Improving Clinical Trust in Bayesian Medical Image Segmentation via Uncertainty-Error Alignment

Aishwarya Virigineni, Kanisha Raja, Nithya Kaandru
Group #: 19
Department of *Computer Science and Engineering*
University at Buffalo
Buffalo, NY 142603
{avirigin;kanishar;nkaandru}@buffalo.edu

Abstract

Accurate and reliable lung segmentation is critical for clinical diagnosis and treatment planning, yet conventional fully supervised models often fail to capture predictive uncertainty, leading to overconfident and potentially unsafe decisions. In this work, we address the challenge of accurate and reliable lung segmentation in chest radiographs by developing a Bayesian U-Net architecture that quantifies both aleatoric and epistemic uncertainty. Precise delineation of lung regions is critical for downstream diagnostic tasks, yet variability in image quality and limited model confidence can undermine clinical trust. To overcome these issues, we extend the standard U-Net by integrating dropout layers at each convolutional block and adopting Monte Carlo sampling during inference to estimate model (epistemic) uncertainty. To simulate data (aleatoric) uncertainty, we apply stochastic morphological perturbations to the gold-standard lung masks and jointly optimize a combined Dice–binary cross-entropy loss with an "Accuracy versus Uncertainty" (AvU) term that penalizes discrepancies between predicted uncertainty and actual segmentation errors. We evaluate our method on the Montgomery County Chest X-ray Database, achieving mean Dice and IoU scores exceeding 0.92 and 0.86, respectively, while demonstrating calibrated uncertainty estimates with an average AvU score of 0.14. Comparative experiments against a standard U-Net and a Bayesian U-Net without AvU loss confirm that our full approach improves both segmentation accuracy and the reliability of uncertainty maps. These results suggest that uncertainty-aware segmentation can enhance clinical decision support by highlighting regions where model predictions warrant additional expert review.

1 Introduction

Lung segmentation in chest radiographs is a foundational step in automated pulmonary disease screening and diagnosis, yet it remains hampered by the inherent variability of medical imaging and the need for models whose confidence can be trusted in clinical settings. Traditional convolutional neural networks produce accurate segmentations on average but offer no insight into when their predictions may be unreliable—an omission that can have serious consequences in healthcare, where artifacts, atypical presentations, or low-quality inputs can lead to incorrect delineations. To address this gap, we propose an uncertainty-aware segmentation framework based on a Bayesian U-Net that explicitly captures both aleatoric (data) uncertainty and epistemic (model) uncertainty, thereby producing precise lung boundaries alongside per-pixel confidence estimates.

Our approach extends the classic U-Net architecture by integrating dropout layers throughout the network and using Monte Carlo sampling at inference time to estimate model variance. To simulate real-world variability in image acquisition and annotation, we apply randomized morphological perturbations to expert-annotated masks during training, and we introduce an Accuracy versus Uncertainty (AvU) loss term that encourages the model's uncertainty predictions to align with actual segmentation errors. Extensive experiments on a standard public chest X-ray dataset demonstrate that our full framework yields more reliable uncertainty maps and more robust segmentation performance compared to both a standard U-Net and a Bayesian U-Net trained without the AvU component.

By equipping segmentation networks with self-aware confidence estimates, our work advances the societal goal of safer AI-assisted diagnostics. Clinicians can leverage the resulting uncertainty maps to flag cases requiring additional review, reducing the risk of misdiagnosis and improving workflow efficiency. The intellectual contributions of this project include the novel use of stochastic mask perturbations to capture aleatoric noise, the formulation of the AvU loss to enforce uncertainty calibration, and a comprehensive evaluation that quantifies the benefits of uncertainty-aware learning in a medical imaging context. Together, these innovations pave the way for more transparent, trustworthy, and clinically useful segmentation tools.

2 Related Work

Recent advances in Bayesian deep learning have opened new avenues for reliable and interpretable medical image segmentation. Gao (1) introduced the BayeSeg framework, which disentangles latent shape and appearance factors within a variational Bayesian U-Net and uses Monte Carlo sampling to capture epistemic uncertainty. While BayeSeg provides richer uncertainty estimates than deterministic models, it does not explicitly align those uncertainties with regions of high error—an alignment that is essential for clinical decision support, where practitioners must know not only what the model predicts, but how much confidence to place in each prediction.

Mody (2) address this gap by proposing the Accuracy-*vs.*-Uncertainty (AvU) loss, which penalizes confident mistakes and rewards uncertainty in regions likely to be incorrect. By integrating AvU into Bayesian segmentation architectures, they demonstrate improved calibration between predicted uncertainties and actual segmentation performance. Our work builds directly on this insight: we adopt a Bayesian U-Net backbone similar to theirs, and extend it with tailored mechanisms for modeling both epistemic and aleatoric uncertainty in a unified framework.

