

A candidate-centred heterogeneous model ensemble improves semi-supervised segmentation of 3–<6 mm pulmonary nodules over a single enhanced Mean Teacher: internal and external testing

Hao Li

Beijing University of Chemical Technology

Zhenhua Zhu

Beijing University of Chemical Technology

Xiaoxiao Wang

Beijing University of Chemical Technology

Yong Cheng

chengyong@mail.buct.edu.cn

Beijing University of Chemical Technology

Fang Li

Beijing University of Chemical Technology

Research Article

Keywords: pulmonary nodule, computed tomography, semi-supervised learning, medical image segmentation, Mean Teacher, deep ensemble, external testing

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10931623/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

A candidate-centred heterogeneous model ensemble improves semi-supervised segmentation of 3–<6 mm pulmonary nodules over a single enhanced Mean Teacher: internal and external testing

Hao Li¹, Zhenhua Zhu¹, Xiaoxiao Wang¹, Yong Cheng^{1*}, Fang Li^{1*}

¹Department of Computer Science and Technology, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, P.R. China.

*Corresponding author(s). E-mail(s): chengyong@mail.buct.edu.cn; lifang@mail.buct.edu.cn;

Abstract

Background: With limited voxel-level annotations, a single Mean Teacher may not address both foreground omission and boundary expansion in candidate-centred segmentation of small pulmonary nodules. We evaluated a development-locked, failure-mode-guided heterogeneous ensemble for semi-supervised segmentation of 3–<6 mm nodules.

Methods: Candidate-centred $48 \times 48 \times 48$ CT patches were analysed with 35 of 350 training CT scans (10%) providing voxel annotations. An enhanced Mean Teacher and three complementary recall- or precision-oriented models were trained independently. Probability weights and the binarisation threshold were fitted only in the development selection set. The locked ensemble was tested across three training seeds in an internal held-out set (51 CT scans, 62 nodules) and external LNDb (176 CT scans, 436 nodules), without LNDb fine-tuning or recalibration. Paired effects were evaluated using CT-cluster bootstrap confidence intervals and two-sided Monte Carlo CT-cluster sign-flip randomisation tests. Homogeneous four-model ensembles were secondary model-count-matched attribution analyses; the equal-weight analysis was exploratory.

Results: Relative to the single enhanced Mean Teacher, the heterogeneous ensemble improved Dice by 0.01488 internally (95% CI, 0.00665–0.02347; Holm-adjusted $P = \mathbf{0.00110}$) and 0.01118 externally (95% CI, 0.00314–0.01892; Holm-adjusted $P = \mathbf{0.00673}$). Dice improved in 77.4% and 71.3% of internal and external nodules, respectively, and all three evaluated seeds showed positive

effects in both test sets. The operating balance shifted towards recall. Attribution analyses did not establish a cross-cohort-stable advantage over homogeneous four-model ensembles.

Conclusions: Relative to a single enhanced Mean Teacher, the development-locked heterogeneous ensemble yielded estimated mean Dice gains of 0.01488 internally and 0.01118 externally, with both confidence intervals excluding zero and concordant effect directions across the three evaluated seeds. These results support a modest ensemble benefit under the present protocol, but do not establish that the population-level Dice improvement exceeds 0.01 or that heterogeneous training objectives provide an independent benefit beyond general deep-ensemble effects. Prospectively locked homogeneous controls and evaluation in new independent cohorts are required.

Keywords: pulmonary nodule, computed tomography, semi-supervised learning, medical image segmentation, Mean Teacher, deep ensemble, external testing

Article type: Research Article

1 Background

Three-dimensional segmentation of pulmonary nodules provides quantitative measurements of volume, equivalent diameter, morphology, and longitudinal change, all of which are closely related to nodule risk assessment and clinical management [1–5]. Architectures based on U-Net, three-dimensional convolution, and residual learning have advanced automated nodule segmentation [6–11]. However, high-quality voxel-level annotations still depend on experienced readers, and slice-by-slice delineation of three-dimensional nodules is particularly labour-intensive. Semi-supervised learning addresses this limitation by combining limited annotated data with unannotated images. Foundational reviews and the Mean Teacher framework established the basis of teacher–student learning [12, 13]. Subsequent studies have investigated uncertainty-aware and dual-task consistency [14, 15], shape constraints and multi-view or pyramid consistency [16–19], and triple uncertainty, contrastive representation learning, ambiguity-aware selection, and cross-pseudo supervision [20–23]. In CT lesion and pulmonary nodule applications, uncertainty pairing, knowledge distillation, and coarse-grained annotations have also been explored [24, 25]. Because these studies differ in datasets, annotation budgets, and evaluation protocols, they provide task-relevant evidence but are not used for direct Dice-based ranking against the present study.

Small nodules, however, are not simply scaled-down versions of larger nodules. Their foreground regions contain few voxels, such that the addition or removal of a small number of boundary voxels can substantially alter Dice. The foreground proportion is further reduced when a nodule is represented within a fixed-size candidate patch. Although a fixed positive pseudo-label threshold may improve pseudo-label precision, it may systematically remove low-probability but spatially contiguous regions of true foreground, thereby increasing foreground omission in small nodules. Meanwhile, multi-reader annotations in LIDC-IDRI demonstrate that both nodule identification and

boundary delineation can involve legitimate inter-reader disagreement [26–32]. Treating model uncertainty and annotation ambiguity uniformly as noise may therefore discard supervision from learnable regions while imposing excessive consistency constraints in regions of reasonable disagreement.

Motivated by these considerations, we initially developed a size-aware threshold and dual-uncertainty control strategy. The positive acceptance threshold was reduced for candidates with small teacher-predicted foreground extents, while Monte Carlo dropout variability and a multi-reader ambiguity proxy represented distinct sources of uncertainty. Although the design had a clear mechanistic rationale, formal development experiments provided an important counterexample: improved local pseudo-label recall did not translate into higher final Dice for small nodules. Nodule-level error auditing further showed that small-nodule errors were not limited to foreground omission. Lowering the threshold recovered some missing foreground but also increased boundary false positives, whereas uniform gating could suppress both forms of correction. The research question therefore shifted from identifying a single improved threshold to preserving complementary recall- and precision-oriented solutions within one framework and locking their combination using a CT-disjoint development set.

In response, we constructed recall- and precision-oriented expert corrections. The recall-oriented branch assigned greater Tversky weight to false negatives to reduce foreground omission, whereas the precision-oriented branch assigned greater weight to false positives to suppress boundary expansion. A voxel-wise gating network selected between the two corrections at a local scale, while balancing and baseline-anchoring constraints were used to prevent gating collapse and excessive deviation from the baseline prediction. Rather than selecting a winner using test data, probability weights and the binarisation threshold for the four-model ensemble were fitted only on the development selection set and frozen before internal and external testing.

Independently trained deep models can exploit predictive diversity arising from different initialisations and optimisation trajectories, providing a general basis for probability averaging and improved stability [33]. Guided by the failure audit, the final heterogeneous ensemble comprised one enhanced Mean Teacher and three independently trained complementary models with directional error objectives. The primary question was whether, with voxel-level annotations available for 10% of CT scans, the development-locked ensemble could produce a positive Dice gain over a single enhanced Mean Teacher in both an internal CT-level held-out test set and external LNDb testing without fine-tuning. An estimated mean Dice gain of at least 0.01 was used as a descriptive development target rather than as a formally tested superiority margin. Concordant effect directions across the three evaluated training seeds were assessed as a descriptive stability criterion. Model-count-matched homogeneous Mean Teacher ensembles were retained as secondary attribution analyses to assess whether the observed benefit could be explained by generic deep-ensemble effects; superiority over these controls was not the primary endpoint.

This study was restricted to candidate-centred three-dimensional segmentation of pulmonary nodules measuring 3-<6 mm, with voxel-level annotations available for 10% of CT scans. The primary conclusions were based on paired comparisons of the same nodules under the same data protocol. Training computation was not fully matched

across methods. Candidate-centred segmentation was not presented as a whole-CT nodule detection system, and results from previous studies using different datasets or evaluation protocols were not treated as directly comparable performance rankings.

2 Methods

2.1 Study design and data separation

2.1.1 Study design, data sources, and data separation

This retrospective computational study used de-identified, publicly available data. The internal data were obtained from low-dose chest CT scans in LUNA16 [34]. Nodule contours were reconstructed from the corresponding multi-reader LIDC-IDRI XML annotations [26, 27], and the original CT images were distributed through The Cancer Imaging Archive [35]. The development training set with available voxel-level contours comprised 350 CT scans and 607 nodules. The development selection set comprised 87 CT scans and 145 nodules, including 69 nodules measuring 3–<6 mm from 47 CT scans. The denominator for the 10% CT-level voxel-annotation budget was restricted to the 350 development training CT scans. Using label-selection seed 0, we fixed a labelled subset of 35 CT scans containing 57 candidate nodules. The unlabelled candidate cache was generated from the 709 CT scans in the original LUNA16 training partition. After excluding the 87 development selection CT scans and the 35 labelled CT scans, the strictly unlabelled pool contained 16,234 candidate patches from 587 CT scans. Of these, 315 were development training CT scans for which the available voxel-level contours were deliberately not used, whereas 272 belonged to the original training partition but were not included in either the contour-based development training set or the development selection set. Consequently, 35 labelled and 587 strictly unlabelled CT scans, totalling 622 mutually exclusive CT scans, contributed to parameter updates. The additional 87 CT scans were used only for development selection, accounting for all 709 CT scans in the original training partition. A code-level audit confirmed that the seriesuid values of the labelled subset, strictly unlabelled pool, development selection set, and internal test set were mutually exclusive as specified. The internal CT-level held-out test set contained 62 nodules measuring 3–<6 mm from 51 CT scans and was not used for training, checkpoint selection, ensemble-weight fitting, or threshold selection. Publicly derived manifests contained only the LUNA16 SeriesInstanceUID or LNDb CT identifier and did not include patient identifiers that could link multiple scans from the same individual. We therefore could not determine whether a participant contributed more than one CT scan. The CT scan was the highest verifiable clustering unit, and data separation, bootstrap resampling, and sign-flip randomisation were performed using CT identifiers. No patient-level analysis is claimed.

