, overlap measures such as Dice should be complemented by a boundary measure when surface accuracy matters; class imbalance may motivate a generalized Dice formulation [7]. Report distributions or confidence intervals, not only means. Paired bootstrap intervals over patients provide one practical choice: sample patients with replacement, recompute the paired difference for each resample, and use its empirical quantiles. Correct for multiple comparisons when several primary hypotheses are tested, and accompany significance tests with effect sizes.

4.3 Results

Present the primary result first and keep metric direction explicit. The table layout in Table 2 separates model selection from evaluation and shows the minimum information needed to interpret a comparison. Replace the descriptive cells with estimates and uncertainty intervals obtained from the locked test analysis.

Table 2. Recommended primary-results layout. An actual manuscript should report numerical estimates with patient-level uncertainty intervals.

Method	Dice ↑	Surface distance ↓	Parameters	Runtime
Standardized baseline	estimate	estimate	count	seconds/case
Proposed method	estimate	estimate	count	seconds/case
Without component A	estimate	estimate	count	seconds/case
Without component B	estimate	estimate	count	seconds/case

The accompanying text should quantify the effect rather than repeat every cell. State whether the main comparison supports the prespecified hypothesis, then describe ablations, external validation, and subgroup analyses. Negative or inconclusive findings are part of the scientific result and should not be hidden.

5 Discussion

Interpret results in relation to the stated clinical or scientific problem. Separate observations directly supported by the experiment from possible mechanisms. Examine representative successes and failures, but do not use selected images as a substitute for aggregate analysis. Discuss performance across relevant subgroups and acquisition conditions, especially when the training cohort is not representative of the intended population.

Limitations should be specific. Typical examples include retrospective data, small or geographically narrow cohorts, imperfect reference standards, lack of reader comparison, untested domain shifts, and computational requirements. Explain how each limitation constrains the conclusion and identify the next experiment needed to address it. Clinical claims should match the study design: technical validation alone does not demonstrate improved patient outcomes.

6 Conclusion

Conclude with the problem, the principal methodological contribution, and the single most important supported result. Avoid introducing new experiments or restating the abstract verbatim. A useful final sentence defines the scope in which the evidence applies and the validation still required before broader use.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer (2016)
2. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR, vol. 70, pp. 1321–1330 (2017)
3. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021)
4. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, vol. 30 (2017)
5. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision, pp. 565–571. IEEE (2016)
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer (2015)
7. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. LNCS, vol. 10553, pp. 240–248. Springer (2017)