Data leakage and external validation in lung CT segmentation

Mihai Nan[1]✉, Cătălin-Mihail Chiru[1], Irina Mocanu[1], and Adina-Magda Florea[1]

Department of Computer Science, Faculty of Automatic Control and Computers, National University of Science and Technology POLITEHNICA Bucharest,
mihai.nan@upb.ro, catalin.chiru@upb.ro, irina.mocanu@upb.ro, adina.florea@upb.ro

Abstract. In medical image segmentation, reported performance is strongly influenced by the evaluation protocol, especially when datasets such as LIDC-IDRI lack a standard testing procedure. We experimentally study how split granularity, information leakage, multi-reader target construction, and Dice aggregation alter reported CT lung nodule segmentation performance. Using the same 2D/2.5D setup, we compare patient-level and random sample-level splitting and evaluate frozen LIDC-IDRI models on MSD Task06 Lung. The leakage audit finds no train–test patient or image overlap for the patient-level protocol, whereas the sample-level protocol shares 568 of 579 test patients and 724 image identifiers with training. The sample-level split obtains higher internal LIDC-IDRI scores, but this apparent advantage reverses externally, where the patient-level model performs better. A controlled same-image experiment further shows that differences among union, majority-vote, and intersection targets shrink substantially when case selection is fixed. These results are empirical observations from the investigated LIDC-IDRI–MSD setting and emphasize the need to report split granularity, leakage, target construction, metric aggregation, and external-validation details explicitly.

Keywords: Medical image segmentation · Data leakage · Dice score · External validation · LIDC-IDRI · Lung CT · Multi-reader annotations

1 Introduction

Lung cancer has a major impact on patient survival, and early detection is particularly important when lesions are still small and appear as pulmonary nodules that are difficult to delineate. Accurate nodule segmentation in CT images supports lesion quantification, disease monitoring, treatment planning, and computer-aided diagnosis. However, in medical imaging, reported performance is meaningful only if it reflects genuine generalisation rather than artefacts introduced by the evaluation protocol.

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [2] is widely used for pulmonary nodule analysis, but its structure makes evaluation non-trivial. A single CT volume can generate multiple correlated 2D or 2.5D samples, and the same lesion may have annotations from different radiologists. Therefore, random sample-level splitting can place correlated information from the same patient or anatomical region in both training and testing sets. At the same time, the way multiple annotations are converted into a target and the strategy used to aggregate Dice scores can alter the interpretation of quantitative results even when the same underlying dataset is used.

In this work, we audit these protocol choices within a controlled LIDC-IDRI lung nodule segmentation study. Using the same 2D/2.5D segmentation setup, we compare patient-level and sample-level splits, analyse metric aggregation strategies, and study target construction from multiple annotations. We quantify train–test contamination at patient and axial-slice image-identifier level and evaluate frozen LIDC-IDRI models, without fine-tuning, on a matched lesion-centred MSD Task06 Lung setting [1]. The results show that sample-level splitting increases internal LIDC-IDRI scores, but this advantage does not transfer externally, where the patient-level model performs better. Our goal is not to propose a new segmentation architecture, but to show how evaluation choices can change the conclusions drawn from an otherwise fixed experimental pipeline.

2 Related Work

Most studies using LIDC-IDRI for lung nodule segmentation focus on improving the segmentation model or the learning strategy. U-Net-based architectures, attention mechanisms, residual encoder–decoder designs, transformer-based variants, and self-configuring pipelines have all been explored for lesion segmentation in medical imaging [8, 4]. In LIDC-IDRI specifically, the availability of multiple radiologist annotations has motivated approaches based on selecting a reference annotation, constructing consensus masks, or explicitly modelling inter-observer disagreement and uncertainty [2, 12, 13]. These studies can differ substantially in case selection, target definition, data splitting, and metric aggregation, which complicates direct comparison of reported Dice scores.

Methodological studies have separately documented several of these evaluation risks. Inappropriate splitting can inflate performance when correlated samples from the same subject appear in both training and testing [11, 9, 10]. External validation has been emphasized as an important test of whether internal performance transfers beyond the development dataset [14]. In parallel, recent work has shown that segmentation metrics depend on implementation details, object size, class imbalance, and aggregation strategy [6, 7, 5]. Unlike work addressing these issues separately, our study examines split granularity, leakage, external validation, multi-annotation target construction, and Dice aggregation within one controlled LIDC-IDRI segmentation setting.

