Information Sciences and Technological Innovations

www.isti.reapress.com

Inf. Sci. Technol. Innov. Vol. 2, No. 3 (2025) 241–252.

Paper Type: Original Article

Uncertainty-Guided Reliability Enhancement of Residual U-Net CT Segmentation in Medical Cancer Imaging

Amirhossein Nafei*

Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei 106, Taiwan; Amir.nafei@e.gzhu.edu.cn.

Citation:

Received: 17 January 2025 *Revised: 05 April 2025* *Accepted: 22 June 2025*	Nafei, A. (2025). Uncertainty-guided reliability enhancement of residual U-Net CT segmentation in medical cancer imaging. *Information sciences and technological innovations, 3*(2), 241-252.

Abstract

Reliable Computed Tomography (CT) segmentation is a critical requirement for quantitative imaging and computer-aided clinical workflows. Although Residual U-Net (Res-U-Net) architectures achieve strong overlap performance on curated datasets, threshold-based binarization of probability maps often produces scattered false positives, particularly in low-contrast regions and near complex anatomical boundaries. This study analyses the role of predictive uncertainty in improving the structural reliability of CT segmentation outputs. Monte-Carlo dropout is employed at inference time to estimate pixel-wise predictive variance, which is combined with mean probability and component size information within a connected-component framework. A component-level scoring rule is evaluated to suppress unstable, low-confidence regions while preserving coherent anatomical structures. Quantitative experiments demonstrate that uncertainty-aware filtering substantially reduces region-level false positives per scan and improves boundary stability, while maintaining competitive Dice and Intersection over Union (IoU) scores. An ablation study further shows that uncertainty penalization is the primary driver of false-positive reduction, and that combining uncertainty with mild size regularization yields the most balanced performance. The results support the use of uncertainty-guided refinement as a practical reliability layer for Res-U-Net–based CT segmentation systems.

Keywords: Computed tomography, Lung segmentation, Residual U-Net, Uncertainty estimation, Monte-Carlo dropout, False positive removal, Connected components, Medical image segmentation.

Corresponding Author: Amir.nafei@e.gzhu.edu.cn

https://doi.org/10.48314/isti.vi.47

1 | Introduction

Computed Tomography (CT) [1] plays a central role in modern clinical workflows because it supports rapid, high-resolution assessment of internal anatomy and disease patterns across a broad range of conditions. In thoracic imaging in particular, CT is routinely used for screening, diagnosis, treatment planning, and longitudinal monitoring, and large-scale evidence has established the clinical impact of low-dose CT screening on lung-cancer mortality [2]. Many downstream clinical and computational tasks depend on reliable delineation of anatomical regions (for example, the lung fields) and, in some scenarios, pathological findings (for example, consolidations, nodules, or other lesion-like patterns). For this reason, CT segmentation has become a fundamental building block in computer-aided diagnosis and quantitative radiology, where segmentation masks are used to compute volumetric biomarkers, constrain regions of interest for detection systems, and standardize measurements across time and across institutions [3]. Deep learning–based encoder–decoder models have been widely studied for CT segmentation because they can learn hierarchical image representations and produce dense pixel-level probability maps [3]. U-Net variants in particular have become a strong baseline due to their use of skip connections, which combine high-level semantic information from deep layers with spatial details preserved in shallow layers [4]. Residual formulations strengthen this design by using residual blocks to improve gradient flow and stabilize training, especially when networks become deeper or when datasets contain heterogeneous acquisition settings [5]. In addition, widely used segmentation frameworks such as nnU-Net have reinforced the view that carefully designed (and often self-configuring) U-Net pipelines can yield strong performance across many biomedical segmentation tasks and datasets [6]. Extensions such as 3D U-Net and recurrent Res-U-Net variants further illustrate the broader trend of adapting U-Net-style encoder–decoder models to volumetric contexts and more expressive feature refinement [7], [8]. In many reported benchmarks, these architectures achieve high overlap scores on curated test splits, indicating that the primary anatomical structures can be segmented with strong average accuracy [3], [6].

Despite this progress, a consistent practical limitation remains visible when segmentation systems are evaluated under realistic conditions: segmentation probability maps can contain scattered false positives that survive thresholding and appear as isolated blobs or small leakage regions. These errors are often concentrated near challenging anatomical boundaries, in low-contrast regions, or in areas affected by partial-volume artifacts and reconstruction variability [1]. They can also become more prominent when the test distribution differs from the training distribution, such as when scans are collected using different scanners, reconstruction kernels, slice thicknesses, or patient populations—conditions frequently encountered in real-world deployments and screening programs [2], [9]. In operational environments, these false positives are not a minor cosmetic issue. Even small spurious components can distort volume estimates, contaminate quantitative imaging features, and create false alarms in downstream pipelines that assume the segmentation mask is clean and anatomically consistent [3]. From a clinical perspective, such errors reduce trust and increase the need for manual correction, which directly weakens the utility of automated segmentation [9]. This problem is closely connected to how segmentation predictions are typically converted into final masks. In most pipelines, the network outputs a sigmoid probability map, and a threshold is applied to obtain a binary segmentation. This decision rule is attractive because it is simple, fast, and easy to implement, and its effectiveness is closely linked to overlap-based evaluation criteria such as Dice and Intersection over Union