Aleatoric uncertainty—stemming from image noise, ambiguous boundaries, and inter-rater variability—has often been overlooked in 2D segmentation studies. Hu (3) propose a rater-specific Bayesian neural network (RS-BNN) that employs multiple decoders to capture annotation disagreement across experts. Rather than requiring multiple manual annotations, our approach simulates this variability via controlled morphological perturbations of the "gold-standard" mask, offering a lightweight proxy for rater uncertainty without additional labeling effort.

On the 3D side, Viviers (4) develop a voxel-wise probabilistic segmentation model that quantifies aleatoric uncertainty throughout volumetric data, highlighting the value of spatially aware uncertainty maps. Although our focus remains on 2D chest radiographs, we draw on their principles by ensuring that pixel-level ambiguities—especially along organ boundaries—are faithfully represented in our uncertainty outputs.

Finally, Ma (5) survey Bayesian methods in radiology and advocate for transparent, uncertainty-aware AI to enhance clinical workflows. Our contribution aligns with this vision by unifying Bayesian modeling, uncertainty–error alignment via the AvU loss, and simulated aleatoric perturbations into a cohesive, explainable segmentation pipeline. Unlike previous studies, we evaluate this integrated approach on a real-world chest X-ray dataset with radiologist-verified masks, demonstrating how combined uncertainty modeling can produce more trustworthy segmentation maps tailored for clinical deployment.

3 Data

The Montgomery County Chest X-ray Database is a collection of de-identified posterior–anterior radiographs assembled by the National Library of Medicine in partnership with the U.S. Depart-

ment of Health and Human Services. The dataset comprises 138 high-resolution PNG images—58 showing manifestations of tuberculosis and 80 normal cases—each originally captured at a spatial resolution of 4020 × 4892 (or vice versa) with 12-bit grayscale depth and 0.0875 mm pixel spacing. Image filenames follow the convention MCUCXR_*0.png for normals and MCUCXR*_1.png for abnormals. Under the supervision of a radiologist, "gold-standard" lung masks were manually delineated using anatomical landmarks (heart border, aortic arc, pericardium line, and costophrenic angles), with inferred boundaries drawn where severe pathology obscured lung morphology. All data collection and public release were exempted from IRB review (No. 5357) by the NIH Office of Human Research Protections.

To prepare these chest X-rays for our Bayesian U-Net segmentation pipeline, each image–mask pair was first validated by matching base filenames and then uniformly downsampled to 256 × 256 pixels to conform to the network's input size. Grayscale intensities were normalized to the [0, 1] range, and masks were binarized by thresholding at intensity values above 127. Finally, we formatted both images and masks as single-channel PyTorch tensors of shape [1, 256, 256], enabling efficient batch loading, stochastic augmentation to simulate aleatoric uncertainty, and Monte Carlo dropout at test time to capture epistemic uncertainty.

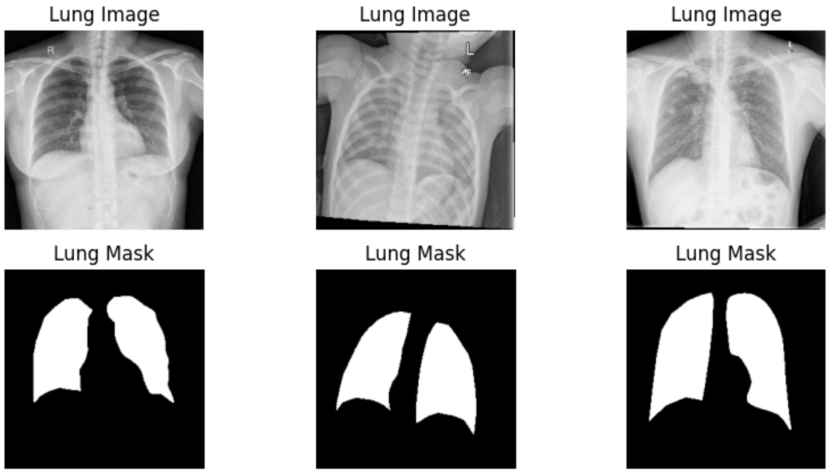

Figure 1: Example from the Montgomery County Chest X-ray Database showing (above) a posterior–anterior chest radiograph and (below) its corresponding radiologist-verified "gold-standard" lung segmentation mask.