Candidate centres were specified using the world coordinates (x, y, z) from the public candidate lists without first rounding them to voxel indices. SimpleITK constructed a $48 \times 48 \times 48$ reference grid with 1.0-mm isotropic spacing centred on each physical coordinate. CT intensities were resampled using linear interpolation, with regions outside the original image filled with -1000 HU; contour masks were resampled using nearest-neighbour interpolation. Arrays were stored in NumPy (z, y, x) axis order.

Table 1 Data flow, analytical purpose, and separation of study subsets

Cohort/ subset	Source	Size	Purpose	Separation rule
Label-budget set	LUNA16 trian	350 CT; 607 nodules	Define the 35-CT labelled subset	Excluded from selection and internal testing; 315 CTs entered the unlabelled pool
Labelled subset	Development training	35 CTs; 57 nodules	Supervised loss	Fixed using label-selection seed 0
Unlabelled pool	LUNA16 candidate cache	587 CTs; 16,234 patches	Consistency training	315 development-training CTs plus 272 other training CTs; all labelled, selection, and test CTs excluded
Development selection	LUNA16 train	87 CT; 145 nodules	Select checkpoints, ensemble weights, and threshold	No parameter updates
Internal held-out test	LUNA16 test	51 CT; 62 small nodules	Locked internal evaluation	No development or selection use
External LNDb	LNDb	176 CT; 436 small nodules	Locked external evaluation	No fine-tuning or recalibration

Intensities were clipped to $[-1000, 400]$ HU and linearly normalised to $[0, 1]$ using $(\text{HU} + 1000)/1400$. Internal nodule diameters were obtained from the LUNA16 annotation metadata. For LNDb, equivalent spherical diameter was calculated from official nodule volume V as $d = (6V/\pi)^{1/3}$. The internal consensus mask included voxels endorsed in at least two of four available reader sessions, whereas external testing used the official LNDb masks. The evaluated task was therefore candidate-centred segmentation of nodules measuring $3\text{--}6$ mm and did not include candidate detection or whole-CT inference. The unlabelled cache was generated from world-coordinate candidates in LUNA16 candidates_V2.csv. Before reading the candidate-class field, a SHA-256 sorting key was constructed from seriesuid, the three-dimensional coordinates, and fixed seed 42. This deterministic ordering retained up to 20,000 candidates from the 709 training CT scans. No additional filtering used candidate class, voxel-level contour availability, or nodule diameter, and coordinates were not separately deduplicated. After excluding the development selection and labelled CT scans, candidates per CT ranged from 2 to 68 (median, 26; interquartile range, 19–35). Candidate patches near image boundaries followed the same extraction procedure, with regions outside the image filled with -1000 HU.

Model development followed a unidirectional information flow. Development training data updated network parameters; the development selection set determined candidate methods and checkpoints and was used to fit ensemble probability weights and the binarisation threshold; and the internal CT-level held-out and external LNDb test sets were used only for final evaluation after locking. Neither test set contributed to model selection, post-processing selection, threshold selection, or stopping-epoch selection. Study subsets, analytical purposes, and separation rules are summarised in Table 1.

Small nodules were defined as having an equivalent diameter of 3–<6 mm. The CT scan was the independent clustering unit.

External testing used publicly available multi-reader LNDb annotations [36]. Candidate patches were extracted at the officially provided nodule centres for 436 nodules measuring 3–<6 mm from 176 CT scans. No LNDb data were used for fine-tuning, threshold recalibration, post-processing selection, or model selection. Both internal and external tests used the models, probability weights, and binarisation threshold frozen during development.

2.1.2 Multi-reader consensus and ambiguity supervision

For nodule i , let M_i denote the total number of reading sessions for its CT scan, and let $r_{ij}(v) \in \{0, 1\}$

denote the foreground vote of reading session j at voxel v . If an actual reading session did not provide a contour matched to the target nodule, that session was represented by a zero mask throughout the candidate patch rather than being removed from the denominator. The voxel-wise foreground proportion was defined as

$$\pi_i(v) = \frac{1}{M_i} \sum_{j=1}^{M_i} r_{ij}(v).$$

Each included internal CT scan was required to have exactly four reading sessions; therefore, $M_i = 4$ for all internal samples. Stratification by three or four matched contours referred only to the number of sessions that provided a contour matched to the target nodule. The foreground-proportion denominator remained four in both groups. The consensus segmentation target used a threshold of at least two votes:

$$y_i(v) = \mathbf{1} \left\{ \sum_{j=1}^{M_i} r_{ij}(v) \geq 2 \right\}.$$

Ambiguity supervision $a_i(v)$ was defined using normalised Bernoulli entropy. The implementation used natural logarithms and divided by $\ln(2)$, which is equivalent to Bernoulli entropy calculated using base-2 logarithms. Its complete piecewise definition was

$$a_i(v) = \begin{cases} -\frac{\pi_i(v) \ln \pi_i(v) + [1 - \pi_i(v)] \ln [1 - \pi_i(v)]}{\ln 2}, & 0 < \pi_i(v) < 1, \\ 0, & \pi_i(v) \in \{0, 1\}. \end{cases}$$

When $\pi_i(v)$ was 0 or 1, $a_i(v)$ was set directly to 0 to avoid evaluating $\ln(0)$. Thus, $a_i(v) \in \{0, 1\}$, where 0 indicates complete agreement among the four reading sessions and 1 corresponds to two foreground and two background votes. Nodules with three matched contours still used four actual reading sessions in the denominator, with the unmatched session contributing a zero vote; the number of matched contours did not replace M_i . This normalisation placed ambiguity supervision for all nodules on the same numerical scale and gave a consistent interpretation to the predicted ambiguity probability ($a(v) = \sigma(z_a(v))$) and its gating threshold of 0.50.

2.2 Model development and training

2.2.1 Network architecture and supervised objectives

The overall architecture of the heterogeneous model ensemble is shown in Fig. 1.

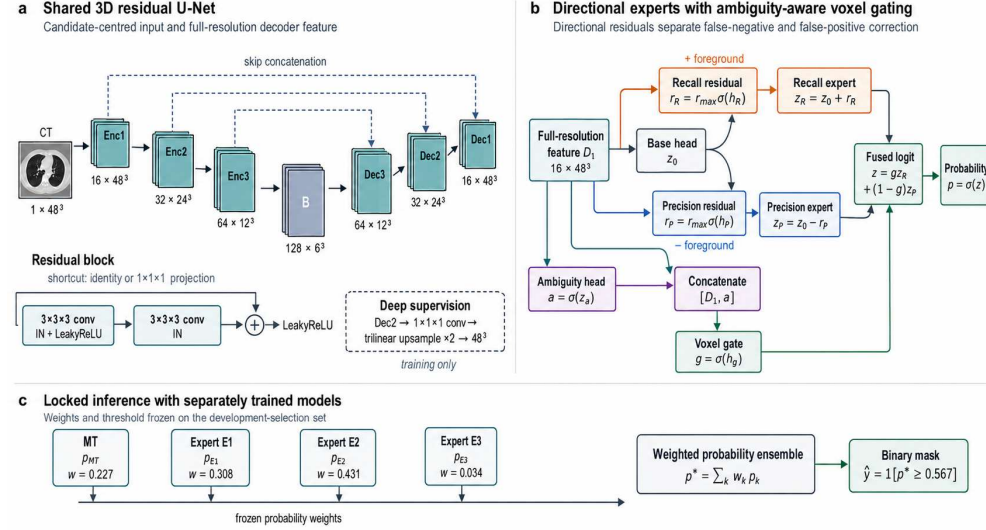


Fig. 1 Heterogeneous model ensemble framework. (a) Four component models used the same three-dimensional residual U-Net backbone and were trained independently. (b) Complementary experts combined the ambiguity logit Z_a with decoder features to generate voxel-wise gates, while recall- and precision-oriented residuals corrected the prediction in opposite directions. (c) Probabilities from the enhanced Mean Teacher, E1, E2, and E3 were combined using development-locked weights and a development-locked threshold.

The backbone was a three-dimensional residual U-Net. Encoder channel widths were 16, 32, 64, and 128. Each residual block contained two $3 \times 3 \times 3$ convolutions, affine InstanceNorm3d, and LeakyReLU with a negative slope of 0.01; a $1 \times 1 \times 1$ projection shortcut was used when the channel width or stride changed. Downsampling was performed by a stride-2 convolution in the first layer of each residual block. The decoder first used $2 \times 2 \times 2$ transposed convolution for upsampling, concatenated the result with the corresponding encoder features, and passed the concatenated features through a residual block. The bottleneck used Dropout3d ($p = 0.2$). The main segmentation and ambiguity heads operated on the full-resolution D1 features (16 channels; 48^3). A deep-supervision head operated on D2 features (32 channels; 24^3) through a $1 \times 1 \times 1$ convolution followed by two-fold trilinear upsampling. All output heads produced logits.

To distinguish logits from probabilities, $\sigma(\cdot)$ denotes the sigmoid function. Binary cross-entropy in the implementation was computed using `binary_cross_entropy_with_logits`. For a voxel-wise logit z and soft target $t \in \{0, 1\}$,

$$\ell_{\text{BCELogit}}(z, t) = -t \ln \sigma(z) - (1 - t) \ln [1 - \sigma(z)].$$

For voxel domain Ω , the mean loss was denoted by

$$L_{\text{BCELogit}}(z, t) = \frac{1}{|\Omega|} \sum_{v \in \Omega} \ell_{\text{BCELogit}}[z(v), t(v)].$$

The soft Dice loss first applied the sigmoid function to the logits and used the implementation smoothing constant of 1:

$$L_{\text{Dice}}(z, y) = 1 - \frac{2 \sum_{v \in \Omega} \sigma[z(v)]y(v) + 1}{\sum_{v \in \Omega} \sigma[z(v)] + \sum_{v \in \Omega} y(v) + 1}.$$

These losses were calculated per sample and then averaged across the minibatch. The main and auxiliary segmentation outputs used the same equally weighted combination:

$$L_{\text{seg}}(z, y) = 0.5L_{\text{BCELogit}}(z, y) + 0.5L_{\text{Dice}}(z, y).$$

The deep-supervision term had weight 0.3 in the total supervised loss. The ambiguity branch used binary cross-entropy with logits and had weight 0.2.

The implementation retained a candidate-presence output head. However, all labelled candidate-centred patches in the final configuration were positive candidates, and the loss weight for this head was 0. It was therefore neither an optimisation objective nor a result-bearing contribution, and no candidate-presence classification conclusion is supported by this output.