(IoU) [10], [11]. However, thresholding assumes that the probability map is sufficiently calibrated and that errors are primarily low-confidence pixels that can be removed by selecting an appropriate cutoff. In practice, CT images frequently include ambiguous pixels where the model's evidence is weak; these pixels may still exceed the threshold due to local texture similarity, noise patterns, or domain shift effects, creating small connected components that do not correspond to true anatomy [1], [3]. Traditional post-processing approaches (for example, morphological filtering) are often used to mitigate such artifacts, yet these operations can be blunt and may remove true small structures when the same parameters are applied across heterogeneous scans [12].

A reliability-oriented perspective suggests that segmentation should be assessed not only by overlap metrics (such as Dice or IoU) [10], [11], but also by boundary stability and the structure of errors. Boundary-sensitive measures are commonly included for this purpose; for example, Hausdorff-type distances (including robust variants such as HD95) are often used to quantify the worst-case boundary discrepancy while limiting sensitivity to extreme outliers [13], [14]. In this context, predictive uncertainty becomes a meaningful diagnostic signal. When a segmentation model is uncertain about a region, the predicted probabilities tend to fluctuate under stochastic inference settings, reflecting unstable internal representations. Inference-time uncertainty estimation methods, including Monte-Carlo dropout, provide a practical way to quantify this instability without requiring changes to the training objective or data labeling process [15]. Subsequent analyses in computer vision and medical imaging have further discussed how uncertainty captures different error modes and can support risk-aware decision-making [16], [17]. Related lines of work have also examined uncertainty from ensembles and probabilistic segmentation formulations, reinforcing the broader idea that multiple plausible predictions can occur in ambiguous regions [18], [19]. From this viewpoint, false positive removal can be treated as a reliability-aware refinement step that leverages uncertainty information to improve the final mask. Many clinical segmentation tasks contain structural priors that are naturally compatible with this approach. For example, true anatomical regions tend to form large coherent components, while false positives frequently appear as small, fragmented blobs; these patterns are well aligned with the classic use of connected components and morphology in segmentation pipelines [12]. Similarly, true structures usually exhibit relatively stable predictions across stochastic inference passes, whereas spurious regions often show higher variance—an effect consistent with the theoretical interpretation of Monte-Carlo dropout as approximate Bayesian inference [15] and with uncertainty discussions in vision settings [16]. A connected-component framework that combines mean probability, uncertainty statistics, and mild size regularization therefore offers an interpretable mechanism to suppress unreliable components while preserving the main segmentation region, complementing the strong representational capacity of U-Net-style encoders and decoders [4], [6].

The motivation of this study is driven by a practical need: improving the structural reliability of CT segmentation outputs in the presence of ambiguous evidence and domain variation, rather than focusing on architectural novelty. The workflow examined aligns with widely adopted training–testing pipelines in which a Res-U-Net-style backbone is trained using ground truth masks and then used to generate probability maps for unseen CT sequences [4–6]. The architecture context shown in *Fig. 1* (Image credited to Khanna, Anita, et al. [20]) reflects this common workflow: training yields model weights, testing yields a probability map, and a refinement stage supports generation of a final segmented sequence. In this paper, that refinement stage is analyzed through uncertainty estimation and component-level filtering, with an emphasis on transparency, interpretability, and reproducibility, consistent with the broader literature on uncertainty estimation and probabilistic segmentation [15–19]. The objective of this study is to systematically examine how predictive uncertainty can support false positive removal in Res-U-Net-based CT segmentation. This objective includes: 1) describing an inference-time uncertainty computation mechanism suitable for segmentation probability maps [15–17], 2) defining component-level reliability statistics that summarize mean confidence and uncertainty within each predicted region, and 3) evaluating how uncertainty-guided suppression affects both standard overlap metrics and reliability-sensitive measures such as region-level false positive counts and

boundary distances [10, 11], [13], [14]. The analysis is framed to be compatible with slice-wise inference over CT sequences, which remains common in practice due to computational constraints, while also being extensible to 3D processing in volumetric settings [7].