3 Protocol-Aware Evaluation Study

Figure 1 summarises the study design and the protocol dimensions investigated in our experiments.

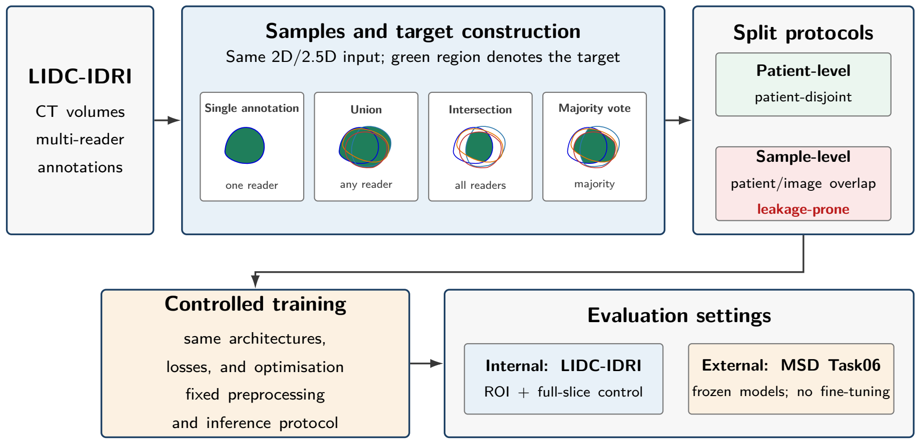

Fig. 1. Overview of the protocol-aware evaluation study.

3.1 Dataset Preparation

To isolate the effect of the evaluation protocol, we kept the data preparation pipeline fixed across paired experiments and varied only the split strategy or the target construction rule. CT volumes were represented as 2D or 2.5D samples and each sample retained patient, case, and axial-slice identifiers. This organisation enabled both patient-level and sample-level splitting and made it possible to audit train–validation–test overlap explicitly.

For LIDC-IDRI, targets were constructed according to the annotation policy under study, including single-annotation, union, intersection, and majority-vote masks. The main 2.5D configuration used 128×128 lesion-centred ROIs with three channels corresponding to the previous, current, and next axial slice; when a neighbouring slice was unavailable, the current slice was repeated. Samples with fewer than 20 foreground pixels were excluded. When applicable, CT intensities were normalised using a lung window of $[-1000, 400]$ HU, and training used simple spatial and intensity augmentation. Lesion-centred crops were defined from the reference mask, so the ROI experiments evaluate segmentation within a known lesion region rather than end-to-end lesion detection.

3.2 Segmentation Setup

We formulated the task as binary pulmonary-lesion segmentation. The primary model was a convolutional U-Net [8]; architecture controls used Attention U-Net and a compact U-Net with a Transformer bottleneck inspired by TransUNet [3]. The purpose of the architecture sweep was to determine whether the effect of the evaluation protocol was specific to a single encoder–decoder design rather than to establish a new state of the art.

For the main U-Net experiments, AdamW was used with learning rate 2×10^{-4}, weight decay 10^{-5}, Dice–Focal loss, batch size 48, and up to 150 epochs. Early stopping had the patience parameter fixed to 25 epochs and the checkpoint with the best validation mean per-case Dice was retained after convergence. The nominal train/validation/test fractions were 70%/10%/20%, and the same overall training and inference pipelines were used across paired protocol comparisons.

3.3 Split Protocols and Leakage Audit

We compare two data-splitting protocols. In the patient-level split, all samples originating from the same patient are assigned exclusively to one of the training, validation, or test partitions. This setting evaluates generalisation to unseen patients. In the random sample-level split, 2D or 2.5D samples are partitioned without enforcing patient-level separation. Because a CT volume can generate multiple correlated samples and a slice can be associated with multiple annotation-derived examples, this protocol is susceptible to information leakage.

The leakage audit was performed at both patient and axial-slice image-identifier level. The patient-level split contained 8,634 training, 965 validation, and 1,500 test samples from 605, 86, and 169 patients, respectively, and had zero patient and image-identifier overlap across all partitions. The sample-level split contained 8,427 training, 1,204 validation, and 1,500 test samples. Its test set represented 579 patients, of whom 568 (98.1%) also occurred in training; moreover, 724 axial-slice image identifiers were shared between train and test. These values directly quantify the contamination of the sample-level protocol rather than inferring leakage solely from performance differences.