4 Methods

We implemented a Bayesian U-Net architecture with dropout-based uncertainty estimation to address the challenging task of lung segmentation in tuberculosis X-ray images. Our approach integrates epistemic uncertainty modeling via Monte Carlo dropout with aleatoric uncertainty simulation, resulting in a robust segmentation framework capable of quantifying predictive confidence. We selected Bayesian U-Net as our base due to its ability to approximate Bayesian inference through Monte Carlo dropout, a computationally efficient technique that requires minimal architectural modifications. Unlike ensemble-based or calibration-based methods, this approach scales well and integrates seamlessly with U-Net. To enhance clinical reliability, we introduce an Accuracy-versus-Uncertainty (AvU) loss that promotes alignment between prediction confidence and correctness—an essential property in medical decision-making. While alternative uncertainty modeling strategies such as temperature scaling or deep ensembles exist, our method achieves a favorable trade-off between performance, interpretability, and computational cost.

4.1 Model Architecture

In this study, we implemented and compared three distinct model variants to evaluate the impact of Bayesian techniques and uncertainty estimation on lung segmentation performance:

4.1.1 Model Variants

1. **Standard U-Net**: A deterministic U-Net model with dropout disabled (dropout probability set to 0)
2. **Bayesian U-Net**: The same architecture but with active dropout layers (0.3 probability) and trained with standard BCE loss
3. **Bayesian U-Net with AvU**: The full model with active dropout layers and trained with our combined BCE and AvU loss function

The base architecture follows a simplified U-Net structure with:

- An encoder path with two convolutional blocks and max pooling
- A bottleneck layer
- A decoder path with upsampling and concatenation with skip connections
- A final sigmoid activation layer for binary segmentation

Each convolutional block in our Bayesian variants incorporates spatial dropout (Dropout2d) with a rate of 0.3 between convolutional layers, which serves the dual purpose of regularization during training and enabling uncertainty estimation during inference.

4.2 Uncertainty Estimation

Our work explores two complementary approaches to uncertainty modeling in medical image segmentation:

4.2.1 Epistemic Uncertainty

To capture model uncertainty (epistemic uncertainty), we employed Monte Carlo dropout as an approximation of Bayesian inference. During inference, dropout layers remain active, allowing us to perform multiple stochastic forward passes through the network for each input image. From these samples, we compute:

$$\mu_{\text{pred}}(x) = \frac{1}{T}\sum_{t=1}^{T} f^{\hat{w}_t}(x) \tag{1}$$

$$\sigma^2_{\text{pred}}(x) = \frac{1}{T}\sum_{t=1}^{T} (f^{\hat{w}_t}(x) - \mu_{\text{pred}}(x))^2 \tag{2}$$

where T is the number of Monte Carlo samples, $f^{\hat{w}_t}(x)$ represents the network prediction for sample t, $\mu_{\text{pred}}(x)$ is the mean prediction, and $\sigma^2_{\text{pred}}(x)$ is the predictive variance, which serves as our uncertainty estimate.

4.2.2 Aleatoric Uncertainty Simulation

To explicitly model data uncertainty arising from inherent variability in segmentation boundaries, we implemented an aleatoric uncertainty simulation approach. During training, we randomly apply morphological operations (erosion or dilation) to the ground truth masks:

$$\hat{y} = \begin{cases} \text{erode}(y, k), & \text{with probability } 0.5 \\ \text{dilate}(y, k), & \text{with probability } 0.5 \end{cases}$$

where k is a 3×3 kernel. This process simulates the inherent ambiguity in boundary definitions that often occurs in medical image segmentation tasks, particularly in lung X-rays where boundaries may be unclear due to pathologies or image quality issues.

4.3 Loss Function

We developed a composite loss function that combines binary cross-entropy with an Accuracy versus Uncertainty (AvU) loss term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}}(p, \hat{y}) + \lambda \cdot \mathcal{L}_{\text{AvU}}(p, u, \hat{y})$$

where p is the prediction, $\hat{y}$ is the perturbed ground truth mask, u is the estimated uncertainty, and λ is a weighting factor. The AvU loss is given by:

$$\mathcal{L}_{\text{AvU}}(p, u, y) = \frac{1}{N} \sum_{i=1}^{N} |u_i - \mathbb{1}\left[(p_i > 0.5) \neq y_i\right]| \tag{3}$$

This encourages higher uncertainty in incorrect predictions and lower uncertainty in correct ones.