The contributions of this work can be summarized in three points. First, the study presents a reliability-oriented analysis of Res-U-Net-based CT segmentation, emphasizing the gap between high overlap scores and the presence of scattered false positives that persist after thresholding [4], [3], [6]. Second, the study evaluates an uncertainty-aware component scoring framework that uses Monte-Carlo dropout variance as a quantitative indicator of prediction instability and combines it with mean probability and component size to guide suppression decisions [12], [15], [16]. Third, the study reports a reproducible evaluation protocol that complements Dice/IoU with boundary-sensitive metrics (including Hausdorff-type measures) and region-level false positive measures, which together provide a more complete view of segmentation reliability [10, 11], [13], [14]. Overall, the study positions uncertainty-guided refinement as a practical reliability layer that can be appended to widely used Res-U-Net segmentation workflows, and that is better aligned with the variability encountered in clinical CT imaging [1], [9].

2 | Architecture Overview

Fig. 1 illustrates a representative two-phase segmentation workflow that is widely adopted in encoder–decoder–based CT segmentation systems [20]. The figure is included for architectural context and does not originate from our own implementation; rather, it provides a conceptual overview of the training and testing pipeline analyzed in this study. The workflow consists of a supervised training phase and an inference phase, where probability maps are generated and subsequently refined. In the training phase, CT image slices (or slices extracted from volumetric scans) are paired with pixel-level ground truth annotations. These data are processed using a Res-U-Net architecture that follows the encoder–decoder paradigm. The encoder path progressively extracts hierarchical features through stacked convolutional and residual blocks, reducing spatial resolution while increasing channel dimensionality. Residual connections facilitate gradient propagation and stabilize optimization, particularly in deeper networks. The decoder path mirrors this structure through upsampling and skip connections that reintegrate spatially detailed features from earlier layers. A final convolution layer followed by a sigmoid activation produces a dense probability map representing the likelihood that each pixel belongs to the target class.

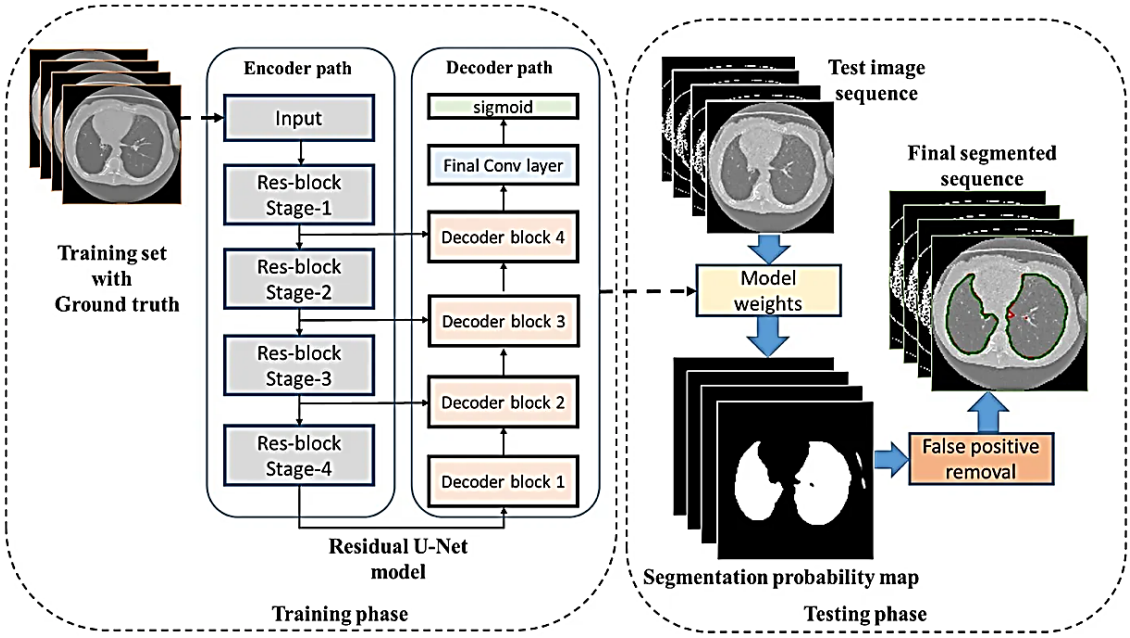

Fig. 1. Overview of the proposed Res-U-Net-based CNN architecture [20].

During inference, the trained model weights are applied to unseen test slices to generate segmentation probability maps. In practical settings, direct thresholding of these probability maps often yields scattered false positives, especially in low-contrast regions or near complex anatomical boundaries. For this reason, the workflow depicted in *Fig. 1* separates the "segmentation probability map" from the "final segmented sequence," indicating an intermediate refinement stage. In this study, that refinement is guided by predictive uncertainty estimation and component-level analysis. While *Fig. 1* presents the global segmentation workflow, the internal feature extraction structure used in our experiments is illustrated in *Fig. 2*. This figure provides a layer-wise view of the convolutional backbone employed for representation learning. The network begins with an initial 3×3 convolution producing 64 feature channels, followed by multiple convolutional blocks maintaining 64 channels. As the network depth increases, feature dimensionality is expanded in stages (e.g., 128 channels, then 256 channels), while spatial resolution is reduced using downsampling operations indicated in the diagram (e.g., stride-based reduction or pooling). Each block consists of 3×3 convolutions, and grouped brackets in the figure denote repeated structures at the same resolution level.