2.2.2 Student–teacher semi-supervised learning

Student parameters are denoted by θ_s and teacher parameters by θ_t . The teacher had the same architecture as the student but did not receive gradients. After each optimisation step, teacher parameters were updated by exponential moving average:

$$\theta_t \leftarrow \alpha \theta_t + (1 - \alpha) \theta_s,$$

where $\alpha = 0.99$. For a labelled patch (x_i, y_i, a_i) , let z , z_{aux} , and z_a denote the main segmentation, auxiliary segmentation, and ambiguity logits, respectively. The supervised objective was

$$L_{\text{sup}} = L_{\text{seg}}(z, y) + 0.30 L_{\text{seg}}(z_{aux}, y) + 0.20 L_{\text{BCELogit}}(z_a, a)$$

Here, L_{seg} is the Dice–binary-cross-entropy combined segmentation loss defined above, with equal weights for logit binary cross-entropy and soft Dice loss. L_{seg} is used consistently throughout the manuscript. The variable a denotes normalised Bernoulli-entropy ambiguity supervision, and L_{BCELogit} denotes mean binary cross-entropy applied directly to logits. The ambiguity probability was defined as $\hat{a}(v) = \sigma(z_a(v))$; $\hat{a}$ is a probability, whereas z_a always denotes the pre-sigmoid logit.

For an unlabelled patch, the teacher generated foreground probabilities $p_t^{(k)}(v)$ during four stochastic forward passes with dropout. To avoid ambiguity in overbar notation across word-processing software, the mean teacher foreground probability is denoted by $\mu_t(v)$:

$$\mu_t(v) = \frac{1}{K} \sum_{k=1}^K p_t^{(k)}(v).$$

Model uncertainty was estimated using Monte Carlo dropout [37–39]. Segmentation-probability variance was calculated across the four stochastic forward passes. The ambiguity-branch logit from pass k was denoted by $z_{a,t}^{(k)}$, and its mean probability was defined separately as $q_t(v) = K^{-1} \sum_k \sigma(z_{a,t}^{(k)}(v))$. The quantity q_t was used only for ambiguity-based acceptance or downweighting during the exploratory stage. By contrast, expert gating used the single-forward-pass probability $\hat{a} = \sigma(z_a)$; the two quantities were not interchanged.

$$u(v) = \frac{1}{K} \sum_{k=1}^K \left[p_t^{(k)}(v) - \mu_t(v) \right]^2.$$

Here, $K = 4$, and $u(v)$ is referred to as model predictive uncertainty.

The same augmented unlabelled patch was provided to the student and teacher. No additional weak–strong augmentation views or spatial alignment procedure was used. Teacher signals were used only to construct acceptance masks, continuous weights, and soft supervision targets; student predictions did not contribute to these audit quantities.

2.2.3 Exploratory size-aware threshold

The first stage tested the original mechanistic hypothesis. Because true nodule size was unavailable for unlabelled patches, foreground extent was approximated solely from the teacher prediction. This quantity represented the foreground extent currently predicted by the teacher; it was not equivalent to true diameter and did not use manual annotations. The following equations define its calculation:

$$F(x) = \sum_{v \in \Omega} \max\{2[\mu_t(v) - 0.5], 0\}.$$

When the teacher-predicted foreground extent was small, the positive threshold was smoothly reduced from the base threshold $\tau_0 = 0.90$ according to

$$\tau(x) = \tau_0 - 0.15 \exp[-F(x)/128].$$

This function constrained the positive acceptance threshold to 0.75–0.90. Smaller teacher-predicted extents produced thresholds closer to 0.75, retaining more potential small-nodule foreground, whereas larger extents returned the threshold towards 0.90. The extent proxy controlled only which voxels entered the soft consistency loss. It was not treated as true size, and the binarised probability was not used as a hard training target. The negative threshold remained 0.10.

2.2.4 Dual-uncertainty gating, weighting, and consistency loss

The mean probability from repeated forward passes through the teacher ambiguity branch was $q_t(v) = K^{-1} \sum_k \sigma(z_{a,t}^{(k)}(v))$. During the exploratory stage, the size-aware

threshold, predictive variance $u(v)$, and $q_t(v)$ were three parallel teacher-side control signals. The ambiguity acceptance condition was $q_t(v) \leq 0.50$; no additional sigmoid transformation was applied to q_t :

$$m(v) = \mathbf{1}\{[\mu_t(v) \geq \tau(x) \vee \mu_t(v) \leq 0.10] \wedge u(v) \leq 0.02 \wedge q_t(v) \leq 0.50\}.$$

Confidence, model predictive uncertainty, and the ambiguity proxy were then combined to form a continuous weight:

$$w(v) = m(v) \frac{\max[\mu_t(v), 1 - \mu_t(v)] - 0.5}{0.5} \exp[-u(v)/0.02] [1 - q_t(v)].$$

The three factors represented teacher confidence, predictive stability, and annotation-ambiguity proxy, respectively. Voxel weights decreased as confidence declined, dropout variance increased, or the ambiguity proxy increased. The aim was to attenuate unreliable consistency continuously rather than discard all difficult boundary voxels. Subsequent failure auditing nevertheless showed that when foreground omission and boundary expansion coexisted, this unidirectional downweighting could not provide corrections in opposing directions.

$$L_u = \frac{\sum_{v \in \Omega} w(v) \ell_{\text{BCELogit}}[z_s^u(v), \text{sg}(\mu_t(v))]}{\sum_{v \in \Omega} w(v) + \varepsilon}.$$

The total loss was $L = L_{\text{sup}} + \lambda_u(t) L_u$. The implementation used a sigmoid ramp-up rather than a linear schedule:

$$\lambda_u(t) = \exp\left\{-5 [1 - \min(t/10, 1)]^2\right\}.$$

The schedule gradually strengthened the unlabelled-data constraint and reached weight 1 after epoch 10. Here, t denotes the training epoch and $\lambda_u(t)$ the consistency-loss weight. The student was updated by backpropagation, whereas the teacher was updated only through the exponential moving average. This initial framework was retained for problem localisation and mechanistic auditing rather than presented as the final superior method.

2.2.5 Complementary experts and locked probability ensemble

Let z_0 denote the baseline segmentation logit, and let the recall- and precision-oriented residuals be $r_R = r_{\text{max}} \sigma(h_R)$ and $r_P = r_{\text{max}} \sigma(h_P)$, respectively. The corrected logits were $z_R = z_0 + r_R$ and $z_P = z_0 - r_P$. The ambiguity probability from a single forward pass was $\hat{a} = \sigma(z_a)$. The gate was generated by concatenating the final decoder feature $D1$ with $\hat{a}$: $g = \sigma(h_g[D1, \hat{a}])$; no repeated normalisation $\sigma(\hat{a})$ was applied. The fused logit was $z = g z_R + (1 - g) z_P$. E1 and E2 used $r_{\text{max}} = 2.0$, whereas E3 used $r_{\text{max}} = 0.5$.

For a sample,

$$p = \sigma(z), \quad y \in \{0, 1\},$$

$$TP = \sum_v p(v)y(v), \quad FP = \sum_v p(v)[1 - y(v)], \quad FN = \sum_v [1 - p(v)]y(v).$$

The implementation used Tversky loss with a smoothing constant of 1 [40]:

$$L_T(z, y; \alpha, \beta) = 1 - \frac{TP + 1}{TP + \alpha FP + \beta FN + 1}.$$

The parameter α weights false positives (FP), and β weights false negatives (FN). The recall-oriented expert used $(\alpha_R, \beta_R) = (0.30, 0.70)$, and the precision-oriented expert used $(\alpha_P, \beta_P) = (0.70, 0.30)$; their losses are denoted by L_R and L_P , respectively.

Gate supervision was calculated only at supported voxels within the consensus boundary band $B = \text{dilate}(y, 2) \setminus \text{erode}(y, 2)$, where dilate and erode denote three-dimensional morphological dilation and erosion with a radius of two voxels, and $\setminus$ denotes set difference. Let $c(v)$ denote the number of reader foreground votes. The disputed-background set $D = \{v \in B : y = 0, c = 1\}$ had a soft gate target of 0.5, whereas the definite-outer set $O = \{v \in B : y = 0, c = 0\}$ had target 0. For E1, the positive-support set was $W = \{v : y = 1, c = 2\}$, with target 1. E2 and E3 replaced W with the complete foreground within the boundary band, $P = \{v \in B : y = 1\}$.

Let S^+ , D , and O denote the positive-support, disputed-background, and definite-outer sets, and let ℓ_{BCE} denote voxel-wise binary cross-entropy with logits. The E1 gate loss was

$$L_g = \text{mean}_i \left\{ \frac{\sum_{v \in (S^+ \cup D \cup O)} \ell_{\text{BCE}}(h_g(v), t_g(v))}{\max(|S^+ \cup D \cup O|, 1)} \right\}.$$

E2 and E3 did not add a separate gate-balancing regulariser. Instead, the three classes were balanced according to the minibatch aggregation implemented in code. For $C \in \{S^+, D, O\}$, let C_i be the voxels of class C in sample i , $I_C = \{i : |C_i| > 0\}$, and $S^+ = \{C : |I_C| > 0\}$. If C_i was empty, sample i did not contribute to the gate loss for class C . If a class was absent from the entire minibatch, it was also excluded from the class average.

$$\ell_C = \frac{1}{|I_C|} \sum_{i \in I_C} \left[\frac{1}{|C_i|} \sum_{v \in C_i} \ell_{\text{BCE}}(h_g(v), t_g(v)) \right].$$

The off-boundary agreement constraint between experts was

$$L_{ag} = \frac{1}{N_L} \sum_{i=1}^{N_L} \frac{\sum_{v \in \Omega_i \setminus B_i} [p_{R,i}(v) - p_{P,i}(v)]^2}{\max(|\Omega_i \setminus B_i|, 1)}.$$

For E3, the baseline-prediction anchoring term was $L_{\text{anchor}} = L_{\text{seg}}(z_0, y)$. During backpropagation of L_R and L_P , the implementation respectively used

$$\text{stopgrad}(z_0) + (z_R - z_0) \quad \text{and} \quad \text{stopgrad}(z_0) + (z_P - z_0).$$

These expressions were numerically unchanged, but the expert losses did not update the baseline head through z_0 ; the anchoring term directly constrained the baseline prediction.