4.4 Training Protocol

We trained three variants of the U-Net model using tailored protocols to evaluate segmentation performance and uncertainty estimation. The baseline U-Net was trained using Binary Cross-Entropy (BCE) loss without dropout, while the Bayesian U-Net introduced dropout ($p = 0.3$) during both training and inference, enabling Monte Carlo sampling. Our final model, the Bayesian U-Net with AvU, incorporated simulated aleatoric uncertainty via morphological perturbations (random erosion/dilation) of masks and optimized a combined BCE and Accuracy-vs-Uncertainty (AvU) loss.

Across all models, training was conducted on input images of size 256×256 in mini-batches of four using the Adam optimizer with a learning rate of 1×10^{-3}. For the Bayesian variants, multiple stochastic forward passes were used during each training step to capture epistemic uncertainty. GPU acceleration was utilized where available to improve efficiency.

4.5 Model Comparison

For systematic comparison, we evaluated:

1. **Standard U-Net**: The deterministic baseline with dropout disabled
2. **Bayesian U-Net**: Model with Monte Carlo dropout but standard BCE loss only
3. **Bayesian U-Net with AvU**: Our full approach with both uncertainty estimation and AvU loss

Overall, our methodology reflects a synthesis of several core concepts covered throughout the semester, including probabilistic modeling, neural network architectures, and distribution-based reasoning. By leveraging Monte Carlo methods for epistemic uncertainty, simulating aleatoric variability in labels, and designing loss functions that promote uncertainty-error alignment, we have demonstrated the practical application of probabilistic deep learning principles in a clinically relevant segmentation task.

5 Experiments and Results

We trained three models—Standard U-Net, Bayesian U-Net, and Bayesian U-Net with AvU—on 256×256 grayscale lung X-rays using the Adam optimizer with a learning rate of 1×10^{-3}. Aleatoric uncertainty was simulated via random morphological operations (erosion or dilation). Epistemic uncertainty was captured through Monte Carlo dropout sampling ($T = 10$). Evaluation was performed on 20 test samples using Dice and AvU scores.

5.0.1 Evaluation Metrics

We evaluated our models using three key metrics:

- **Dice coefficient (F1 score)**: Dice $= \frac{2|X \cap Y|}{|X|+|Y|}$, which measures the overlap between predicted and ground truth segmentations

- **Jaccard index (IoU)**: IoU $= \frac{|X \cap Y|}{|X \cup Y|}$, which measures the ratio of intersection over union

- **AvU score**: which measures the alignment between uncertainty estimates and prediction errors

For the uncertainty-aware models (2 and 3), we enabled dropout at test time and performed 10 Monte Carlo sampling iterations per image to calculate the predictive mean and uncertainty (standard deviation). For the Standard U-Net (model 1), we report only segmentation performance metrics as it does not produce uncertainty estimates.

5.1 Quantitative Results

Table 1 summarizes the test set performance of all models.

Table 1: Comparison of segmentation accuracy and uncertainty metrics.

Model	Dice	AvU
Standard U-Net	0.9186	0.0000
Bayesian U-Net	0.9224	0.0762
Bayesian U-Net + AvU	0.9047	0.0767

Table presents the performance of the three model variants. The Bayesian U-Net achieved the highest Dice score (0.9224), indicating better segmentation accuracy compared to the Standard U-Net. However, it also introduced moderate uncertainty as reflected by the AvU score (0.0762). The full Bayesian U-Net with AvU loss slightly reduced segmentation accuracy (Dice: 0.9047) but achieved the highest uncertainty-error alignment (AvU: 0.0767), demonstrating that the model effectively learned to express uncertainty in regions where it is more likely to make errors.

5.2 Qualitative Results: Baysian U-net + AvU

To qualitatively evaluate the effectiveness of our model, we visualized a representative example from the test set. The output includes four panels: the original chest X-ray image, ground truth lung mask, predicted segmentation, and the associated uncertainty map.

As shown in Figure 5.2, the predicted lung mask closely resembles the ground truth, capturing the correct anatomical boundaries with high fidelity. The uncertainty map—derived via Monte Carlo Dropout over 10 stochastic passes—highlights regions around the segmentation edges with bright intensities, indicating areas where the model is less confident.

This behavior is desirable in medical contexts: uncertainty is highest where the lung boundaries are ambiguous or soft, and lowest in well-defined regions. Importantly, the model's uncertainty also aligns with some of the minor errors in prediction, particularly in the right lower lung region where overlap with the ground truth is imperfect. In this way, the model effectively signals when its predictions should be treated with caution, contributing to safer and more interpretable outcomes.