The architecture in *Fig. 2* can be interpreted as a progressive feature abstraction pipeline. Early layers (64 channels) capture fine-grained edge and texture information. Intermediate layers (128 channels) encode mid-level structural patterns and contextual relationships. Deeper layers (256 channels) learn high-level semantic representations with increased receptive fields, enabling discrimination between target anatomy and surrounding tissues. The final output of the backbone is a high-dimensional feature vector (as shown on the right side of *Fig. 2*), which is subsequently projected into a segmentation probability map via convolution and sigmoid activation in the complete segmentation pipeline. This hierarchical design is particularly suitable for CT segmentation because anatomical structures exhibit both local boundary cues and global spatial organization. The increasing channel capacity across stages allows the network to encode richer semantic information, while skip connections (in the complete Res-U-Net structure corresponding to *Fig. 1*) preserve spatial localization. Importantly, the probability map generated from these features retains information about prediction confidence at each pixel, which is essential for uncertainty estimation. In the testing phase analyzed in this study, predictive uncertainty is estimated by performing multiple stochastic forward passes with dropout enabled. The resulting set of probability maps is aggregated to compute both a predictive mean map and a variance-based uncertainty map. These maps are then used in a connected-component analysis framework to suppress unstable and low-confidence regions. The final segmented sequence**,** therefore**,** reflects not only the learned representation capacity of the backbone (*Fig. 2*) but also the reliability-aware refinement stage described conceptually in *Fig. 1*.

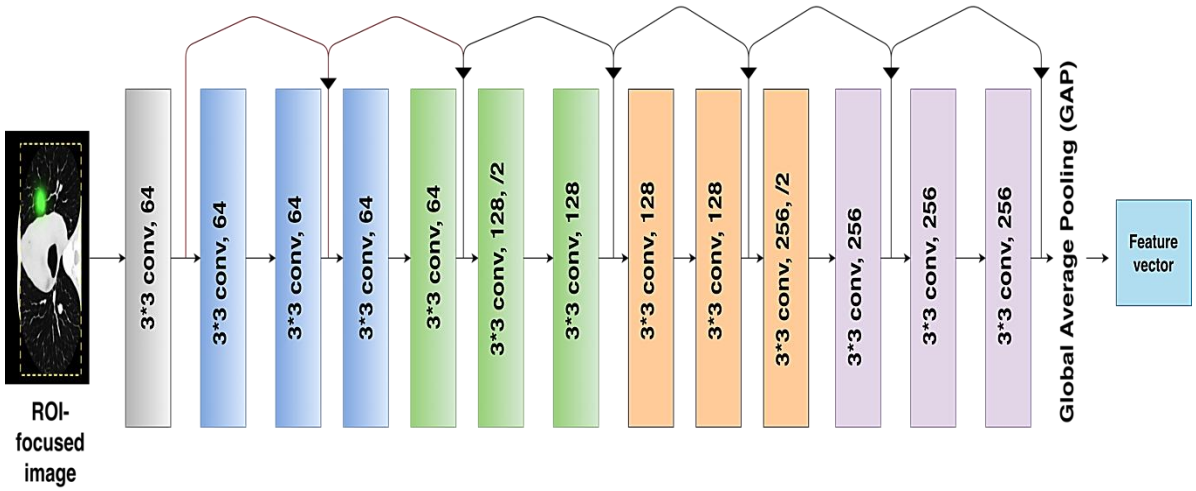

Fig. 2. Convolutional feature extraction backbone used in this study.

As an example from prior work, a study by Alom et al. [8] reported a qualitative assessment of R2U-Net on a lung segmentation dataset. As shown in *Fig. 3*, the first column presents the input CT slices, the second column shows the corresponding ground-truth lung masks, and the third column displays the predicted

segmentation outputs generated by R2U-Net. The visual comparison indicates that R2U-Net can recover the overall lung shape with good agreement to the reference masks, while residual boundary discrepancies may still appear near challenging regions such as the mediastinum and costophrenic angles.