Let $L_{core}(z) = L_{seg}(z, y) + 0.30 L_{seg}(z_{aux}, y) + 0.20 L_{BCELogit}(z_a, a)$, and let L_u denote the unlabelled consistency term. The complete losses for the three complementary models were

$$\begin{aligned} L_{E1} &= L_{core} + \lambda_u(t) L_u + 0.05 (L_R + L_P) + 0.05 L_g + 0.01 L_{ag}, \\ L_{E2} &= L_{core} + \lambda_u(t) L_u + 0.05 (L_R + L_P) + 0.05 L_{g,bal} + 0.01 L_{ag}, \\ L_{E3} &= L_{core} + \lambda_u(t) L_u + 0.05 (L_R + L_P) + 0.05 L_{g,bal} + 0.01 L_{ag} + 0.50 L_{anchor}. \end{aligned}$$

The coefficient $0.05(L_R + L_P)$ arose from $0.10 \times 0.5(L_R + L_P)$ in the implementation. Apart from the stated differences, E1–E3 used identical data, semi-supervised terms, and optimisation settings. Other candidate losses, including size-aware Tversky, boundary-shell, excess-volume, and distance supervision, were disabled in the final configurations.

The enhanced Mean Teacher was defined as a Mean Teacher with the dual-expert branch disabled but with deep supervision and the ambiguity auxiliary head retained: $L_{enhMT} = L_{core} + \lambda_u(t) L_u$. The teacher was an exponential moving average of the student parameters with decay 0.99. It performed four Monte Carlo dropout passes on the same augmented unlabelled patch, and the segmentation probabilities were averaged to obtain μ_t . The enhanced Mean Teacher did not use predictive variance or ambiguity probability for exclusion and did not lower the acceptance threshold. It accepted only voxels for which $\mu_t \geq 0.90$ or $\mu_t \leq 0.10$, used $(w(v) = 2|\mu_t(v) - 0.5|)$ as a continuous confidence weight, and used the teacher soft probability as the binary-cross-entropy target. The consistency weight λ_u followed a sigmoid ramp-up to 1 over the first 10 epochs.

The enhanced Mean Teacher, E1, E2, and E3 were trained independently from random initialisation; the experts were not obtained by fine-tuning the enhanced Mean Teacher. Optimisation used AdamW with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-4} . ReduceLROnPlateau was scheduled using teacher Dice for 3-<6 mm nodules in the development selection set (factor, 0.5; patience, 4). Training used at most 80 epochs, early-stopping patience of 15, batch size 2, and 300 updates per epoch. Each spatial axis was flipped with probability 0.5, and Gaussian noise with standard deviation 0.01 was added with probability 0.3. The actual epoch counts for seeds 0/1/2 were 23/32/36 for the enhanced Mean Teacher, 27/31/27 for E1, 30/27/29 for E2, and 22/27/29 for E3. The corresponding best-teacher checkpoints occurred at epochs 8/17/21, 12/16/12, 15/12/14, and 7/12/14. Checkpoints were selected only by development-selection Dice for 3-<6 mm nodules. Training was recorded using Python 3.12.10, PyTorch 2.12.0+cu130, CUDA 13.0, and an NVIDIA GeForce RTX 4060 Laptop GPU (8,188 MiB; driver 595.79). Statistical analysis and plotting used Python 3.12.10, NumPy 2.5.0, pandas 3.0.5, SciPy 1.18.0, and Matplotlib 3.11.1. Within outer run $s \in \{0, 1, 2\}$, the enhanced Mean Teacher, E1, E2, and E3 were run in separate training processes using the same numerical seed s . Model initialisation, RandomSampler, and

augmentation generators were reset to s in each process. This paired design controlled the random stream so that member differences primarily reflected the loss and gating objectives rather than additional seed offsets.

2.2.6 Training implementation and semi-supervised sampling

One epoch was defined by 300 batches from the unlabelled DataLoader. Each optimiser update paired one unlabelled batch with one labelled batch; both had nominal batch size 2. At the candidate-patch level, the strictly unlabelled pool used RandomSampler with replacement to draw 600 patches per epoch. Candidate-class weighting, foreground mining, and CT balancing were disabled. The 57 labelled candidate patches were shuffled without weighting, iterated without replacement, reshuffled after exhaustion, and cycled as required. Incomplete final batches were retained, resulting in 590 labelled patch presentations per epoch. Candidate patches from multi-nodule CT scans were treated as separate training samples, without CT-equalised sampling. DataLoader used `num_workers = 0`. The unlabelled sampler and DataLoader generators used the corresponding training seed, whereas selection of the 35 labelled CT scans remained fixed by label-selection seed 0 for all three training seeds. The enhanced Mean Teacher, E1, E2, E3, and standard Mean Teacher controls used the same sampling protocol.

The heterogeneous probability ensemble on the development selection set was defined as $p^* = \sum_k w_k p_k$, where p_k denotes the probabilities from the enhanced Mean Teacher, E1, E2, and E3, respectively, with $w_k \geq 0$ and $\sum_k w_k = 1$. The locked weights were 0.226690, 0.308426, 0.430849, and 0.034035, respectively, and the binarisation threshold was $\tau^* = 0.567185$. Thus, the ensemble was a probability-level weighted average of four independently trained models. The balanced-gating E2 received the largest weight, while the enhanced Mean Teacher provided a stable anchor.

Fitting used only the $N = 69$ nodules measuring $3 < 6$ mm in the development selection set and predictions from the four seed-0 members. Each member was loaded from its best teacher checkpoint. A checkpoint replaced the previous best only when development-selection Dice increased strictly; ties therefore retained the earlier epoch. Weights were parameterised by a softmax to form a non-negative continuous simplex and were initialised equally; the threshold was initialised at 0.55. Let $s_i(v) = \text{sigmoid}(40[p_i^*(v) - \tau])$. The continuous fitting objective was $L_{\text{fit}} = -N^{-1} \sum_i \frac{2 \sum_v s_i(v) y_i(v) + 10^{-6}}{\sum_v s_i(v) + \sum_v y_i(v) + 10^{-6}}$. Adam (learning rate 0.03; 500 steps; fitting seed 0) jointly updated the weight logits and τ , with τ constrained to $[0.30, 0.80]$. After continuous optimisation gave $\tau_{\text{soft}} = 0.583185$, the weights were fixed and hard Dice was scanned at increments of 0.001 over $[\max(0.30, \tau_{\text{soft}} - 0.15), \min(0.80, \tau_{\text{soft}} + 0.15)) = [0.433185, 0.733185]$. Candidate thresholds were evaluated in ascending order, and the threshold was updated only when Dice increased strictly; ties therefore selected the lower threshold. The resulting weights and τ^* were frozen before inspection of internal held-out or LNDb results and were applied unchanged to the seed-0, seed-1, and seed-2 member sets obtained using the same checkpoint rule.

The primary reported endpoint was Dice for $3 < 6$ mm nodules. Secondary metrics were intersection over union, precision, recall, average surface distance, 95th-percentile Hausdorff distance, and absolute relative volume error [41]. Effect size, 95% confidence

interval, two-sided Monte Carlo randomisation P value, nodule and CT counts, proportion of improved nodules, median paired difference, and seed-specific directions were reported together. A Dice difference of at least 0.01, confidence-interval lower bound above 0, adjusted $P < 0.05$, improvement in more than 55% of nodules, positive median difference, and concordant seed directions were used only as a completeness audit and were not claimed as prospectively registered criteria. No a priori power calculation was performed; all nodules meeting the frozen 3–<6 mm criterion were included.

All method comparisons were paired within the same nodule. For nodule j in CT scan c and training seed s , the paired difference d_{cjs} was first calculated and then averaged across the three seeds within each nodule, $d_{cj} = \frac{1}{3} \sum_s d_{cjs}$. The reported overall effect was the nodule-weighted mean $\Delta = \frac{\sum_c \sum_j d_{cj}}{\sum_c n_c}$. Each CT-cluster bootstrap replicate sampled the same number of CT scans as the original test set with replacement and carried all nodules from each sampled CT; 95% confidence intervals were the 2.5th and 97.5th percentiles of 10,000 replicates [42]. The null hypothesis for the two-sided Monte Carlo CT-cluster sign-flip test was that the joint distribution of paired differences within CT clusters was symmetric about zero. Under this null, all d_{cj} values from a CT scan could be multiplied by the same randomly sampled sign. We used $B = 100,000$ sign flips and calculated the two-sided Monte Carlo randomisation P value as $(b + 1)/(B + 1)$. This test does not require random assignment of models, but its interpretation depends on CT-level sign exchangeability. We therefore also report CT-cluster bootstrap intervals, which do not require the sign-symmetry assumption, and do not describe the randomisation P values as exact. Holm correction was applied across the internal and external primary comparisons [43]. These analyses were performed after locking models, weights, and thresholds on development data but were not formally preregistered. P values for equal-weight, constrained-residual, and oracle analyses were interpreted only within their exploratory analysis families. Inference was conditional on the three trained model sets; neither CT-cluster procedure incorporated the population variance due to retraining or selection of the fixed labelled subset. Seed-specific directions were descriptive stability checks.

2.3 Comparator and secondary analysis design

2.3.1 Model-count-matched homogeneous four-model ensemble controls

To distinguish directional objectives from generic deep-ensemble effects, we evaluated two homogeneous four-model ensembles matched only by model count. First, the four-enhanced-Mean-Teacher ensemble used members with the same backbone, deep supervision, ambiguity auxiliary head, and semi-supervised objective as the single enhanced Mean Teacher. The three outer seed groups were $\{0, 10, 20, 30\}$, $\{1, 11, 21, 31\}$, and $\{2, 12, 22, 32\}$; selection of the labelled CT scans was fixed by label-selection seed 0 for all members. Using only seed-0 development-selection predictions, the same continuous weight optimisation and hard-threshold scan as for the heterogeneous ensemble locked weights of 0.330440, 0.599598, 0.022352, and 0.047610 and a threshold of 0.600289. These settings were applied unchanged to all three outer

seed groups. Second, standard Mean Teacher retained only main-segmentation supervision and exponential-moving-average consistency: $L_{\text{std}} = L_{\text{seg}}(z_s, y) + \lambda_u(t)L_{u,\text{std}}$, where $m(v) = \mathbf{1}[\mu_t(v) \geq 0.90 \text{ or } \mu_t(v) \leq 0.10]$, $w(v) = 2m(v)|\mu_t(v) - 0.5|$, and $L_{u,\text{std}} = \frac{\sum_v w(v)\ell_{\text{BCELogit}}(z_s(v), \mu_t(v))}{\sum_v w(v) + 10^{-6}}$. Standard Mean Teacher used one teacher forward pass ($K = 1$) and did not use deep supervision, an ambiguity branch, Monte Carlo uncertainty control, or the size-aware threshold module. Its three seed groups were likewise $\{0, 10, 20, 30\}$, $\{1, 11, 21, 31\}$, and $\{2, 12, 22, 32\}$. Development seed 0 locked weighted-ensemble weights of 0.1232, 0.4612, 0.0994, and 0.3163 and a threshold of 0.5601; the exploratory equal-weight threshold was 0.596. Both controls used the same 10% labelled subset, strictly unlabelled pool, maximum of 80 epochs, and checkpoint-selection metric. They matched model count only and did not imply equality of architecture, parameter count, training FLOPs, inference FLOPs, or wall-clock time.