For robustness, we also evaluated repeated patient-level holdout partitions generated independently from the main split. Their results are reported in the "Discussion" section to assess whether the patient-level estimate is dominated by a single favourable partition.

3.4 Multi-Annotator Target Construction

To analyse annotation variability, we considered three consensus rules. The union labels a pixel as foreground if it is included by at least one radiologist, the intersection retains only pixels marked by all available radiologists, and majority voting retains pixels supported by the majority of annotations. These rules represent increasingly different definitions of the reference region and can therefore affect both target size and measured Dice.

Target construction is also coupled to case selection. Because we exclude samples below a minimum foreground size, different target rules can retain different subsets of images. We therefore evaluated both independently filtered variants and a controlled same-image subset. For the controlled experiment, the same 756 image/slice samples and the same patient-level split were used for all three target definitions, and a separate model was trained for each target rule. This design controls case selection while retaining the end-to-end effect of changing the target during both training and evaluation.

3.5 Metric Computation and Aggregation

For each evaluated sample i, prediction probabilities were binarised at a fixed threshold of 0.25, while reference masks were binarised at 0.5. The prediction threshold was kept unchanged across all experiments and was not adapted to individual test sets or to MSD. Let P_i and G_i denote the resulting masks. The per-sample Dice score is $d_i = \frac{2|P_i \cap G_i|}{|P_i|+|G_i|}$. Mean Dice, $D_{\text{mean}} = \frac{1}{N}\sum_i d_i$, gives every sample equal weight, whereas median Dice provides a robust summary of the typical case. Reference-size-weighted Dice, $D_{\text{weighted}} = \frac{\sum_i |G_i| d_i}{\sum_i |G_i|}$, gives greater influence to larger targets. Global pooled Dice, $D_{\text{global}} = \frac{2\sum_i |P_i \cap G_i|}{\sum_i |P_i| + \sum_i |G_i|}$, pools foreground counts before computing the score. We additionally report the failure rate $\frac{1}{N}\sum_i \mathbf{1}[d_i < 0.20]$. For mean Dice, 95% confidence intervals (CIs) were estimated using 1,000 bootstrap resamples over the evaluated samples.

4 Experiments and Results

The experiments examine four complementary aspects of the evaluation protocol: internal versus external behaviour under different split granularities, a full-slice control, target construction from multiple annotations, and sensitivity to the segmentation architecture. The training and inference settings are kept as consistent as possible within each paired comparison.

4.1 Internal LIDC-IDRI Performance versus External Generalisation

For the external experiment, 63 MSD Task06 training volumes yielded 1,613 lesion-containing axial slices. The reference mask was used to identify lesion-containing slices and centre the same lesion-focused ROI, after which the frozen LIDC-IDRI models were applied without fine-tuning. Thus, the experiment is a controlled cross-dataset segmentation stress test within known lesion regions rather than an end-to-end lesion localisation experiment. Both models were evaluated on the identical MSD samples.

Table 1 shows a clear internal–external reversal. On the internal LIDC-IDRI test set, the model trained with a random sample-level split obtains substantially higher mean, median, and global Dice than the model trained with patient-level separation. If only the internal benchmark were considered, the random split would therefore appear to produce the stronger model. On the external MSD samples, however, the direction reverses: the patient-level model achieves higher performance across all three reported aggregation strategies.

The 95% bootstrap CIs for mean Dice were 0.430 [0.410, 0.448] versus 0.518 [0.499, 0.538] internally, and 0.531 [0.517, 0.544] versus 0.464 [0.448, 0.479] on MSD for patient-level and sample-level training, respectively. The internal performance increase is therefore observed under substantial measured train–test contamination and does not translate into better cross-dataset performance. Because the internal protocols also induce different test-set compositions, we interpret the LIDC comparison primarily as an audit of the reported benchmark estimate; the matched MSD evaluation provides the cleaner same-test-set comparison of the two frozen models.

Table 1. Internal LIDC-IDRI and external MSD performance. "P/I" denotes train–test overlap in patient/image identifiers.