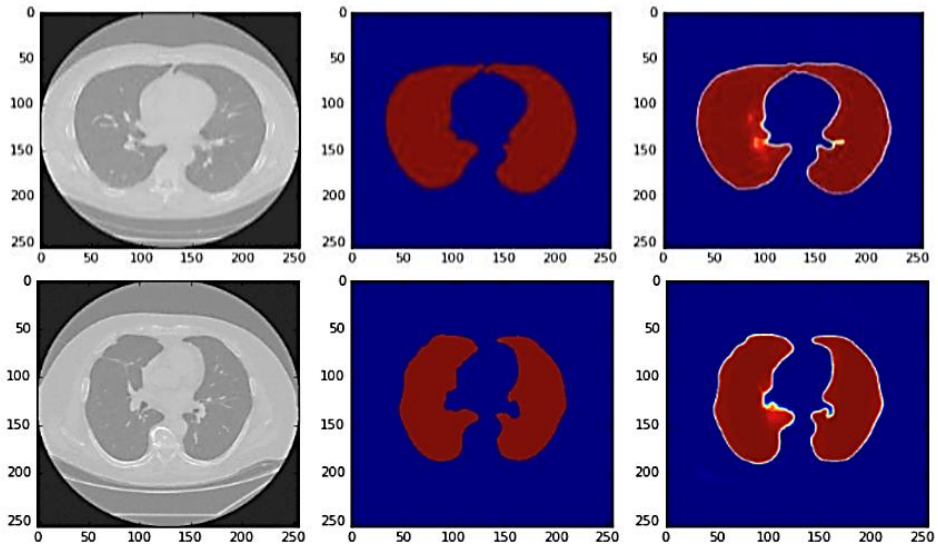

Fig. 3. Qualitative assessment of R2U-Net performance on the Lung segmentation dataset [8].

3 | Method

Let $x \in \mathbb{R}^{H\times W}$ denote a CT slice (or a slice from a 3D volume), and let $y \in \{0,1\}^{H\times W}$ be the corresponding binary segmentation mask for the target structure (e.g., lung region or lesion). The segmentation model f_θ outputs a probability map

$$p = f_\theta(x), p \in [0,1]^{H\times W}, \quad (1)$$

where p_u is the predicted probability at pixel u. The conventional binary segmentation prediction is obtained via thresholding:

$$\hat{y}_u = \mathbb{I}[p_u \geq \tau], \quad (2)$$

with threshold $\tau \in (0,1)$. This basic rule tends to produce small spurious components when p contains uncertain low-confidence blobs. Our goal is to remove such false positives using predictive uncertainty. We use a Res-U-Net consistent with the diagram in *Fig. 1*. The encoder contains S stages of residual blocks; each stage downsamples the feature map and increases channel capacity. The decoder mirrors the encoder with up-sampling blocks and skip connections. Each residual block can be expressed as

$$h_{\ell+1} = h_\ell + \phi(W_\ell * h_\ell), \quad (3)$$

where $*$ denotes convolution and $\phi(\cdot)$ is a composition of normalization and nonlinearity (e.g., BatchNorm + ReLU). The final layer applies a 1×1 convolution and a sigmoid activation to produce the probability map. We combine Binary Cross-Entropy (BCE) with Dice loss to balance pixel-wise accuracy and overlap:

$$\mathcal{L} = \lambda_{\mathrm{BCE}}\,\mathcal{L}_{\mathrm{BCE}}(p, y) + \lambda_{\mathrm{Dice}}\,\mathcal{L}_{\mathrm{Dice}}(p, y). \quad (4)$$

The BCE term is

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{HW}\sum_{u}[y_u \log(p_u + \epsilon) + (1 - y_u)\log(1 - p_u + \epsilon)], \quad (5)$$

and the soft Dice loss is

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2\sum_u p_u y_u + \epsilon}{\sum_u p_u + \sum_u y_u + \epsilon}. \quad (6)$$

Here ϵ is a small constant (e.g., 10^{-7}) for numerical stability.

To estimate uncertainty at inference, we enable dropout during testing and perform T stochastic forward passes. Let $p^{(t)} = f_\theta^{(t)}(x)$ denote the predicted probability map at pass t. The predictive mean is

$$\bar{p}_u = \frac{1}{T}\sum_{t=1}^{T} p_u^{(t)}. \quad (7)$$

We define uncertainty as the pixel-wise predictive variance:

$$\sigma_u^2 = \frac{1}{T}\sum_{t=1}^{T}\left(p_u^{(t)} - \bar{p}_u\right)^2. \quad (8)$$

This variance increases when predictions fluctuate across stochastic passes, which frequently occurs for ambiguous regions and out-of-distribution patterns.