2.3.2 Development-locked constrained residual recovery

To test whether complementary experts could recover only adjacent foreground voxels missed by the equal-weight four-standard-Mean-Teacher ensemble, we scanned a low-complexity add-only residual rule in the development selection set. Let P_B be the equal-weight four-standard-Mean-Teacher probability, P_E the locked complementary-expert probability, and $B = \mathbf{1}[P_B \geq 0.596]$ the baseline mask. Candidate additions were restricted to the one-voxel six-connected dilation band surrounding B , excluding B itself. A voxel was added only when $P_E \geq 0.625$, $P_E - P_B \geq 0$, and $P_B \geq 0.40$. The final mask was the union of B and the added region. The rule was selected from 975 candidate combinations on the development selection set and then frozen; no further tuning used internal or LNDb data. Because the rule never removed baseline-mask voxels, its intended effect was to improve recall rather than optimise precision symmetrically.

2.4 Evaluation, statistical analysis, and reporting

2.4.1 Evaluation metrics and statistical analysis

The primary endpoint was Dice similarity coefficient for nodules measuring 3–<6 mm. Secondary endpoints included intersection over union, precision, recall, average surface distance, 95th-percentile Hausdorff distance, and absolute relative volume error, defined according to three-dimensional medical-segmentation metric guidance [41]. All effects were paired within nodules, averaged across the three training seeds within each nodule, and clustered by CT scan. The internal and external primary comparisons of the heterogeneous ensemble with the single enhanced Mean Teacher used 10,000 CT-cluster bootstrap replicates, 100,000 two-sided Monte Carlo CT-cluster sign flips, and Holm correction [42, 43]. The development-locked weighted four-standard-Mean-Teacher ensemble was a secondary model-count-matched control. The development-locked four-enhanced-Mean-Teacher ensemble was treated as an exploratory model-count-matched attribution control. Its internal and external direct comparisons were interpreted descriptively and were not used to support superiority claims; the reported P values were nominal. Equal-weight four-standard-Mean-Teacher, constrained-residual, and oracle analyses were exploratory or descriptive; their P values were nominal or adjusted

within the corresponding exploratory family, did not have the same evidential status as the primary comparisons, and did not support new superiority claims.

2.4.2 Generative artificial intelligence assistance and human verification

OpenAI Codex assisted with language editing and drafting of plotting code. All study design, experimental results, statistical interpretations, references, and final text were verified item by item by the named authors, who accept responsibility for the work. Generative artificial intelligence was not listed as an author, did not generate experimental data, and did not alter the original medical images.

3 Results

3.1 Development audit and method selection

3.1.1 The initial framework did not resolve bidirectional errors in small nodules

The size-aware threshold and dual-uncertainty framework improved local pseudo-label recall but did not improve final Dice for small nodules. Relative to the fixed-threshold enhanced Mean Teacher, its mean development-selection Dice decreased by 0.013. Audit of paired nodules showed that lowering the positive threshold recovered some low-probability foreground voxels but simultaneously expanded false-positive boundary regions. Uniform uncertainty gating could attenuate both types of voxels but could not correct omission and expansion in opposite directions. This negative result motivated the transition from a single thresholding mechanism to complementary recall- and precision-oriented objectives.

3.1.2 Development shifted from a single mechanism to complementary model solutions

On the development selection set, E1 and E2 each improved Dice by 0.00292 relative to the enhanced Mean Teacher, whereas E3 changed Dice by -0.00055 . A volume-oriented candidate improved Dice by 0.00384 but reduced recall and was therefore not retained. For training seed 0, the locked four-model heterogeneous ensemble improved Dice by 0.01160; across the three evaluated seeds, the mean development-selection gain was 0.00888. Adding a fifth volume-oriented model did not improve the fitted development objective (Dice, 0.817146 versus 0.817179 for the four-model ensemble). The four-member configuration was therefore frozen before confirmatory testing (Fig. 2).

3.2 Primary internal and external evaluation

3.2.1 Internal held-out testing

In the internal CT-level held-out test set, the mean Dice of the single enhanced Mean Teacher was 0.76428 and that of the locked heterogeneous ensemble was 0.77915. The paired improvement was 0.01488 (CT-cluster 95% CI, 0.00665–0.02347; two-sided

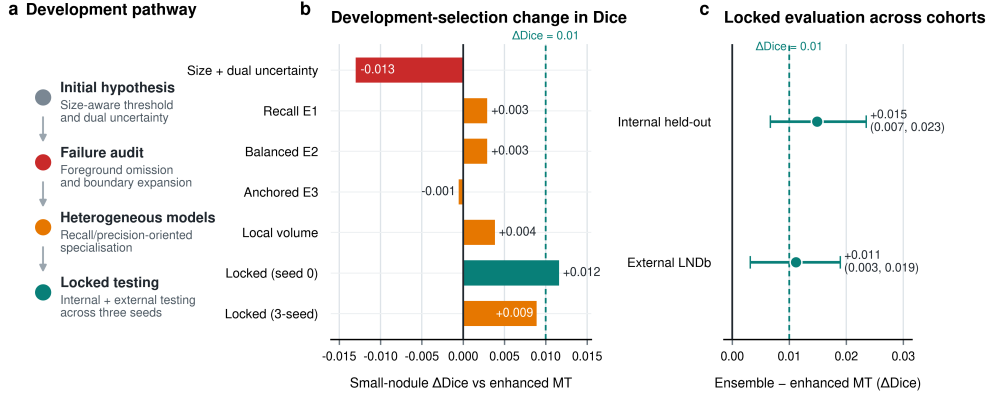


Fig. 2 Development audit and locked ensemble evaluation. The original size-aware and dual-uncertainty mechanism improved local pseudo-label recall but did not improve final small-nodule Dice. Paired error auditing identified coexisting foreground omission and boundary expansion, motivating independently trained recall- and precision-oriented models. Probability weights and the binarisation threshold were fitted only in the development selection set and then frozen for internal and external testing. MT, Mean Teacher.

Table 2 Primary small-nodule Dice results

Dataset	Small nodules/ CT scans	Enhanced MT Dice	Locked ensemble Dice	Δ Dice (CT-cluster 95% CI)
Internal CT-level held-out test	62/51	0.76428	0.77915	0.01488 (0.00665–0.02347)
External LNDb	436/176	0.52586	0.53705	0.01118 (0.00314–0.01892)

Monte Carlo randomisation $P = 0.00055$; Holm-adjusted $P = 0.00110$). Dice improved in 77.4% of nodules, and the median paired difference was 0.00995. Seed-specific effects were 0.01424, 0.01944, and 0.01095 for seeds 0, 1, and 2, respectively; all were positive (Table 2 and Fig. 3).

3.2.2 External testing without LNDb adaptation

In the external LNDb test set, the mean Dice of the single enhanced Mean Teacher was 0.52586 and that of the locked heterogeneous ensemble was 0.53705. The paired improvement was 0.01118 (CT-cluster 95% CI, 0.00314–0.01892; two-sided Monte Carlo randomisation $P = 0.00673$; Holm-adjusted $P = 0.00673$). Dice improved in 71.3% of nodules, and the median paired difference was 0.01108. Seed-specific effects were 0.01647, 0.01094, and 0.00615; all were positive despite differences in magnitude. No LNDb image, label, threshold, or post-processing decision was used during development (Table 2 and Fig. 3).

Values are means across the three evaluated training seeds. Confidence intervals were obtained using 10,000 CT-cluster bootstrap replicates. Adjusted P values used Holm correction across the internal and external primary comparisons. CI, confidence interval; MT, Mean Teacher.

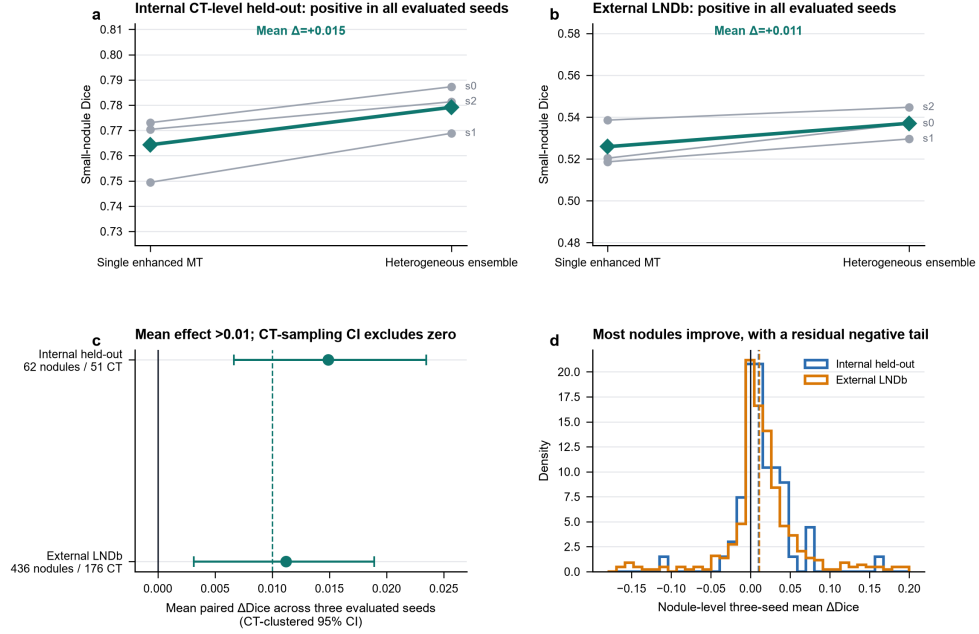


Fig. 3 Internal and external gains over a single enhanced Mean Teacher. (a, b) Mean Dice for the single enhanced Mean Teacher and the development-locked heterogeneous ensemble. (c, d) CT-cluster bootstrap distributions of the paired Dice effect. Error bars show CT-cluster 95% confidence intervals. Effects were averaged across the three evaluated training seeds within each nodule before CT-cluster inference. LNDb was tested without fine-tuning, recalibration, or threshold reselection. MT, Mean Teacher.