Training split	Overlap P/I	LIDC			MSD		
		Mean	Median	Global	Mean	Median	Global
Patient-level	0/0	0.430	0.425	0.390	**0.531**	**0.558**	**0.500**
Sample-level	568/724	**0.518**	**0.673**	**0.597**	0.464	0.479	0.416

4.2 Full-Slice Control

To test whether the difference between split protocols is driven only by lesion-centred cropping, we conducted an additional full-slice experiment. This setting is substantially harder because the lesion occupies a much smaller fraction of the model input. As shown in Table 2, performance decreases overall, but the sample-level split remains higher across all reported Dice variants and has a lower failure rate. The persistence of the gap indicates that the split-dependent effect is not solely an artefact of ROI cropping.

The patient-level median Dice of 0.000, together with a 0.550 failure rate, indicates a strongly zero-inflated failure pattern in this harder 2D setting. We therefore interpret this experiment as a control showing that the split-dependent difference persists without lesion-centred cropping, rather than as a competitive full-slice segmentation baseline.

Table 2. Full-slice control; the split-dependent gap remains visible without lesion-centred ROI cropping.

Split protocol	Mean	Median	Global	Vol.-weighted	Fail < 0.20
Patient-level	0.347	0.000	0.458	0.453	0.550
Sample-level	**0.408**	**0.315**	**0.555**	**0.558**	**0.489**

4.3 Target Construction from Multiple Annotations

When union, majority vote, and intersection were first evaluated on independently filtered subsets, the differences were large: mean Dice increased from 0.507 for union to 0.762 for majority vote and 0.790 for intersection, while the failure rate for Dice < 0.20 decreased from 0.349 to 0.038 and 0.026, respectively. These results, however, combine the effect of the target definition with the effect of retaining different image subsets.

To separate these factors, we repeated the comparison on the controlled 756-sample subset. Table 3 shows that the differences become much smaller when case selection is fixed: mean Dice varies only between 0.750 and 0.771, median Dice between 0.829 and 0.843, and failure rate between 0.025 and 0.037. Because this subset contains only images eligible under all three target definitions, its absolute Dice values should not be compared directly with the main LIDC benchmark; the relevant result is the within-subset comparison. The experiment indicates that an apparently large "target-construction effect" can in part be a case-selection effect.

Table 3. Multi-annotator target construction on the same 756 image/slice samples. A separate model was trained for each target rule using the same patient-level split.

Target	Mean Dice	Median Dice	Global Dice	Vol.-weighted Dice	Fail < 0.20
Union	0.769	0.831	0.828	**0.836**	**0.025**
Majority vote	0.750	0.829	0.810	0.833	0.037
Intersection	**0.771**	**0.843**	**0.819**	0.833	0.032

4.4 Architecture Effects under Different Split Protocols

To verify whether the observed pattern is specific to a single architecture, we repeated the patient-level versus sample-level comparison with U-Net, Attention U-Net, and a compact TransUNet-inspired variant. Table 4 shows that differences between architectures are comparatively small when the split protocol is fixed. In the patient-level setting, mean Dice ranges from 0.433 to 0.443, whereas all sample-level runs increase to approximately 0.52. The median Dice and failure rates show the same overall pattern.

This control suggests that, within the architectures studied here, the effect associated with the split protocol is larger than the differences between the models themselves. The architecture sweep used separate control runs from the U-Net checkpoints in Table 1, which explains the small difference in the reported patient-level U-Net value (0.440 versus 0.430). Architectural comparisons on LIDC-IDRI should therefore be interpreted together with split granularity and explicit leakage reporting.

Table 4. Effect of split protocol across separate architecture-control runs.

Architecture	Split	Mean Dice	Median Dice	Global Dice	Fail < 0.20
U-Net	Patient-level	0.440	0.454	0.437	0.434
	Sample-level	0.518	0.665	0.580	0.320
Attention U-Net	Patient-level	0.433	0.401	0.391	0.443
	Sample-level	0.523	0.659	**0.608**	0.310
TransUNet-variant	Patient-level	0.443	0.447	0.428	0.443
	Sample-level	**0.526**	**0.668**	0.604	**0.306**

5 Discussion

The main finding is that a random sample-level split yields higher internal LIDC-IDRI scores while exhibiting substantial measured train–test contamination, and that this apparent advantage does not transfer to the matched external MSD samples. In contrast, patient-level separation provides a more conservative internal estimate but the resulting model performs better externally. A repeated patient-level holdout control using two additional partitions yielded mean Dice values of 0.420 and 0.448 (mean 0.434; 1,000 test samples from 162 and 163 patients, respectively). This variation is smaller than the 0.088 difference between the main patient-level and sample-level internal mean Dice values, supporting the interpretation that the observed gap is not explained solely by one favourable patient partition.