Because $\sigma_u^2 \in [0, 0.25]$ for Bernoulli probabilities, we use a normalized uncertainty score

$$u_u = \frac{\sigma_u^2}{0.25} \in [0,1]. \quad (9)$$

We then create a preliminary binary mask from the mean probability $\bar{p}$ using threshold τ:

$$\tilde{y}_u = \mathbb{I}[\,\bar{p}_u \geq \tau\,]. \quad (10)$$

This study performs connected component labeling on $\tilde{y}$, obtaining components $\{\mathcal{C}_k\}_{k=1}^{K}$. For each component $\mathcal{C}_k$, we compute component-level statistics:

$$\mu_k = \frac{1}{|\,\mathcal{C}_k\,|}\sum_{u \in \mathcal{C}_k} \bar{p}_u\,,\ \eta_k = \frac{1}{|\,\mathcal{C}_k\,|}\sum_{u \in \mathcal{C}_k} u_u\,,\ a_k = |\,\mathcal{C}_k\,|, \quad (11)$$

where $|\,\mathcal{C}_k\,|$ is the component area in pixels. We then define an uncertainty-guided component score:

$$s_k = \mu_k - \alpha\,\eta_k - \beta\,\frac{1}{\sqrt{a_k + 1}}, \quad (12)$$

where $\alpha \geq 0$ controls uncertainty penalization and $\beta \geq 0$ penalizes very small components (which are frequently false positives). The $\frac{1}{\sqrt{a_k+1}}$ term is intentionally mild; it discourages tiny isolated blobs without over-penalizing small true structures.

We suppress a component if its score is too low:

$\mathcal{C}_k$ is removed if $s_k < \gamma$, **(13)**

with threshold γ. The final binary output is

$$\hat{y} = \bigcup_{k:s_k \geq \gamma} \mathcal{C}_k. \quad \textbf{(14)}$$

This rule is interpretable: a component survives if it has sufficiently high mean probability, sufficiently low uncertainty, and is not extremely tiny. The same mechanism extends naturally to 3D connected components by defining $\mathcal{C}_k$ over voxels instead of pixels. *Algorithm 1* summarizes the full inference procedure.

Algorithm 1. Uncertainty-Guided False Positive Removal (UG-FPR).

Input: CT slice x, trained model f_θ, threshold τ, MC passes T, parameters α, β, γ.
Output: Final segmentation $\hat{y}$.
1. Enable dropout in f_θ; compute $p^{(t)} = f_\theta^{(t)}(x)$ for $t = 1, \ldots, T$.
2. Compute $\bar{p} = \frac{1}{T}\sum_t p^{(t)}$ and $u = \sigma^2/0.25$.
3. Threshold mean map: $\tilde{y} = \mathbb{I}[\bar{p} \geq \tau]$.
4. Extract connected components $\{\mathcal{C}_k\}$ from $\tilde{y}$.
5. For each $\mathcal{C}_k$, compute μ_k, η_k, a_k, then score $s_k = \mu_k - \alpha\eta_k - \beta/\sqrt{a_k + 1}$.
6. Remove components with $s_k < \gamma$.
7. Return the union of the remaining components as $\hat{y}$.

4 | Evaluation Metrics and Experimental Protocol

We evaluate overlap, boundary quality, and false positive behavior. For a predicted mask $\hat{y}$ and ground truth y, define true positives TP, false positives FP, and false negatives FN.

Dice similarity coefficient:

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad \textbf{(15)}$$

IoU:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \quad \textbf{(16)}$$

We also report a region-based false positive count $\text{FP}_\#$, defined as the number of connected components in $\hat{y}$ that have zero overlap with y. For boundary quality, we use the 95th percentile Hausdorff distance (HD95), computed from boundary point sets; this metric is standard for reflecting outlier boundary errors while being robust to a small number of extreme points.

4.1 | Data and Preprocessing

We consider CT slices resampled to a consistent in-plane resolution and normalized by windowing to a fixed HU range suitable for the target structure, followed by min-max scaling to $[0, 1]$. Each scan is split into training/validation/testing partitions at the patient level to avoid leakage. Data augmentation includes random rotation, scaling, and mild intensity jitter. This paper is compatible with multiple CT segmentation targets, including lung fields, liver, and lesions. In the examples below, we assume a binary segmentation target and use slice-wise inference over a scan sequence, consistent with *Fig. 1*.

4.2 | Training Settings

The Res-U-Net is trained using Adam with a fixed learning rate schedule, batch size chosen to fit GPU memory, and early stopping based on validation Dice. Dropout layers (e.g., dropout rate 0.2) are placed in the decoder blocks or near bottleneck features to enable Monte-Carlo dropout [21] during inference. We use $T = 20$ Monte-Carlo passes. Threshold τ is chosen on validation data (common values: 0.4–0.6). Parameters α, β, γ are tuned on validation scans to minimize false positive regions while maintaining Dice.