3.3 Attribution and exploratory analyses

3.3.1 Model-count-matched attribution analysis

The homogeneous ensemble of four enhanced Mean Teachers improved internal Dice from 0.76428 to 0.78309 but reduced external Dice from 0.52586 to 0.50748, with negative external effects in all three outer seed groups. Compared directly with this exploratory control, the heterogeneous ensemble had a mean internal difference of -0.00394 (95% CI, -0.01577 to 0.00468 ; nominal $P = 0.54035$) and an external difference of 0.02956 (95% CI, 0.01939 – 0.04015 ; nominal $P = 0.00001$). These opposing results did not establish a stable cross-cohort advantage attributable to heterogeneous training objectives and were not used to support superiority claims.

The development-locked weighted ensemble of four standard Mean Teachers improved internal Dice by 0.00813 relative to a single standard Mean Teacher but changed external Dice by -0.00534 ; the external direction reversed in outer seed group 2. Relative to this weighted control, the heterogeneous ensemble differed by -0.00525 internally and 0.00671 externally; the external interval crossed zero (95% CI, -0.00207 to 0.01558 ; $P = 0.13684$). In the exploratory equal-weight analysis, four standard Mean Teachers improved Dice by 0.01091 internally and 0.01489 externally relative to the

Table 3 Model-count-matched attribution analysis

Method	Internal Dice	External LNDb Dice	Interpretation relative to single model
Single enhanced MT	0.76428	0.52586	Primary single-model comparator
Four enhanced MTs (locked weighted)	0.78309	0.50748	Internal gain; external decrease
Heterogeneous model ensemble	0.77915	0.53705	Primary method; gain over single enhanced MT in both cohorts
Four standard MTs (locked weighted)	0.78441	0.53034	Secondary model-count-matched control
Four standard MTs (equal weight)	0.78719	0.55057	Exploratory; highest external Dice

corresponding single standard model. Its external Dice was 0.55057, exceeding the heterogeneous ensemble by 0.01352 (heterogeneous minus equal-weight standard ensemble: 95% CI, -0.02122 to -0.00604 ; nominal $P = 0.00048$). Thus, the primary gain over a single enhanced Mean Teacher was supported, but a general heterogeneous-objective advantage over homogeneous deep ensembles was not (Table 3; Figs. 4 and 5).

Values are means across the three evaluated outer seed groups. The equal-weight four-standard-Mean-Teacher analysis was exploratory. These controls were matched by model count, not by architecture, training computation, or parameter count. MT, Mean Teacher.

3.3.2 Constrained residual recovery and oracle analysis

The development-locked add-only residual rule produced a mean Dice difference of 0.00183 relative to the equal-weight four-standard-Mean-Teacher ensemble in the internal test set (95% CI, -0.00076 to 0.00458 ; $P = 0.18742$; Holm-adjusted $P = 0.56225$). Dice improved in 54.8% of nodules, and all three seed-specific effects were positive. In LNDb, the mean difference was 0.00210 (95% CI, -0.00037 to 0.00465 ; $P = 0.10608$; Holm-adjusted $P = 0.31824$); 34.4% of nodules improved, the median difference was -0.00152 , and all three seed-specific mean effects were positive. Because both confidence intervals crossed zero and the nodule-level distributions were inconsistent, the rule did not establish a reliable improvement.

Relative to the single enhanced Mean Teacher, the heterogeneous ensemble shifted the operating balance towards recall. Internally, recall increased by 0.01546 (95% CI, 0.01191 – 0.01941 ; Holm-adjusted $P = 0.00007$), whereas precision decreased by 0.01185 (95% CI, -0.01536 to -0.00864). Externally, recall increased by 0.02377 (95% CI, 0.02067 – 0.02705), whereas precision decreased by 0.01545 (95% CI, -0.01735 to -0.01359); average surface distance and absolute relative volume error also worsened. These results indicate that the Dice gain was obtained through a controlled shift in the recall–precision balance rather than uniform improvement in all metrics.

A nodule-level oracle that selected the best member for each nodule provided an upper-bound gain of 0.01447 on the seed-0 development selection set. Across the three evaluated seeds, the oracle gain ranged from 0.00547 to 0.02109 internally and

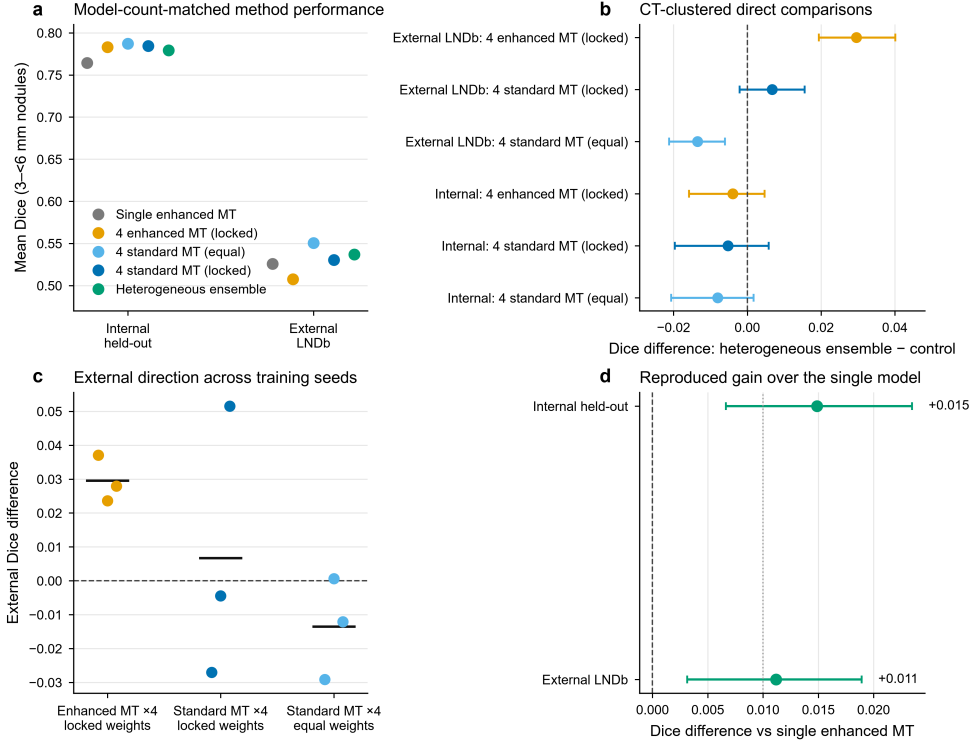


Fig. 4 Model-count-matched attribution analysis for small-nodule segmentation. (a) Internal and external Dice for the single enhanced Mean Teacher, development-locked four-enhanced-Mean-Teacher ensemble, development-locked weighted four-standard-Mean-Teacher ensemble, exploratory equal-weight four-standard-Mean-Teacher ensemble, and heterogeneous ensemble. (b) CT-clustered direct comparisons of the heterogeneous ensemble with each four-model control. (c) Seed-specific external directions for the same direct comparisons. (d) Primary heterogeneous-ensemble gains over the single enhanced Mean Teacher. The four-enhanced-Mean-Teacher comparisons were exploratory and their P values were nominal. MT, Mean Teacher.

from 0.02068 to 0.03942 externally. This remaining gap suggests exploitable nodule-level complementarity, but it does not constitute a deployable routing result because the oracle used the ground-truth mask for selection (Fig. 6). Extended seed-specific and secondary-metric results are provided in Additional file 1, with machine-readable supporting data in Additional file 2.

3.3.3 Qualitative error patterns

Qualitative LNDb examples illustrated the attribution comparison with the strong exploratory equal-weight four-standard-Mean-Teacher ensemble. In some nodules, the heterogeneous ensemble recovered peripheral foreground missed by the homogeneous ensemble and improved overlap. In others, their predictions were similar. Failure cases showed that the heterogeneous ensemble could extend into adjacent parenchyma or vascular structures, reducing precision despite recovering additional foreground (Fig. 7).

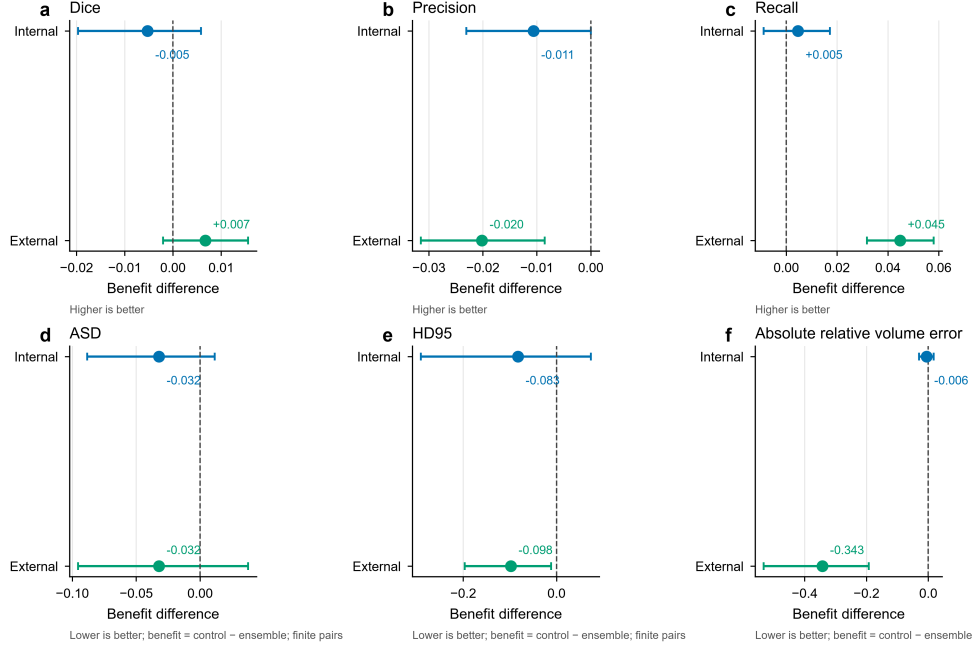


Fig. 5 Metric trade-offs in the model-count-matched attribution analysis. The panels show direct paired comparisons of the locked heterogeneous ensemble with the development-locked weighted four-standard-Mean-Teacher ensemble. For Dice, precision, and recall, benefit was defined as heterogeneous ensemble minus control. For ASD, HD95, and absolute relative volume error, benefit was defined as control minus heterogeneous ensemble because lower values are preferable. Positive values therefore favour the heterogeneous ensemble in every panel. Error bars show CT-cluster 95% confidence intervals. ASD, average surface distance; HD95, 95th-percentile Hausdorff distance; MT, Mean Teacher.