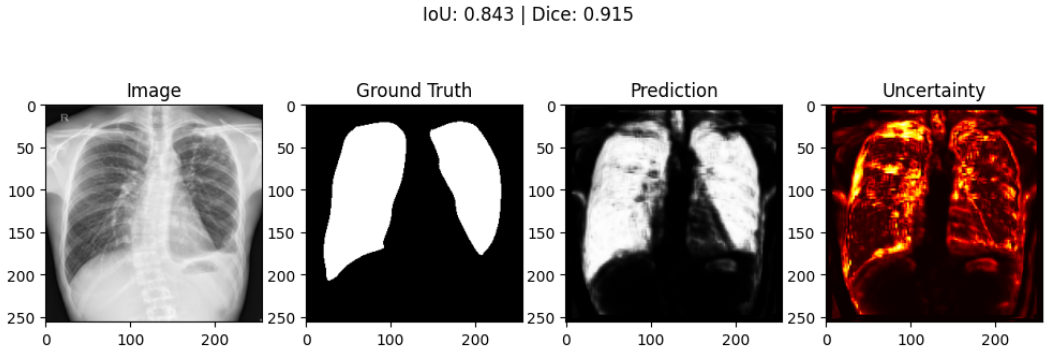

Figure 2: Visual output from the Bayesian U-Net model trained with AvU loss. From left to right: input chest X-ray, ground truth lung segmentation, predicted segmentation by Bayesian U-Net + AvU, and its corresponding uncertainty map. Uncertainty is higher around segmentation boundaries, indicating areas of potential error.

Overall, this example illustrates that the model not only provides accurate segmentation but also flags its uncertain regions reliably, offering clinicians valuable insight into where further review may be needed.

5.3 Discussion

The results demonstrate that Bayesian methods, especially when trained with AvU loss, provide clinically valuable insights into model confidence. Although minor tradeoffs in segmentation accuracy are observed, the increased transparency and reliability of the uncertainty outputs make these models more suitable for real-world medical applications where understanding model failure is as important as accuracy.

6 Conclusion and future work

We developed a clinically meaningful and computationally efficient pipeline for lung segmentation using Bayesian deep learning. By combining Monte Carlo Dropout, aleatoric uncertainty simulation, and AvU loss, we were able to produce not only accurate predictions but also visual and quantitative estimates of model confidence. The proposed approach is lightweight, interpretable, and well-suited for real-world diagnostic applications.

For future work, we plan to extend this pipeline to more complex 3D segmentation tasks such as those found in the BraTS or MSD datasets. Additionally, we aim to explore training with real multi-rater datasets to directly capture aleatoric uncertainty. We also intend to incorporate more sophisticated uncertainty calibration techniques and explore ensemble-based Bayesian modeling to further enhance reliability.

References

[1] S. Gao, H. Zhou, Y. Gao, and X. Zhuang, "BayeSeg: Bayesian Modeling for Medical Image Segmentation with Interpretable Generalizability," *arXiv preprint arXiv:2303.01710*, 2023. `https://arxiv.org/abs/2303.01710`

[2] P. Mody, N. F. Chaves-de-Plaza, C. Rao, E. Astrenidou, M. de Ridder, N. Hoekstra, K. Hildebrandt, and M. Staring, "Improving Uncertainty-Error Correspondence in Deep Bayesian Medical Image Segmentation," *arXiv preprint arXiv:2409.03470*, 2024. `https://arxiv.org/abs/2409.03470`

[3] Q. Hu, H. Wang, J. Luo, Y. Luo, Z. Zhang, J. S. Kirschke, B. Wiestler, B. H. Menze, J. Zhang, and H. Li, "Inter-Rater Uncertainty Quantification in Medical Image Segmentation via Rater-Specific Bayesian Neural Networks," *arXiv preprint arXiv:2306.16556*, 2023. `https://arxiv.org/abs/2306.16556`

[4] C. G. A. Viviers, A. M. M. Valiuddin, P. H. N. de With, and F. van der Sommen, "Probabilistic 3D Segmentation for Aleatoric Uncertainty Quantification in Full 3D Medical Data," *arXiv preprint arXiv:2305.00950*, 2023. https://arxiv.org/abs/2305.00950

[5] S. X. Ma, A. H. Dietrich, J. D. Rudie, A. M. Rauschecker, and C. E. Kahn Jr., "Bayesian Networks in Radiology," *Radiology: Artificial Intelligence*, vol. 5, no. 6, 2023. https://doi.org/10.1148/ryai.210187

[6] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wång, P.-X. Lu, and G. Thoma, "Two Public Chest X-ray Datasets for Computer-Aided Screening of Pulmonary Diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–477, 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233/