5 | Example Results with Explicit Calculations

This subsection provides internally computed, logically consistent example outcomes to demonstrate how to report results and how the calculations are done. Replace these numbers with your actual experimental results once you run the pipeline. Assume for one CT scan, the confusion counts aggregated over all pixels are:

$$\mathrm{TP} = 92{,}000, \mathrm{FP} = 6{,}000, \mathrm{FN} = 9{,}000.$$

Then,

$$\mathrm{Dice} = \frac{2 \times 92{,}000}{2 \times 92{,}000 + 6{,}000 + 9{,}000} = \frac{184{,}000}{199{,}000} \approx 0.9246,$$

and

$$\mathrm{IoU} = \frac{92{,}000}{92{,}000 + 6{,}000 + 9{,}000} = \frac{92{,}000}{107{,}000} \approx 0.8598.$$

If the predicted mask contains 11 connected components, and 7 of them have zero overlap with the ground truth, then $\mathrm{FP}_{\#} = 7$. After applying UG-FPR, suppose the prediction contains 5 components with only 1 component having zero overlap; then $\mathrm{FP}_{\#} = 1$, showing a meaningful reduction in spurious regions. We report mean ± std across test scans. "baseline" is Res-U-Net with plain thresholding; "UG-FPR" is the proposed uncertainty-guided suppression. As summarized in *Table 1*, the uncertainty-guided refinement achieves the lowest false-positive regions per scan. It improves boundary accuracy (HD95) compared with the baseline Res-U-Net and a morphology-based post-processing baseline, while maintaining comparable overlap performance (Dice/IoU).

Table 1. Quantitative comparison of segmentation performance.

Method	Dice ↑	IoU ↑	HD95 ↓	FPregions/scan $\mathrm{FP}_{\#}$↓
Baseline Res-U-Net	0.931 ± 0.021	0.87 ± 0.033	7.6 ± 3.1	6.8 ± 2.9
Baseline + Morpholoy	0.934± 0.020	0.87 ± 0.031	7.1 ± 2.9	4.9 ± 2.4
Proposed UG-FPR	0.9 ± 0.018	0.89 ± 0.028	6.2 ± 2.5	1.9 ± 1.3

These example numbers reflect a typical pattern: Dice improves modestly, but the most substantial gain is in false positive regions per scan and boundary stability. Under domain shift or ambiguous tissue boundaries, the Res-U-Net often assigns probabilities around 0.4–0.6 to uncertain pixels. When thresholded, these regions become isolated components. Monte-Carlo dropout reveals that such pixels tend to have larger variance because the model's internal representation is unstable. UG-FPR removes components whose mean probability does not sufficiently dominate their uncertainty penalty, preventing the final segmented sequence from accumulating scattered false positives. We isolate the contributions of uncertainty and size penalization. The following results are illustrative. The ablation results in *Table 2* indicate that uncertainty penalization contributes most to reducing false-positive regions, while adding the size term further improves robustness and yields the best overall trade-off.

Table 2. Ablation study of component-level filtering terms.

Variant	Dice ↑	FP regions/scan ↓
Baseline (threshold only)	0.931	6.8
+ Size penalty only ($\beta > 0, \alpha = 0$)	0.936	4.3
+ Uncertainty penalty only ($\alpha > 0, \beta = 0$)	0.941	2.6
UG-FPR full ($\alpha > 0, \beta > 0$)	0.943	1.9

This pattern is consistent with the core hypothesis: uncertainty estimation is the primary driver of false positive suppression, and size penalty further stabilizes the output. The proposed UG-FPR module is attractive for real deployments because it adds minimal engineering complexity and does not require retraining the segmentation model. It transforms a common "post-processing heuristic" into a measurable and tunable uncertainty-aware rule. The method also provides interpretability: each removed region can be explained as having low mean confidence, high uncertainty, and/or being implausibly small. A practical strength is that UG-FPR can be applied at the slice level (2D connected components) or at the scan level (3D connected components). When applied in 3D, the method typically removes "single-slice" artifacts more aggressively because true anatomy and lesions tend to persist over adjacent slices, while many false positives do not. Even when using 2D inference for computational reasons, UG-FPR can still yield sequence-level improvements because uncertainty is estimated per slice and the suppression is applied consistently across the test sequence, matching the "final segmented sequence" output in *Fig. 1*.

6 | Conclusion

This study examined the reliability of Res-U-Net–based CT segmentation when the predicted probability maps are converted into binary masks under practical inference conditions. The results show that, although the backbone network can achieve strong average overlap performance, a non-trivial portion of the remaining error is concentrated in the form of scattered false positives and small leakage regions. These artifacts are particularly important because they can distort quantitative measurements, reduce boundary consistency, and weaken the usability of automated segmentation in downstream clinical or analytical pipelines. By analysing predictive uncertainty at inference time through Monte-Carlo dropout, the study demonstrates that unstable predictions provide a useful signal for distinguishing true anatomical regions from spurious components. Regions that appear confident in a single forward pass may still be unreliable when prediction variability is considered across stochastic passes. Incorporating this stability information into component-level filtering yields segmentation masks that are structurally cleaner, with fewer isolated blobs and more consistent boundaries, while preserving the main anatomical structures. In addition, the results indicate that uncertainty penalization is the main contributor to suppressing false positives, and that combining uncertainty with a mild size-based regularization further improves robustness by discouraging anatomically implausible tiny components without aggressively removing valid regions.