3.4 Contextual comparison with previous studies

Published pulmonary-nodule segmentation studies differ substantially in candidate definition, nodule-size range, label budget, train-test separation, and external-testing protocol [19, 24, 25, 39, 44, 45]. We therefore did not rank methods using cross-paper Dice values. The appropriate comparison for the present study was paired testing within the same nodules and frozen protocol. Under that protocol, the heterogeneous ensemble exceeded the single enhanced Mean Teacher by more than 0.01 Dice in both the internal and external cohorts, but its advantage over model-count-matched homogeneous ensembles was cohort dependent.

4 Discussion

This study evaluated whether a development-locked heterogeneous model ensemble could improve semi-supervised segmentation of candidate-centred 3–<6 mm pulmonary nodules relative to a single enhanced Mean Teacher. The primary finding was a small but consistent Dice improvement in both the internal CT-level held-out test set and external LNDb without adaptation. The mean gains were 0.01488 and 0.01118,

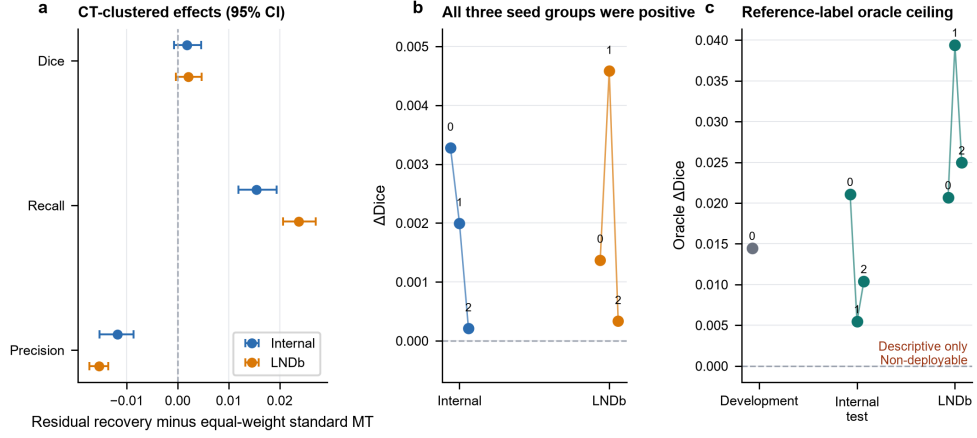


Fig. 6 Constrained residual effects and nodule-level oracle upper bound. (a) CT-clustered Dice, recall, and precision effects for constrained residual recovery relative to the equal-weight four-standard-Mean-Teacher ensemble. (b) Seed-specific Dice effects of the same residual rule. (c) Performance of a ground-truth-informed nodule-level oracle, shown only as a descriptive upper bound. The oracle is not deployable and was not used for the primary analysis. Error bars show CT-cluster 95% confidence intervals. MT, Mean Teacher.

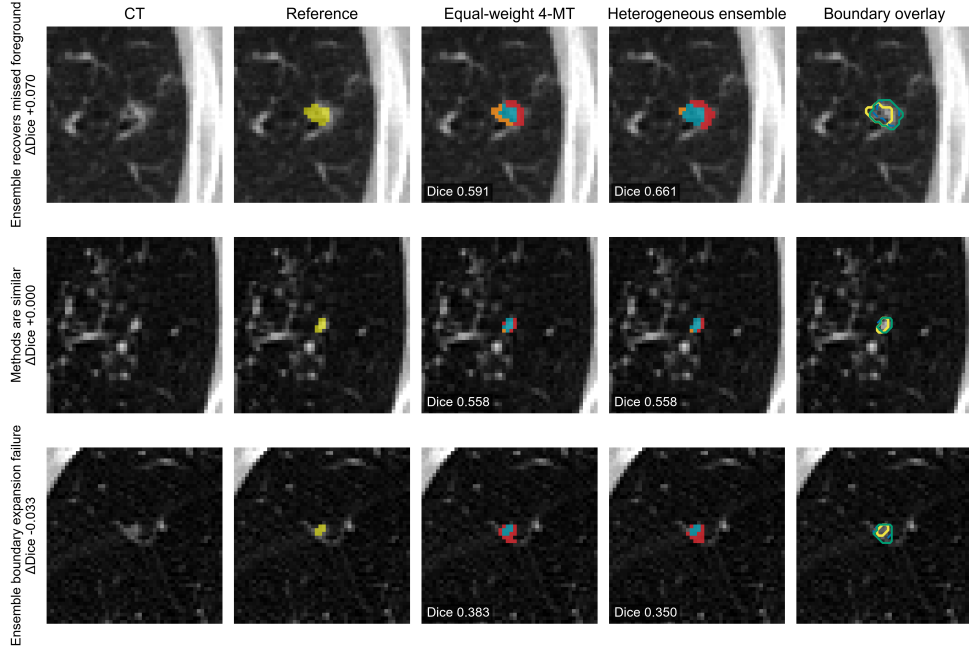


Fig. 7 Qualitative segmentation and error overlays in external LNDb. Representative nodules show improved peripheral foreground recovery, similar predictions, and boundary over-expansion. Columns display the CT patch, reference mask, equal-weight four-standard-Mean-Teacher prediction, heterogeneous-ensemble prediction, and boundary overlay. Examples were selected to illustrate observed error modes rather than estimate their prevalence. MT, Mean Teacher.

respectively; both CT-cluster confidence intervals excluded zero, adjusted P values were below 0.05, most nodules improved, and all three evaluated training seeds had positive effects. These results support the bounded claim that the locked multi-model system improved on its single enhanced-Mean-Teacher baseline under the present protocol.

The development history clarifies why the final method differs from the original hypothesis. A size-aware positive threshold directly addressed the low foreground fraction of small nodules, while dual-uncertainty control separated predictive instability from an ambiguity proxy. However, improving pseudo-label recall alone did not improve final Dice. Small nodules combined at least two opposing errors: foreground omission and boundary expansion. Lowering the threshold could correct the former while aggravating the latter, and uniform confidence gating attenuated both. The subsequent expert models preserved solutions with different error preferences instead of forcing a single voxel-selection rule to resolve both simultaneously.

The secondary metrics nevertheless define an important boundary of the benefit. The ensemble consistently shifted the operating point towards recall, with corresponding reductions in precision and worsening of some surface and volume metrics externally. The Dice gain should therefore not be interpreted as uniform contour improvement. For applications in which missing a small peripheral component is costlier than limited over-segmentation, this operating shift may be useful. For precise volumetry or growth assessment, however, the additional boundary and volume errors may be clinically undesirable and require explicit calibration.

The model-count-matched analyses also constrain causal interpretation. The heterogeneous ensemble did not show a stable advantage over four homogeneous enhanced Mean Teachers across cohorts, and the exploratory equal-weight four-standard-Mean-Teacher ensemble achieved the highest external Dice. Accordingly, the present experiments cannot separate gains due to directional expert objectives from generic variance reduction and probability averaging in a deep ensemble [33]. The four-member heterogeneous system remains the development-locked primary method because its weights and threshold were locked using development data and it improved on the single enhanced Mean Teacher in both test cohorts; nevertheless, heterogeneity itself should be viewed as a plausible design rationale rather than an independently established source of benefit.

External testing is a strength because no LNDb data were used for fine-tuning, recalibration, threshold selection, or post-processing. The concordant direction across the internal and external cohorts reduces the likelihood that the primary result reflects only one held-out split. CT-cluster inference also avoided treating multiple nodules from the same scan as statistically independent. However, the available public identifiers did not permit verification of whether any person contributed more than one CT scan. The highest auditable clustering level was therefore the CT scan rather than the patient.

Several limitations remain. First, this was candidate-centred segmentation and did not evaluate whole-volume detection. Second, only 35 of 350 development-training CT scans contributed voxel-level supervision, and the fixed labelled subset was not repeated; the statistical intervals were conditional on that selection and on three trained model sets. Third, internal and external reference masks were constructed differently, and LNDb was not an independent prospective clinical cohort. Fourth, the

four-model system increases training, storage, and inference requirements relative to a single model. Fifth, model-count matching did not equalise architecture, parameter count, training FLOPs, inference FLOPs, or wall-clock time. Finally, ensemble fitting used seed-0 development predictions, and the equal-weight standard-Mean-Teacher analysis was exploratory. Future work should prospectively lock the heterogeneous and strongest homogeneous ensembles, repeat label-subset selection, evaluate additional independent cohorts, test calibration for clinically relevant precision–recall operating points, and develop routing or distillation methods that preserve complementarity with lower inference cost.

5 Conclusions

Under the present candidate-centred, 10%-label protocol, a development-locked heterogeneous model ensemble improved Dice for 3–<6 mm pulmonary nodules by more than 0.01 relative to a single enhanced Mean Teacher in both an internal CT-level held-out test set and external LNDb without adaptation. The effect direction was concordant across the three evaluated training seeds and was accompanied by a shift towards higher recall and lower precision. These findings support a modest, externally observed ensemble benefit, but they do not establish an independent advantage of heterogeneous training objectives beyond general deep-ensemble effects. Prospective comparison with a locked strong homogeneous ensemble and testing in new independent cohorts are needed.

List of abbreviations

ASD: average surface distance; CI: confidence interval; CT: computed tomography; FP: false positive; FN: false negative; HD 95: 95th-percentile Hausdorff distance; HU: Hounsfield unit; LNDb: Lung Nodule Database; MT: Mean Teacher; ReLU: rectified linear unit; SHA-256: Secure Hash Algorithm 256-bit; TP: true positive.