The results also illustrate why segmentation performance cannot be fully characterised by a single Dice number. Mean, median, global pooled, and reference-size-weighted Dice emphasize different aspects of prediction quality. This is particularly relevant for lung nodules, where target sizes vary considerably and small lesions can strongly influence per-case scores. Reporting multiple aggregations together with a failure rate makes the distribution of outcomes more visible and reduces the risk that a single headline value dominates interpretation.

The multi-reader experiment highlights a related evaluation issue. Union, majority vote, and intersection define different reference masks, but they can also alter which samples pass foreground-based eligibility criteria. The large differences observed on independently filtered subsets shrink markedly once the same 756 images are used for all three rules. Thus, target construction and sample selection are coupled: a comparison of consensus strategies should report not only the rule itself, but also whether the evaluated case set changes as a consequence.

The scope of these conclusions is deliberately limited to the investigated LIDC-IDRI–MSD Task06 setting. LIDC nodules and MSD lung tumours differ in lesion characteristics, and the MSD experiment uses reference masks to define lesion-centred ROIs; it should therefore be interpreted as cross-dataset segmentation validation within known lesion regions rather than a clinical end-to-end pipeline. The audit quantifies patient and exact axial-slice image overlap, but lesion-level overlap and partially overlapping 2.5D windows were not separately enumerated.

6 Conclusion and Future Work

The higher internal Dice scores obtained with random sample-level splitting were observed under substantial measured train–test contamination and did not translate into better external performance, whereas patient-level splitting yielded stronger MSD performance. Metric aggregation and multi-reader target construction also changed the interpretation of reported results. Comparisons in this setting should therefore report split granularity and overlap statistics, target construction, sample eligibility, metric aggregation, and the exact external-validation protocol.

Future work will extend the analysis to additional datasets and fully 3D pipelines, including lesion-level leakage auditing, partially overlapping 2.5D windows, and evaluation strategies that represent multi-reader disagreement more explicitly.

Acknowledgements This research was supported by the project “Romanian Hub for Artificial Intelligence – HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 351416.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
5. Kocak, B., Klontzas, M.E., Stanzione, A., Meddeb, A., Demircioğlu, A., Bluethgen, C., Bressem, K., Ugga, L., Mercaldo, N., Díaz, O., et al.: Evaluation metrics in medical imaging ai: fundamentals, pitfalls, misapplications, and recommendations. European journal of radiology artificial intelligence p. 100030 (2025)
6. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M., Büttner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: Recommendations for image analysis validation. arxiv 2023. arXiv preprint arXiv:2206.01653 (2023)
7. Reinke, A., Tizabi, M.D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A.E., Rädsch, T., Sudre, C.H., Acion, L., Antonelli, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nature methods **21**(2), 182–194 (2024)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
9. Tampu, I., Eklund, A., Haj-Hosseini, N.: Inflation of test accuracy due to data leakage in deep learning-based classification of oct images. scientific data, 9 (1), 580 (2022)
10. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine **5**(1), 48 (2022)
11. Yagis, E., Atnafu, S.W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., Diciotti, S.: Effect of data leakage in brain mri classification using 2d convolutional neural networks. Scientific reports **11**(1), 22544 (2021)
12. Yang, H., Shen, L., Zhang, M., Wang, Q.: Uncertainty-guided lung nodule segmentation with feature-aware attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 44–54. Springer (2022)
13. Yang, H., Wang, Q., Zhang, Y., An, Z., Liu, C., Zhang, X., Zhou, S.K.: Lung nodule segmentation and uncertain region prediction with an uncertainty-aware attention mechanism. IEEE Transactions on Medical Imaging **43**(4), 1284–1295 (2023)
14. Yu, A.C., Mohajer, B., Eng, J.: External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiology: Artificial Intelligence **4**(3), e210064 (2022)