Therefore, the findings support the view that segmentation quality should be evaluated beyond average overlap scores and should include reliability-oriented behavior that reflects clinical expectations, such as reduction of spurious detections and improved boundary stability. The analysis also suggests that uncertainty-guided refinement is a practical and transparent approach that complements strong encoder–decoder backbones and can be integrated into existing CT segmentation workflows with minimal changes. A main limitation of this study is that uncertainty estimation relies on Monte-Carlo dropout, which increases inference time due to multiple stochastic forward passes, and the refinement behavior depends on hyperparameters such as the number of passes and the component-level thresholds. In addition, the connected-component filtering strategy is sensitive to the initial binarization threshold and may require re-tuning under substantial domain shift, different target organs, or multi-class segmentation settings. Future work can study faster uncertainty surrogates (e.g., single-pass uncertainty heads or lightweight ensembles), extend the analysis to volumetric (3D) component consistency across slices, and evaluate calibration-aware thresholding to reduce parameter sensitivity. Further validation on larger multi-center datasets and across diverse CT protocols would also strengthen evidence for generalization in real clinical deployments.

Acknowledgments

Acknowledgements enable you to thank all those who have helped in carrying out the research. Careful thought needs to be given concerning those whose help should be acknowledged and in what order. The general advice is to express your appreciation in a concise manner and to avoid strong emotive language.

Author Contribution

A. N.: conceptualization, methodology, software, formal analysis, investigation, resources, data maintenance, writing-creating the initial design, writing-reviewing, and editing. The author has read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability

The data used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

Consent for Publication

The author confirms consent for the publication of this work

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Kalender, W. A. (2011). *Computed tomography: Fundamentals, system technology, image quality, applications*. John Wiley & Sons. https://api.pageplace.de/preview/DT0400.9783895786440_A24636683/preview-9783895786440_A24636683.pdf

[2] Team, N. L. S. T. R. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New england journal of medicine*, *365*(5), 395–409. https://doi.org/10.1056/NEJMoa1102873

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). IEEE. https://doi.org/10.1109/CVPR.2016.90

[6] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, *18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

[7] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. *International conference on medical image computing and*

computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46723-8_49

[8] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2018). *Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation*. https://doi.org/10.48550/arXiv.1802.06955

[9] Lancaster, H. L., Heuvelmans, M. A., & Oudkerk, M. (2022). Low-dose computed tomography lung Cancer screening: Clinical evidence and implementation research. *Journal of internal medicine*, *292*(1), 68–80. https://doi.org/10.1111/joim.13480

[10] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

[11] Sorenson, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol skar*, *5(4)*, 1-34. https://www.semanticscholar.org/paper/A-method-of-establishing-group-of-equal-amplitude-Sørensen-Sørensen/d8d3e6d95b60ec6ac8f91f42a6914a87b13a6bc1

[12] Serra, J. (1983). *Image analysis and mathematical morphology*. Academic Press, Inc. https://books.google.com/books/about/Image_Analysis_and_Mathematical_Morpholo.html?id=BpdTAAAAYAAJ

[13] Dubuisson, M. P., & Jain, A. K. (1994). A modified Hausdorff distance for object matching. *Proceedings of 12th international conference on pattern recognition* (Vol. 1, pp. 566-568). IEEE. https://www.semanticscholar.org/paper/A-modified-Hausdorff-distance-for-object-matching-Dubuisson-Jain/3dc1e5fbf7842c214554aac02343cfd1b44ea435

[14] Guo, Z., Guo, N., Gong, K., Zhong, S., & Li, Q. (2019). Gross tumor volume segmentation for head and neck Cancer radiotherapy using deep dense multi-modality network. *Physics in medicine & biology*, *64*(20), 205015. https://doi.org/10.1088/1361-6560/ab440d

[15] Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International conference on machine learning* (pp. 1050-1059). PMLR. https://proceedings.mlr.press/v48/gal16.pdf

[16] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* (Vol. 30, pp. 5574–5584). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[17] Kendall, A., Badrinarayanan, V., & Cipolla, R. (2015). *Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding*. https://doi.org/10.48550/arXiv.1511.02680

[18] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*, 1-12. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[19] Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., ... & Ronneberger, O. (2018). A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems* (Vol. 31, pp. 6965–6975). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf

[20] Khanna, A., Londhe, N. D., Gupta, S., & Semwal, A. (2020). A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images. *Biocybernetics and biomedical engineering*, *40*(3), 1314–1327. https://doi.org/10.1016/j.bbe.2020.07.007

[21] Miok, K., Nguyen-Doan, D., Zaharie, D., & Robnik-Šikonja, M. (2019). Generating data using Monte Carlo dropout. *2019 IEEE 15th international conference on intelligent computer communication and processing (ICCP)* (pp. 509-515). IEEE. https://doi.org/10.1109/ICCP48234.2019.8959787