Declarations

Ethics approval and consent to participate

Ethical approval and informed consent were not required for this secondary analysis of fully de-identified, publicly available datasets. The original data were collected under the ethics and governance procedures of their respective repositories.

Consent for publication

Not applicable. No identifiable individual data are presented.

Availability of data and materials

The LUNA16/LIDC-IDRI and LNDb datasets used in this study are available from their respective public repositories, subject to the applicable terms of use [26, 34–36]. Redistributable derived materials, including fixed split files, candidate-centre

manifests, locked probability-combination weights and thresholds, and analysis tables, are publicly available in the archived reproducibility release on Zenodo (version 0.1.0; <https://doi.org/10.5281/zenodo.22152092>), with the corresponding development repository hosted on GitHub (<https://github.com/wodiui/reproduce->). Raw CT images and third-party annotations are not redistributed; instructions for accessing the source datasets and reconstructing the analysis inputs are provided in the repository documentation.

Competing interests

The authors declare that they have no competing interests.

Funding

Supported by "the Fundamental Research Funds for the Central Universities" (PT2026-18).

Acknowledgements

Not applicable.

Authors' contributions

HL conceived and designed the study, developed the methodology and software, conducted the experiments, analysed the data, prepared the figures, and drafted the manuscript. ZHZ contributed to data curation, experimental validation, statistical analysis, and critical revision of the manuscript. XXW supervised the study, contributed to the interpretation of the results, and critically revised the manuscript. YC contributed to supervision, methodological review, and critical revision of the manuscript. FL contributed to project administration, funding acquisition, supervision, and critical revision of the manuscript. All authors read and approved the final manuscript. All authors read and approved the final manuscript.

References

- [1] de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *N Engl J Med*. 2020;382(6):503-513. <https://doi.org/10.1056/NEJMoa1911793>.
- [2] MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology*. 2017;284(1):228-243. <https://doi.org/10.1148/radiol.2017161659>.
- [3] Wormanns D, Kohl G, Klotz E, Marheine A, Beyer F, Heindel W, et al. Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo

- p reproducibility.
- Eur Radiol.*
- 2004;14(1):86-92.
- <https://doi.org/10.1007/s00330-003-2132-0>
- .
- [4] de Hoop B, Gietema H, van Ginneken B, Zanen P, Groenewegen G, Prokop M. A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: What is the minimum increase in size to detect growth in repeated CT examinations. *Eur Radiol.* 2009;19(4):800-808. <https://doi.org/10.1007/s00330-008-1229-x>.
 - [5] Zhi L, Jiang W, Zhang S, Zhou T. Deep neural network pulmonary nodule segmentation methods for CT images: Literature review and experimental comparisons. *Comput Biol Med.* 2023;164:107321. <https://doi.org/10.1016/j.compbiomed.2023.107321>.
 - [6] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science.* 2015:234-241. https://doi.org/10.1007/978-3-319-24574-4_28.
 - [7] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *Lecture Notes in Computer Science.* 2016:424-432. https://doi.org/10.1007/978-3-319-46723-8_49.
 - [8] Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV). 2016:565-571. <https://doi.org/10.1109/3DV.2016.79>.
 - [9] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognit (CVPR). 2016:770-778. <https://doi.org/10.1109/CVPR.2016.90>.
 - [10] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2021;18(2):203-211. <https://doi.org/10.1038/s41592-020-01008-z>.
 - [11] Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Med Image Anal.* 2017;40:172-183. <https://doi.org/10.1016/j.media.2017.06.014>.
 - [12] Jiao R, Zhang Y, Ding L, Xue B, Zhang J, Cai R, et al. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Comput Biol Med.* 2024;169:107840. <https://doi.org/10.1016/j.compbiomed.2023.107840>.
 - [13] Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems.* 2017;30:1195-1204.

- [14] Yu L, Wang S, Li X, Fu CW, Heng PA. Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation. *Lecture Notes in Computer Science*. 2019;605-613. https://doi.org/10.1007/978-3-030-32245-8_67.
- [15] Luo X, Chen J, Song T, Wang G. Semi-supervised Medical Image Segmentation through Dual-task Consistency. *Proc AAAI Conf Artif Intell*. 2021;35(10):8801-8809. <https://doi.org/10.1609/aaai.v35i10.17066>.
- [16] Li S, Zhang C, He X. Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images. *Lecture Notes in Computer Science*. 2020;552-561. https://doi.org/10.1007/978-3-030-59710-8_54.
- [17] Xia Y, Yang D, Yu Z, Liu F, Cai J, Yu L, et al. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Med Image Anal*. 2020;65:101766. <https://doi.org/10.1016/j.media.2020.101766>.
- [18] Luo X, Wang G, Liao W, Chen J, Song T, Chen Y, et al. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med Image Anal*. 2022;80:102517. <https://doi.org/10.1016/j.media.2022.102517>.
- [19] Wu Y, Ge Z, Zhang D, Xu M, Zhang L, Xia Y, et al. Mutual consistency learning for semi-supervised medical image segmentation. *Med Image Anal*. 2022;81:102530. <https://doi.org/10.1016/j.media.2022.102530>.
- [20] Wang K, Zhan B, Zu C, Wu X, Zhou J, Zhou L, et al. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Med Image Anal*. 2022;79:102447. <https://doi.org/10.1016/j.media.2022.102447>.
- [21] You C, Zhou Y, Zhao R, Staib L, Duncan JS. SimCVD: Simple Contrastive Voxel-Wise Representation Distillation for Semi-Supervised Medical Image Segmentation. *IEEE Trans Med Imaging*. 2022;41(9):2228-2237. <https://doi.org/10.1109/TMI.2022.3161829>.
- [22] Xu Z, Wang Y, Lu D, Luo X, Yan J, Zheng Y, et al. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Med Image Anal*. 2023;88:102880. <https://doi.org/10.1016/j.media.2023.102880>.
- [23] Chen X, Yuan Y, Zeng G, Wang J. Semi-supervised semantic segmentation with cross pseudo supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit*. 2021:2613-2622.
- [24] Qiao P, Li H, Song G, Han H, Gao Z, Tian Y, et al. Semi-Supervised CT Lesion Segmentation Using Uncertainty-Based Data Pairing and SwapMix. *IEEE Trans Med Imaging*. 2023;42(5):1546-1562. <https://doi.org/10.1109/TMI.2022.3232572>.

- [25] Liu W, Zhang L, Li X, Liu H, Feng M, Li Y, et al. A semisupervised knowledge distillation model for lung nodule segmentation. *Sci Rep.* 2025;15:10562. <https://doi.org/10.1038/s41598-025-94132-9>.
- [26] McNitt-Gray MF, Armato SG, Meyer CR, Reeves AP, McLennan G, Pais RC, et al. The Lung Image Database Consortium (LIDC) Data Collection Process for Nodule Detection and Annotation. *Acad Radiol.* 2007;14(12):1464-1474. <https://doi.org/10.1016/j.acra.2007.07.021>.
- [27] Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Med Phys.* 2011;38(2):915-931. <https://doi.org/10.1118/1.3528204>.
- [28] Armato SG, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR, et al. The Lung Image Database Consortium (LIDC): An Evaluation of Radiologist Variability in the Identification of Lung Nodules on CT Scans. *Acad Radiol.* 2007;14(11):1409-1421. <https://doi.org/10.1016/j.acra.2007.07.008>.
- [29] Lin H, Huang C, Wang W, Luo J, Yang X, Liu Y. Measuring Interobserver Disagreement in Rating Diagnostic Characteristics of Pulmonary Nodule Using the Lung Imaging Database Consortium and Image Database Resource Initiative. *Acad Radiol.* 2017;24(4):401-410. <https://doi.org/10.1016/j.acra.2016.11.022>.
- [30] Kohl SAA, Romera-Paredes B, Meyer C, De Fauw J, Ledsam JR, Maier-Hein KH, et al. A probabilistic U-Net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems.* 2018;31:6965-6975.
- [31] Ji W, Yu S, Wu J, Ma K, Bian C, Bi Q, et al. Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognit (CVPR). 2021:12336-12346. <https://doi.org/10.1109/CVPR46437.2021.01216>.
- [32] Zhang L, Tanno R, Xu M, Huang Y, Bronik K, Jin C, et al. Learning from multiple annotators for medical image segmentation. *Pattern Recognit.* 2023;138:109400. <https://doi.org/10.1016/j.patcog.2023.109400>.
- [33] Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems.* 2017;30:6402-6413.
- [34] Setio AAA, Traverso A, de Bel T, Berens MSN, van den Bogaard C, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med Image Anal.* 2017;42:1-13. <https://doi.org/10.1016/j.media.2017.06.015>.

- [35] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. <https://doi.org/10.1007/s10278-013-9622-7>.
- [36] Pedrosa J, Aresta G, Ferreira C, Atwal G, Phoulady HA, Chen X, et al. LNDb challenge on automatic lung cancer patient management. *Med Image Anal*. 2021;70:102027. <https://doi.org/10.1016/j.media.2021.102027>.
- [37] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*. 2016;48:1050-1059.
- [38] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*. 2017;30:5574-5584.
- [39] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf Fusion*. 2021;76:243-297. <https://doi.org/10.1016/j.inffus.2021.05.008>.
- [40] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Wang Q, Shi Y, Suk H-I, Suzuki K, editors. *Machine Learning in Medical Imaging*. Cham: Springer; 2017. p. 379-387. https://doi.org/10.1007/978-3-319-67389-9_44.
- [41] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15(1):29. <https://doi.org/10.1186/s12880-015-0068-x>.
- [42] Efron B. Bootstrap Methods: Another Look at the Jackknife. *Ann Stat*. 1979;7(1):1-26. <https://doi.org/10.1214/aos/1176344552>.
- [43] Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.
- [44] Pham TQ, Le TT, Lai KH, Ngo DV, Pham TD, Hua Q, et al. CAAF-ResUNet: Adaptive Attention Fusion with Boundary-Aware Loss for Lung Nodule Segmentation. *Medicina (Kaunas)*. 2025;61(7):1126. <https://doi.org/10.3390/medicina61071126>.
- [45] Zhang K, Lan X, Wang Y, Wang L, Liu Y, Guo F. CA-3DTransUNet with dynamic cross-scale fusion for pulmonary nodule segmentation. *Sci Rep*. 2026;16:16033. <https://doi.org/10.1038/s41598-026-47436-3>.